Universidad ORT Uruguay

Facultad de Ingeniería

Análisis comparativo de portales de Infraestructuras de Datos Espaciales en base a técnicas de *Text Mining*

Entregado como requisito para la obtención del título

Ingeniera en Sistemas

Noelia Bentancor - 242970

Tutor: Esther Hochsztain

2025

Declaratoria de Autoría

Yo, Noelia Bentancor, declaro que el trabajo que se presenta en esa obra es de mi propia mano. Puedo asegurar que:

- La obra fue producida en su totalidad mientras realizaba el Proyecto de Grado.
- Cuando he consultado el trabajo publicado por otros, lo he atribuido con claridad;
- Cuando he citado obras de otros, he indicado las fuentes. Con excepción de estas citas, la obra es enteramente mía;
- Cuando la obra se basa en trabajo realizado conjuntamente con otros, he explicado claramente qué fue contribuido por otros, y qué fue contribuido por mi;
- Ninguna parte de este trabajo ha sido publicada previamente a su entrega, excepto donde se han realizado las aclaraciones correspondientes.

Noelia Bentancor
16/10/2025

Agradecimientos

A mi familia, por estar presente y acompañarme incondicionalmente a lo largo de todo el proceso.

A mis amigos, por su constante apoyo y por ser sostén emocional en los momentos más exigentes.

A mi tutora, Esther Hochsztain, por su guía y acompañamiento durante todo el proceso de investigación. Por haberme transmitido sus conocimientos y acercado al mundo de la investigación.

A Juan Gabito y Cecilia Belleti que me permitieron desarrollar mi primer proyecto de investigación en la materia Pasantía Académica, donde comencé a descubrir mi interés por la investigación, motivo por el cual hoy elijo continuar por ese camino.

Abstract

Vivimos en la era de la información, en la que la *web* se ha consolidado como un vasto repositorio de datos. Entre ellos, las Infraestructuras de Datos Espaciales (IDEs) destacan por integrar tecnologías, políticas, estándares y recursos humanos para la gestión de información geoespacial. La información contenida en estos repositorios evoluciona constantemente y, además, gran parte de ella es textual, lo que dificulta su análisis. Este escenario plantea un desafío tanto para mantenerse a la vanguardia de los datos como para comparar distintas IDEs. Con el objetivo de abordar esta problemática, esta investigación propone un marco conceptual para comparar portales de Infraestructuras de Datos Espaciales mediante técnicas de *Text Mining*.

Se propone llevar a cabo la identificación y relevamiento de categorías que sean semánticamente comparables de cada IDE. Luego, los datos se extraen a través del proceso consolidado de *Web Scraping* y se exploran los datos a través de un análisis de índole exploratorio con el fin de obtener un primer acercamiento a la información que se va a comparar. Una vez realizado el estudio inicial, se emplean métricas de validación interna y técnicas de *clustering*, permitiendo una evaluación objetiva de la estructura y cohesión de los *clusters* resultantes. Finalmente, todo este flujo se integra en la herramienta IDE *Comparator*, que centraliza el proceso completo y facilita la visualización y comparación de los resultados obtenidos.

Este enfoque permite comprender cómo se relacionan los datos dentro de cada IDE, facilita la comparación de su estructura interna y reduce la dificultad asociada al procesamiento de información textual. Además, el prototipo demuestra la factibilidad de automatizar parcialmente el relevamiento de información geoespacial y el marco propuesto es extensible a otras IDEs y a distintos campos.

Palabras Clave

Infraestructura de Datos Espaciales (IDE), Geoportales, *Text Mining*, *Web Scraping*, Métricas de validación interna, *Silhouette Score*, Calinski-Harabasz, *Elbow Method*, Davies-Bouldin, *Clustering*, *KMeans*, *Principal Components Analysis(PCA)*.

Índice general

Índice de tablas **18**

Índice de ilustraciones **26**

Glosario **27**

1 Introducción **31**

2 Marco teórico **33**

2.1 *Web Scraping* 33

2.1.1 Definición de *Web Scraping* 34

2.1.2 Proceso de *Web Scraping* 35

2.1.3 Aplicaciones de *Web Scraping* 36

2.2 Infraestructura de Datos Espaciales (IDE) 37

2.2.1 Definición de Infraestructura de Datos Espaciales(IDE) 38

2.2.2 Componentes principales de Infraestructura de Datos Espaciales (IDE) 38

2.2.3 Evolución de las IDEs 41

2.2.3.1 Primera Generación: IDE Centrada en Datos (década de 1990) 42

2.2.3.2 Segunda Generación: IDE Centrada en Procesos (principios de 2000) 44

2.2.3.3 Tercera Generación: IDE Centrada en el Usuario (actualidad) 44

2.3 *Text Mining* . . . 45

2.3.1 Definición de *Text Mining* . . . 46

2.3.2 Proceso de *Text Mining* . . . 46

2.3.3 *World clouds* . . . 47

2.3.4 *Clustering* . . . 48

2.3.4.1 Definición de *Clustering* . . . 48

2.3.4.2 Métodos de *Clustering* . . . 48

2.3.4.3 Métricas de validación interna . . . 50

2.4 *Principal Component Analysis(PCA)* . . . 56

2.4.1 Conceptos básicos de *Principal Component Analysis(PCA)* . . . 56

2.4.2 Aplicaciones principales de *Principal Component Analysis (PCA)* . . . 57

3 Definición del problema **59**

4 Diseño metodológico **61**

4.1 Objetivos . . . 61

4.1.1 Objetivos generales . . . 61

4.1.2 Objetivos específicos . . . 61

4.2 Metodología propuesta . . . 62

4.2.1 *Business Understanding* . . . 63

4.2.2 *Data understanding* 63
4.2.2.1 Identificación de los datos disponibles 63
4.2.2.2 Análisis exploratorio 67
4.2.3 *Data preparation* 68
4.2.4 *Modeling* 69
4.2.5 *Evaluation* 71
4.2.5.1 Análisis de métricas 71
4.2.5.2 Análisis de *Clustering* 72
4.2.6 *Deployment* 72
4.3 Arquitectura del prototipo 73
4.3.1 Descripción general del prototipo 73
4.3.2 Componentes y funciones 73
4.3.3 Flujos principales 74
4.3.3.1 Metadatos 74
4.3.3.2 Scraper 75
4.3.3.3 Análisis descriptivo 76
4.3.3.4 Análisis de métricas 77
4.3.3.5 Análisis de *clustering* 78
4.3.4 Tecnologías utilizadas 79
4.3.4.1 Base de datos 79

4.3.4.2 Docker 80

4.3.4.3 Lenguaje de programación 80

5 Resultados obtenidos **81**

5.1 Proceso de *Web Scraping* 81

5.1.1 Evolución del *scraping* de los datos 81

5.1.1.1 Enfoque de la etapa 1: *Scraping* mediante Selenium 82

5.1.1.2 Enfoque de la etapa 2: *Scraping* mediante Requests 83

5.1.1.3 Comparación de los enfoques tomados 83

5.1.2 Esquema consolidado del proceso de *Web Scraping* 84

5.2 Análisis descriptivo 84

5.2.1 IDE España 85

5.2.2 IDE Uruguay 87

5.2.3 Análisis comparativo 88

5.3 Análisis de métrica *Silhouette Score* 89

5.3.1 Análisis evolutivo de la métrica *Silhouette Score* mediante el *scraping* progresivo de los datos 90

5.3.1.1 IDE España 91

5.3.1.2 IDE Uruguay 94

5.3.1.3 Análisis comparativo 96

5.3.2 Análisis de la tasa de cambio de *Silhouette Score* 98

5.3.2.1 IDE España 99

5.3.2.2 IDE Uruguay 101

5.3.2.3 Análisis comparativo 104

5.3.3 Análisis comparativo de la estructura 105

5.3.4 Análisis comparativo del *k* óptimo 105

5.4 Análisis de métrica Davies-Bouldin 106

5.4.1 Análisis evolutivo de métrica Davies-Bouldin mediante el *scraping* progresivo de los datos 107

5.4.1.1 IDE España 107

5.4.1.2 IDE Uruguay 110

5.4.1.3 Análisis comparativo de la evolución 113

5.4.2 Análisis de la tasa de cambio de valor Davies-Bouldin 114

5.4.2.1 IDE España 115

5.4.2.2 IDE Uruguay 118

5.4.2.3 Análisis comparativo de la tasa de cambio 121

5.4.3 Análisis comparativo del índice 121

5.5 Análisis de métrica Calinski-Harabasz 122

5.5.1 Análisis evolutivo de métrica Calinski-Harabasz mediante el *scraping* progresivo de los datos 122

5.5.1.1 IDE España 123

5.5.1.2 IDE Uruguay 125

5.5.1.3 Análisis comparativo 128

5.5.2 Análisis y discusión final 128

5.6 Análisis de métrica *Elbow Method* 129

5.6.1 Análisis evolutivo de la inercia (*Elbow Method*) mediante el *scraping* progresivo de los datos 130

5.6.1.1 IDE España 130

5.6.1.2 IDE Uruguay 133

5.6.1.3 Análisis comparativo de la evolución 135

5.6.2 Análisis comparativo del *Elbow* 136

5.6.3 Análisis comparativo de la curva de la inercia 137

5.7 Análisis de *clustering* 139

5.7.1 IDE España 140

5.7.1.1 *Clustering* según la métrica *Silhouette Score* 140

5.7.1.2 *Clustering* según la métrica *Davies-Bouldin* 145

5.7.1.3 Clustering según métrica Calinski-Harabasz 148

5.7.1.4 *Clustering* según *Elbow Method* 151

5.7.2 IDE Uruguay 154

5.7.2.1 *Clustering* según *Silhouette Score* 155

5.7.2.2 *Clustering* según Davies-Bouldin,Calinski-Harabasz y *Elbow Method* . . . 158

5.7.3 Análisis comparativo . . . 161

5.7.3.1 *Silhouette Score* . . . 162

5.7.3.2 Davies-Bouldin . . . 163

5.7.3.3 Calinski-Harabasz . . . 164

5.7.3.4 Elbow Method . . . 164

5.7.3.5 Consideraciones finales . . . 165

5.8 IDE *Comparator* . . . 166

5.8.1 Objetivo . . . 166

5.8.2 Alcance . . . 166

5.8.3 Cómo IDE Comparator guía la investigación . . . 167

5.8.3.1 Sobre la fase de *Data Understanding* . . . 167

5.8.3.2 Sobre la fase de *Data Preparation* . . . 172

5.8.3.3 Sobre la fase de *Modeling+Evaluation* . . . 173

5.9 Ayuda y documentación . . . 180

6 Conclusiones **181**

6.1 Recapitulación de los objetivos . . . 181

6.1.1 Objetivo General . . . 181

6.1.2 Objetivos Específicos . . . 181

6.2 Resumen de los hallazgos 183

6.3 Contribuciones del estudio 185

7 Lecciones aprendidas **187**

7.1 Sobre investigar 187

7.2 Sobre la forma de trabajo 188

7.3 Sobre las herramientas 188

8 Futuros trabajos **190**

8.1 Sobre la fase de *Business Understanding* 190

8.2 Sobre la fase de Data *Understanding* 191

8.3 Sobre la fase de *Data Preparation* 191

8.4 Sobre la fase de *Modeling* 192

8.5 Sobre la fase de *Evaluation* 192

8.6 Sobre la fase de *Deployment* 193

8.6.1 Documentación 193

8.6.2 Prototipo 193

Bibliografia **214**

A Anexos **215**

A.1 Resultados análisis métricas 215

A.1.1 *Silhouette Score* . . . 215
A.1.1.1 IDE Uruguay . . . 215
A.1.1.2 IDE España . . . 216
A.1.2 Davies Bouldin Index . . . 216
A.1.2.1 IDE Uruguay . . . 216
A.1.2.2 IDE España . . . 217
A.1.3 Calinski-Harabasz . . . 217
A.1.3.1 IDE Uruguay . . . 217
A.1.3.2 IDE España . . . 218
A.1.4 *Elbow Method* (inercia) . . . 218
A.1.4.1 IDE Uruguay . . . 218
A.1.4.2 IDE España . . . 219
A.1.5 Cálculo derivada de la inercia . . . 220
A.1.5.1 IDE Uruguay . . . 220
A.1.5.2 IDE España . . . 221
A.2 Repositorio del proyecto . . . 222
A.3 Manuales . . . 224
A.3.1 Manual de instalación . . . 224
A.3.1.1 Prerequisitos . . . 224
A.3.1.2 Clonación del repositorio . . . 225

A.3.1.3 Ejecutar docker 225
A.3.1.4 Agregar archivo .env 225
A.3.1.5 Instalación de dependencias 225
A.3.1.6 Ejecutar la aplicación 226
A.3.2 Manual de usuario 226
A.3.2.1 Inicio 226
A.3.2.2 Metadatos 227
A.3.2.3 Scraper 229
A.3.2.4 Análisis descriptivo 231
A.3.2.5 Análisis de métricas 235
A.3.2.6 Análisis de *clustering* 241
A.3.2.7 Ayuda y documentación 242

Índice de tablas

2.1 Interpretación práctica del *Silhouette Score* 53

2.2 Interpretación cualitativa orientativa del índice Davies-Bouldin 54

2.3 Resumen de métricas de validación interna 56

5.1 Cantidad de documentos por idioma. 87

5.2 Cantidad de documentos por idioma 87

5.3 Comparación entre los metadatos de la IDE de España y Uruguay 88

5.4 Evolución del valor *Silhouette* y de k en función de la cantidad de datos en España. 92

5.5 Evolución del valor *Silhouette Score* en función de la cantidad de datos de la IDE de Uruguay 95

5.6 Derivada discreta $\Delta S/\Delta N$ para España - *Silhouette Score* 100

5.7 Derivadas discretas del *Silhouette Score* para los datos de Uruguay 103

5.8 Comparación de *Silhouette Score* para España y Uruguay usando todos los datos *scrapeados* 105

5.9 Comparación del valor de k óptimo para España y Uruguay usando todos los datos *scrapeados* 105

5.10 Evolución del índice Davies-Bouldin en función de la cantidad de documentos de la IDE de España 109

5.11 Valores de Davies-Bouldin para Uruguay en función de la cantidad de datos . . 112

5.12 Derivadas discretas del índice Davies-Bouldin para los datos de Uruguay . . . 117

5.13 Derivadas discretas aproximadas del índice Davies-Bouldin para Uruguay según la función $f(N)$ . . . 120

5.14 Comparación del número óptimo de *clusters* (k) y del valor Davies-Bouldin entre España y Uruguay utilizando todos los datos *scrapeados* . . . 121

5.15 Evolución del índice de Calinski-Harabasz para los datos de la IDE de España . 124

5.16 Evolución del índice de Calinski-Harabasz para los datos de la IDE de Uruguay 127

5.17 Comparación del número óptimo de *clusters* (k) y valor máximo del índice Calinski-Harabasz entre España y Uruguay usando todos los datos *scrapeados* . 129

5.18 Valores de inercia y k óptimo para España en función de la cantidad de documentos 132

5.19 Valores de inercia y k óptimo para España en función de la cantidad de documentos 134

A.1 Evolución del valor Silhouette Score(sin redondeo) en función de la cantidad de datos en Uruguay. . . . 215

A.2 Evolución de *Silhouette Score* y de k en función de la cantidad de datos en España 216

A.3 Evolución del índice Davies Bouldin y de k en función de la cantidad de datos en Uruguay . . . 216

A.4 Evolución del índice Davies Bouldin y de k en función de la cantidad de datos en España. . . . 217

A.5 Resultados del índice de Calinski-Harabasz para Uruguay . . . 217

A.6 Resultados del índice de Calinski-Harabasz para España, con distintos límites de documentos analizados . . . 218

A.7 Valores de inercia para Uruguay en función de la cantidad de documentos, . . . 218

A.8 Inercia para Uruguay 219

A.9 Evolución de la inercia (*Elbow Method*) en función de la cantidad de documentos en España. 219

A.10 Inercia completa para España considerando 2200 documentos. 220

Índice de ilustraciones

2.1 Proceso de *Web Scraping*. 35

2.2 Componentes de IDE 39

2.3 Las tres generaciones de las IDEs. 41

2.4 Proceso de *Text Mining*. 47

2.5 (a) Una curva visual con un *elbow* explícito.(b) Una curva visual bastante suave con un *elbow* ambiguo. 52

4.1 CRISP-DM [1] *Methodology* 62

4.2 Páginas principales de las Infraestructuras de Datos Espaciales (IDE) de España y Uruguay 64

4.3 Catálogos de metadatos de la IDE de Uruguay y España 65

4.4 Categorías exploradas en los catálogos de metadatos de las IDE de España y Uruguay. 66

4.5 Categorías de metadatos elegidas para el caso de estudios de la IDE de España y Uruguay. 67

4.6 *Pipeline* de la fase *Data preparation* 68

4.7 Diagrama de componentes 73

4.8 Diagrama de caso de uso de metadatos 75

4.9 Diagrama de caso de uso de scraper 76

4.10 Diagrama de caso de uso de análisis descriptivo 76

4.11 Diagrama de caso de uso de análisis de métricas 77

4.12 Diagrama de caso de uso de análisis de *clustering* 78

5.1 Inspección de las peticiones de red realizadas por las IDEs. 83

5.2 Nube de palabras para la categoría de metadatos de España de *Land Cover* para diferentes tamaños de conjunto de datos. 85

5.3 Nube de palabras para la categoría de metadatos de Uruguay de cobertura con mapas básicos e imágenes para diferentes tamaños de conjunto de datos. 87

5.4 Estimación de la cantidad de *clusters* (k) de la IDE de España con la métrica *Silhouette Score* variando el tamaño de la muestra (100 a 1700 datos). 91

5.5 Estimación de la cantidad de *clusters* (k) de la IDE de España con la métrica *Silhouette Score* variando el tamaño de la muestra (1900 a 2200 datos) 92

5.6 Interpretación de colores y símbolos de la Tabla 5.4 93

5.7 Estimación de la cantidad de *clusters* (k) de la IDE de Uruguay con la métrica *Silhouette Score* variando el tamaño de la muestra (100 a 7000 datos) 94

5.8 Estimación de la cantidad de *clusters* (k) de la IDE de Uruguay con la métrica *Silhouette Score* variando el tamaño de la muestra (5000 a 9391 datos) 95

5.9 Interpretación de colores y símbolos de la Tabla 5.5 96

5.10 Evolución del *Silhouette Score* para la IDE de Uruguay y España en función de la cantidad de datos 97

5.11 Evolución del k óptimo resultante de calcular la métrica *Silhouette Score* para España y Uruguay en función de la cantidad de datos. 98

5.12 Tasa de cambio del valor *Silhouette Score*-IDE España 101

5.13 Tasa de cambio del valor *Silhouettte Score* -IDE Uruguay 104

5.14 Estimación de la cantidad de *clusters* de la IDE de España con la métrica Davies Bouldin variando el tamaño de la muestra(100-1100 datos) 107

5.15 Estimación de la cantidad de *clusters* de la IDE de España con la métrica Davies Bouldin variando el tamaño de la muestra(1300-2200 datos) 108

5.16 Interpretación de colores de la Tabla 5.10 109

5.17 Interpretación de símbolos de la Tabla 5.10 109

5.18 Evolución del Davies-Bouldin Index (DBI) para diferentes volúmenes de datos de la IDE de Uruguay 110

5.19 Evolución del Davies-Bouldin Index (DBI) para diferentes volúmenes de datos de la IDE de Uruguay 111

5.20 Interpretación de colores de la Tabla 5.11 112

5.21 Interpretación de símbolos de la Tabla 5.11 112

5.22 Comparación de la evolución del índice Davies-Bouldin para Uruguay y España 114

5.23 Comparación de la evolución del número de *clusters* óptimo (k) para la IDE de Uruguay y España 114

5.24 Tasa de cambio del valor Davies-Bouldin(España) 118

5.25 Tasa de cambio del valor Davies-Bouldin(Uruguay) 120

5.26 Interpretación de colores de la Tabla 5.14 121

5.27 Estimación de la cantidad de *clusters* (k) de la IDE de España con la métrica Calinski-Harabasz variando el tamaño de la muestra(100-1700 datos) 123

5.28 Estimación de la cantidad de *clusters* (k) de la IDE de España con la métrica Calinski-Harabasz variando el tamaño de la muestra(1900-2200 datos) 124

5.29 Interpretación de símbolos de la Tabla 5.15 125

5.30 Evolución de métrica Calinski-Harabasz en función de la cantidad de datos de la IDE de Uruguay 126

5.31 Interpretación de símbolos de la Tabla 5.16 127

5.32 Evolución de la inercia (*Elbow Method*) variando el tamaño de la muestra para la IDE de España(100-1100 datos) 130

5.33 Evolución de la inercia (*Elbow Method*) variando el tamaño de la muestra para la IDE de España(1300-2200 datos) 131

5.34 Interpretación de símbolos de la Tabla 5.18 132

5.35 Evolución de la inercia (*Elbow Method*) para diferentes volúmenes de datos en Uruguay 133

5.36 Interpretación de símbolos de la Tabla 5.19 134

5.37 Comparación de la evolución del k óptimo en Uruguay y España en función de la cantidad de datos 135

5.38 Comparación de la evolución de la inercia entre Uruguay y España 136

5.39 Análisis comparativo del *Elbow* para Uruguay y España 136

5.40 Inercia para España y Uruguay en función del número de *clusters* (k) 138

5.41 $I'(k)$ de España y Uruguay 139

5.42 Clustering de la IDE de España evaluado con Silhouette Score (k=15) luego de aplicar PCA para visualización bidimensional 141

5.43 Distribución de documentos por *cluster* tras cálculo del *Silhouette Score* 142

5.44 Nube de palabras para *clusters* (k=15, resultado de calcular el k óptimo con la métrica *Silhouette Score*) - IDE España 142

5.45 Nube de palabras para *clusters* (k=15, resultado de calcular el k óptimo con la métrica *Silhouette Score*) - IDE España 143

5.46 *Clustering* de la IDE de España evaluado con Davies-Bouldin Index (k=3) . . . 146

5.47 Distribución de documentos por *cluster* tras cálculo de Davies-Bouldin 147

5.48 *Clusters* obtenidos según Davies-Bouldin Index - IDE España 147

5.49 *Clustering* de la IDE de España evaluado con Calinsky-Harabasz(k=2) 149

5.50 Distribución de documentos por *cluster* tras cálculo de Calinski-Harabasz . . . 150

5.51 Nube de palabras para *clusters* obtenidos según Calinski-Harabasz Index-IDE España 150

5.52 *Clustering* de la IDE de España evaluado con *Elbow Method*(k=6) 151

5.53 Distribución de documentos por *cluster* tras cálculo de *Elbow Method* 152

5.54 *Clusters* obtenidos según el *Elbow Method* 153

5.55 *Clustering* de la IDE de Uruguay evaluado con *Silhouette Score* (k=4) 155

5.56 Distribución de documentos por *cluster* tras cálculo de *Silhouette Score* 156

5.57 *Clusters* obtenidos según *Silhouette Score* para la IDE de Uruguay 157

5.58 *Clustering* de la IDE de Uruguay evaluado con Davies-Bouldin, Calinski-Harabasz y *Elbow Method* 158

5.59 Distribución de documentos por *cluster* tras cálculo de Davies-Bouldin, Calinski-Harabasz y *Elbow Method* 159

5.60 *Clusters* obtenidos según Davies-Bouldin, Calinski-Harabasz y *Elbow Method* para la IDE de Uruguay 160

5.61 Pantalla de visualización de metadatos - IDE Comparator 168

5.62 Visualización de *tooltips* en la tabla de metadatos - IDE Comparator. 169

5.63 Acciones de descarga de metadatos - IDE Comparator 169

5.64 Datos descargados-IDE Comparator 170

5.65 Pantallas del módulo de análisis descriptivo - IDE Comparator. 170

5.66 Resultados de análisis descriptivo individual - IDE Comparator 171

5.67 Resultados de análisis descriptivo individual en PDF - IDE Comparator 171

5.68 Resultados del análisis comparativo descriptivo - IDE Comparator. 172

5.69 Resumen en PDF del análisis comparativo - IDE Comparator 172

5.70 Pantalla del módulo Scraper - IDE Comparator. 173

5.71 Pantalla principal del módulo de análisis de métricas 174

5.72 Resultados del análisis individual de análisis de métricas-IDE Comparator . . . 175

5.73 Resumen en PDF del análisis de métricas-IDE Comparator 176

5.74 Resultados del análisis comparativo de métricas-IDE Comparator 177

5.75 Resumen en PDF del análisis comparativo de métricas-IDE Comparator. 178

5.76 Pantalla principal del módulo de análisis de *clustering* 179

5.77 Resultados del análisis de *clustering*, mostrando cantidad de documentos por *cluster* y visualización en PCA. 180

5.78 Pantalla ayuda y documentación-IDE Comparator 180

6.1 Presentación Jornadas IPGH. 185

A.1 Dependencias-IDE Comparator. 226

A.2 Pantalla principal-IDE Comparator 227

A.3 Pantalla de visualización de metadatos - IDE Comparator 228

A.4 Visualización de *tooltips* en la tabla de metadatos - IDE Comparator. 228

A.5 Acciones de descarga de metadatos - IDE Comparator. 229

A.6 Datos descargados-IDE Comparator 229

A.7 Pantalla del módulo Scraper - IDE Comparator. 230

A.8 Mensajes de *feedback* del módulo Scraper. 231

A.9 Pantallas del módulo de análisis descriptivo - IDE Comparator. 231

A.10 Feedback de análisis descriptivo individual - IDE Comparator. 232

A.11 Resultados de análisis descriptivo individual - IDE Comparator. 233

A.12 Resultados de análisis descriptivo individual en PDF - IDE Comparator. 233

A.13 Feedback durante análisis comparativo - IDE Comparator 234

A.14 Resultados del análisis comparativo - IDE Comparator. 234

A.15 Resumen en PDF del análisis comparativo - IDE Comparator. 235

A.16 Pantalla principal del módulo de análisis de métricas. 235

A.17 Feedback de procesamiento durante análisis individual. 236

A.18 Resultados del análisis individual de análisis de métricas-IDE Comparator. . . . 237

A.19 Resumen en PDF del análisis de métricas-IDE Comparator 237

A.20 Resultados del análisis comparativo de métricas-IDE Comparator. 239

A.21 Resumen en PDF del análisis comparativo de métricas-IDE Comparator. 240

A.22 Pantalla principal del módulo de análisis de clustering. 241

A.23 Resultados del análisis de *clustering*, mostrando cantidad de documentos por *cluster* y visualización en PCA. 242

A.24 Pantalla ayuda y documentación-IDE Comparator 242

Glosario

A

***API (Application Programming Interface)*:** Conjunto de reglas y protocolos que permite la comunicación entre diferentes aplicaciones o sistemas, facilitando el intercambio de datos y funcionalidades de forma estructurada.

B

Business Inteligence (BI): Conjunto de software, procesos y herramientas que analizan datos empresariales con el fin de transformarlos en información útil y apoyar la toma de decisiones.

C

Calinski-Harabasz: Índice utilizado para evaluar la calidad de un *clustering*, midiendo la relación entre la dispersión entre *clusters* y la dispersión dentro de los *clusters*. Valores más altos indican una mejor separación y cohesión de los grupos.

Clusters: Conjuntos de elementos que comparten características similares dentro de un conjunto de datos, obtenidos mediante técnicas de *clustering*.

***Clustering*:** Técnica de análisis no supervisado que organiza los datos en grupos o *clusters*, de modo que los elementos dentro de cada grupo presentan mayor similitud entre sí que con los de otros grupos.

D

Data Mining: Proceso de análisis de grandes volúmenes de datos para descubrir patrones, relaciones o tendencias relevantes, transformando datos en conocimiento útil.

Davies-Bouldin *Index* (DBI): Métrica que evalúa la calidad de *clusters*, considerando su

cohesión interna y separación entre *clusters*. Valores bajos indican *clusters* compactos y bien separados.

Davies-Bouldin: En el proyecto refiere a Index Davies Bouldin.

E

Elbow: Refiere al punto donde la mejora comienza a disminuir significativamente en *Elbow Method*.

Elbow Method: Métrica de validación interna de *clustering* para determinar el número óptimo de *clusters*, identificando el *elbow* en donde las mejoras dejan de ser significativas.

H

HTML: Acrónimo de *HyperText Markup Language*. Es el lenguaje de marcado estándar utilizado para crear y estructurar contenido en la *web*, como textos, imágenes, enlaces y otros elementos multimedia.

I

Infraestructura de Datos Espaciales (IDE): Conjunto de tecnologías, normas, políticas y recursos humanos que permiten la adquisición, procesamiento, almacenamiento, distribución y aprovechamiento eficiente de datos geográficos en la red.

Inteligencia Artificial: Conjunto de tecnologías que permiten a las computadoras realizar funciones avanzadas como percibir su entorno, comprender y procesar lenguaje natural, analizar datos, generar recomendaciones y ejecutar tareas.

J

JSON: Acrónimo de *JavaScript Object Notation*. Es un formato ligero de intercambio de datos

K

k: Número de *clusters*.

M

***Machine Learning*:** Rama de la inteligencia artificial que utiliza datos y algoritmos para permitir que los sistemas aprendan y mejoren su desempeño de manera similar a como lo hacen los humanos.

Métricas: En este documento, se consideran equivalentes a métricas de validación interna.

Métricas de validación interna: Criterios utilizados para evaluar la calidad de un *clustering* basándose únicamente en la información de los datos, sin requerir conocimiento previo de las clases.

N

N: Tamaño de la muestra.

P

PCA(Principal Component Analysis): Procedimiento matemático que realiza una reducción de dimensionalidad mediante la extracción de los componentes principales de los datos multidimensionales.

Python: Lenguaje de programación de alto nivel, claro y versátil, ampliamente usado en ciencia de datos, desarrollo de software y automatización.

S

***Silhouette Score*:** Métrica de validación que mide qué tan similares son los elementos dentro de un *cluster* en comparación con otros *clusters*. Su valor varía entre -1 y 1: valores cercanos a 1 indican buena cohesión y separación, valores cercanos a 0 indican superposición de *clusters*, y valores negativos sugieren asignaciones incorrectas.

***Stopwords*:** Palabras de alta frecuencia en los textos que no aportan información significativa para el análisis, y cuya lista varía según el idioma y la aplicación.

T

***Text Mining* :** Es el proceso de transformar datos de texto no estructurados en un formato estructurado para descubrir patrones ocultos, tendencias y nuevos conocimientos.

W

Web Scraping: Técnica para extraer datos de sitios web de forma automatizada utilizando *software*.

1 Introducción

Este proyecto de investigación se centra en el análisis comparativo de portales de Infraestructuras de Datos Espaciales. Este análisis comparativo es empleado a través de técnicas de *Text Mining*, empleando métricas de validación interna y técnicas de *clustering*.

El objetivo principal del proyecto es desarrollar un marco conceptual para comparar portales de Infraestructuras de Datos Espaciales en base a técnicas de *Text Mining*. Por tanto, los objetivos específicos para lograr este objetivo son: diseñar un marco conceptual, especificar los módulos que lo integran y sus principales características, proponer una metodología semiautomática de relevamiento de datos textuales de portales de Infraestructuras de Datos Espaciales, proponer una metodología para la comparación de portales de Infraestructuras en base a técnicas de *Text Mining*, desarrollar una prueba de concepto y construir un prototipo para validar la propuesta y luego validarlo con un caso de estudio.

El documento se organiza en 8 capítulos.

En el capítulo 2 (ver capítulo 2) se presenta el marco teórico, donde se explora el estado del arte en cuatro temas principales: *Web Scraping*, IDE (Infraestructura de Datos Espaciales), *Text Mining* y *PCA* (*Principal Component Analysis*).

En el capítulo 3 (ver capítulo 3) se define el problema. En este capítulo se destaca la necesidad de encontrar una forma de comparación adecuada ante la escasa cantidad de estudios e investigaciones previas sobre el tema en base a técnicas de *Text Mining*.

En el capítulo 4 (ver capítulo 4), se describen los objetivos generales y específicos del proyecto. En adición, se detalla la metodología adoptada en el presente proyecto y la arquitectura del prototipo en términos generales, especificando los componentes que lo integran y las tecnologías utilizadas.

El capítulo 5 (ver capítulo 5) tiene como propósito principal presentar de manera detallada

los hallazgos de la presente investigación, mostrando cómo cada etapa metodológica contribuyó a la obtención de información significativa. Se incluyen los resultados del proceso de *Web Scraping*, donde se describe el mecanismo consolidado del proceso de *Web Scraping*. Luego, se realiza un análisis descriptivo de la información obtenida que permite un primer acercamiento de los datos recopilados. Seguidamente, se exponen los resultados de las métricas de evaluación interna aplicadas, incluyendo *Silhouette Score*, Davies-Bouldin y Calinski-Harabasz, así como el análisis a través del *Elbow Method*, lo que permite indagar en el estudio de relacionamiento de los datos. Posteriormente, se presenta un análisis de *clustering* más detallado que aporta información sobre la naturaleza de los datos y las relaciones entre ellos. Finalmente, se introduce la herramienta IDE *Comparator*, la cual integra los hallazgos anteriores.

En el capítulo 6 (ver capítulo 6), se elaboran las conclusiones del estudio en función de los objetivos generales y específicos establecidos. Se proporciona un análisis de logro de cada objetivo y se presentan los principales hitos.

En el capítulo 7 (ver capítulo 7) se describen los conocimientos y lecciones aprendidas en el marco del presente proyecto de investigación.

Por último, en el capítulo 8 (ver capítulo 8) se presentan las principales líneas de investigación futuras, enmarcadas dentro de la metodología propuesta basada en el modelo CRISP-DM.

2 Marco teórico

El marco teórico de este proyecto estudia el estado del arte, según diversos autores, en cuatro áreas principales: *Web Scraping* (ver sección 2.1), IDE (Infraestructura de Datos Espaciales) (ver sección 2.2), *Text Mining* (ver sección 2.3) y *PCA* (*Principal Component Analysis*) (ver sección 2.4). El objetivo de esta sección es proporcionar una comprensión sólida de estos temas, estableciendo las bases conceptuales necesarias para el desarrollo del presente trabajo de investigación.

De esta manera, el marco teórico no solo establece un contexto conceptual, sino que también permite identificar los fundamentos metodológicos que guiarán el análisis de datos y la extracción de información, garantizando que las decisiones e interpretaciones tomadas durante la investigación se apoyen en conocimiento previo y en prácticas reconocidas en la literatura especializada.

2.1. *Web Scraping*

> *"If programming is magic, then Web Scraping is surely a form of wizardry. By writing a simple automated program, you can query web servers, request data, and parse it to extract the information you need."*
>
> – Mitchell [2]

Vivimos en una era digital, en donde la *web* se ha convertido en un vasto repositorio de información de todo tipo, desde datos científicos hasta contenidos comerciales y sociales. Extraer esta información de manera manual resulta impráctico e ineficiente, lo que hace imprescindible el uso de técnicas automatizadas como *Web Scraping*.

En esta sección se presentan los conceptos esenciales de *Web Scraping*. En el apartado 2.1.1 se ofrece su definición, en 2.1.2 se detalla el proceso de extracción de datos, y en 2.1.3 se muestran sus principales aplicaciones.

2.1.1. Definición de *Web Scraping*

Según K. Parikh et al. [3] anteriormente, los datos solo estaban disponibles en los *web browsers* y no podían ser copiados; actualmente, esto es logrado gracias a la presente técnica.

Según Mitchell [2], en la teoría, *Web Scraping* es la práctica de encontrar datos mediante un programa interactuando con una API, o un humano utilizando un *web browser*.

Esto es logrado desarrollando un programa automático que consulta a un *web server*, realiza la petición de datos y luego *parsea* estos datos para extraer la información requerida.

De acuerdo con Q. Niu et al. [4] *Web Scraping* se refiere al proceso de extraer datos de un sitio web de manera eficiente y rápida.

Según Lawson, consiste en extraer información de sitios *web* por medio de software. Este software puede copiar la forma en la que el individuo explora en los sitios *web*. [5]

Conforme a Banerjee [6], es una técnica utilizada para convertir datos *web* no estructurados en datos estructurados que pueden ser guardados y analizados en una hoja de cálculo o una base de datos. Esto habilita a que se pueda obtener grandes volúmenes de datos en una cantidad corta de tiempo, convirtiéndose en una ventaja en el mundo en que vivimos, en donde la información está siendo constantemente modificada y actualizada.

De acuerdo con K. V. Rajkumar et al. [7] es la forma de reducir el tiempo, devolviendo solo las etapas necesarias y eliminando los datos innecesarios.

A. V. Saurkar et al. [8] entienden que las técnicas de *Web Scraping* habilitan a los usuarios para que a partir de datos de varios sitios *web* puedan tenerlo o bien en una base de datos o en una hoja de cálculo. Como resultado, los datos pueden ser vistos muy rápido y analizados a posteriori.

E. J. Farley y L. Pierotte [9] exponen que *Web Scraping* incluye la creación e implementación de dos programas de software: un *crawler* y un *scraper*. El *crawler* descarga información de Internet de una manera sistemática; el *scraper* luego extrae la información importante de los

datos descargados, los estructura y luego los almacena en una base de datos o archivo de acuerdo a la estructura y formatos preestablecidos por el usuario. Este nuevo archivo es luego evaluado de una manera que los datos iniciales presentados en Internet no lo permiten.

2.1.2. Proceso de *Web Scraping*

Para Persson [10], el proceso (ilustrado en la Figura 2.1) de *Web Scraping* se divide en tres etapas.

1. **Etapa de obtención:** El sitio deseado con la información relevante debe primero ser accedido en lo que se conoce como etapa de obtención, esto es logrado utilizando el protocolo HTTP, protocolo para enviar y recibir peticiones de los servidores *webs*. Los servidores utilizan métodos similares para obtener material de páginas *webs*.

2. **Etapa de extracción:** Luego de obtener la información de la página, los datos importantes deberían ser extraídos. Herramientas tales como expresiones regulares, librerías de *HTML parsing* son usadas en esta fase.

3. **Etapa de transformación:** En este punto, la información relevante es adquirida; por tanto, puede ser convertida en un formato estructurado para presentación o almacenamiento.

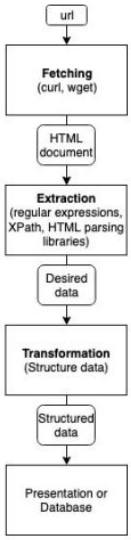


Figura 2.1: Proceso de *Web Scraping*.

Fuente: Persson [10].

2.1.3. Aplicaciones de *Web Scraping*

Web Scraping es una técnica versátil utilizada en distintos campos, tales como:

- ***Business Intelligence*:** R. Banerjee [6] señala que en el ámbito empresarial, *Web Scraping* se utiliza ampliamente para el análisis de precios, estudios de mercado y evaluación de opiniones de los consumidores (*sentiment analysis*). Por ejemplo, en el comercio electrónico, donde los precios cambian con frecuencia, monitorear manualmente a los competidores no es una solución viable. *Web Scraping* automatiza la extracción de información clave sobre precios y tendencias, consolidando estos datos en bases de datos o documentos accesibles que pueden informar estrategias comerciales en tiempo real.

- **Inteligencia Artificial (AI):** El aprendizaje automático (*Machine Learning*), una rama de la inteligencia artificial que requiere grandes volúmenes de datos para optimizar sus modelos y mejorar su precisión. En este contexto, S. D. S. Sirisuriya [11] sitúa a *Web Scraping* como una técnica esencial para recopilar información diversa y en grandes cantidades, como publicaciones en redes sociales, noticias y contenido de sitios *web*. Estos datos pueden ser procesados y utilizados para entrenar algoritmos, potenciando aplicaciones en predicción, clasificación y personalización. La capacidad de acceder a datos actualizados y variados posiciona a la técnica de *Web Scraping* como una herramienta clave para la evolución de *Machine Learning*.

- **Salud y Medicina:** Según Khan y Bide [12], *Web scraping* se ha convertido en una herramienta esencial en el sector de la salud y proporciona vastas oportunidades para el acceso y análisis de datos

- **Ciberseguridad:** Según M et al. [13], *Web Scraping* tiene aplicaciones críticas en el ámbito de seguridad y monitoreo, donde diversas organizaciones gubernamentales y privadas utilizan esta tecnología para supervisar actividades maliciosas que ocurren en internet

- **Investigación:** En el ámbito académico, C. Lotfi et al. [14] conciben que *Web Scraping*

ha demostrado ser una herramienta valiosa para analizar citas bibliográficas, recopilar publicaciones científicas, generar informes automatizados e identificar vacíos en la literatura existente. Estas aplicaciones facilitan la sistematización y organización de información relevante, lo que a su vez permite identificar tendencias emergentes y realizar análisis más exhaustivos. Además, la creación de conjuntos de datos estructurados con técnicas de *Web Scraping* optimiza el acceso y la interpretación de grandes volúmenes de datos, apoyando investigaciones en múltiples disciplinas.

En definitiva, su capacidad para automatizar la recopilación de grandes volúmenes de datos permite optimizar procesos analíticos y de toma de decisiones.

2.2. Infraestructura de Datos Espaciales (IDE)

"Vivimos en una era de la información, en la que esta información resulta esencial para afrontar los desafíos de la sociedad actual. En particular, la información espacial es uno de los elementos más críticos que sustentan la toma de decisiones en muchas disciplinas."

– Williamson et al. [15]

El crecimiento exponencial de la información digital ha impulsado el desarrollo de infraestructuras y tecnologías orientadas a su gestión eficiente. En este contexto, la información espacial adquiere un rol central al servir como base para la planificación territorial, la gestión de recursos naturales, el transporte, la seguridad y múltiples áreas de decisión estratégica.

Para dar soporte a estas necesidades surge el concepto de Infraestructura de Datos Espaciales (IDE).

En este apartado, se describe la definición de IDE según distintos autores (ver 2.2.1), los componentes principales de una IDE (ver 2.2.2) y la evolución de las IDEs (ver 2.2.3).

2.2.1. Definición de Infraestructura de Datos Espaciales(IDE)

Según Schweers et al. [16], una Infraestructura de Datos Espaciales (IDE) representa un marco integral diseñado para facilitar el intercambio y la utilización de información geográfica entre múltiples organizaciones y actores. En su esencia, una IDE está compuesta por acuerdos institucionales, políticas y tecnologías que crean un entorno propicio para el intercambio de recursos de información geográfica, fomentando en última instancia comunidades de intercambio de información.

Según Saab[17], ”la Infraestructura de Datos Espaciales (IDE) es un *framework* informativo común que proporciona una estructura de datos mediante un protocolo estandarizado y que permite a organizaciones e instituciones compartir información espacial utilizando tecnologías de sistemas de información.”

Pashova y Bandrova [18] desarrollan que la Infraestructura de Datos Espaciales (IDE) es un concepto institucional que tiene como objetivo responder de manera más efectiva a las necesidades de la sociedad de información geoespacialmente referenciada en diversos ámbitos de resolución de problemas.

2.2.2. Componentes principales de Infraestructura de Datos Espaciales (IDE)

Rajabifard y Williamson [19] identifican cinco componentes principales de una Infraestructura de Datos Espaciales (IDE): política, red de acceso, normas técnicas, personas (incluyendo asociaciones) y datos.

Estos componentes se agrupan en dos grandes temas: la interacción humano-datos (datos y personas) y las tecnologías facilitadoras (política, acceso y normas). En la Figura 2.2 se ilustran los componentes principales de la IDE.

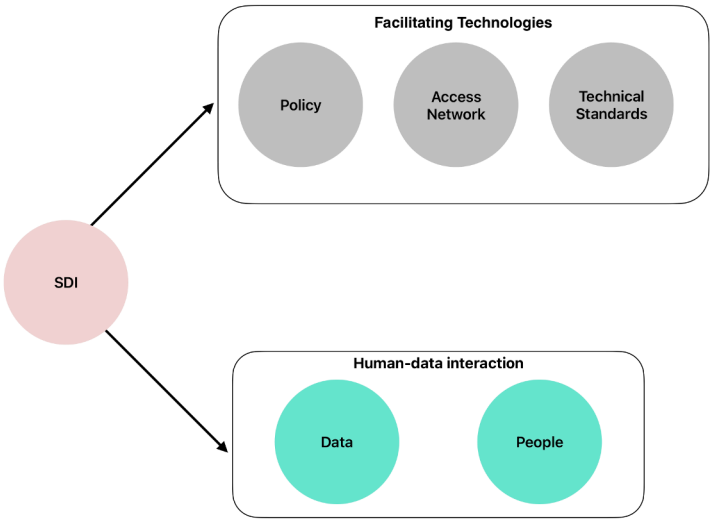


Figura 2.2: Componentes de IDE

Fuente: Elaboración propia basado en Rajabifard y Williamson[19].

A continuación, se sintetizan los principales componentes de una IDE :

- **Personas:** Según Borba[20], este componente es el foco principal de cualquier IDE moderna y consiste en actores que, de manera general, pueden ser clasificados como:
  - **Instituciones gubernamentales**: organismos del Estado a nivel nacional, regional o local, responsables de la generación, custodia y provisión de información geoespacial.
  - **Academia**: estudiantes, profesores e investigadores de universidades, institutos y centros de investigación que participan en la producción de nuevo conocimiento, el desarrollo de metodologías de análisis y la formación de profesionales capacitados en el uso y gestión de información geoespacial.
  - **Iniciativa privada**: empresas y organizaciones del sector productivo y de servicios que generan, procesan o utilizan datos espaciales para sus actividades. Su participación impulsa la innovación tecnológica, el desarrollo de soluciones

aplicadas y la integración de la información geoespacial en ámbitos como la planificación territorial, la logística o el mercado inmobiliario.

- **Sociedad**: ciudadanos y organizaciones de la sociedad civil que se benefician del acceso abierto a la información geoespacial para la toma de decisiones, la participación en procesos de gobernanza, la vigilancia ciudadana y la generación de iniciativas colaborativas que potencien el uso social de los datos.

- **Datos:** Borba [20] destaca que este componente consiste en el *framework* de datos, ya sea fundamental, temático, especial u otro que forme la base de datos de una IDE.

- **Políticas:** Borba [20] señala que este componente puede ser formal o informal, y tiene como objetivo establecer el entorno dentro del cual la IDE será desarrollada y gestionada.

- **Normas y estándares:** McLaughlin y Nichols [21] conciben que este componente define las restricciones, convenciones y metas generales y, de cierta manera, delimita los medios mediante los cuales se alcanzarán los objetivos. En esta línea Paixâo et al. [22] destacan que es el núcleo de una IDE en la búsqueda de la interoperabilidad, ya que, además de proveer interoperabilidad, evita el uso específico de una determinada tecnología o modelo. Las normas y estándares permiten el descubrimiento, intercambio, integración y usabilidad de la información espacial y tienen un impacto sobre todos los demás componentes.

- **Redes de acceso:** Borba [20] señala que este componente representa la infraestructura de *hardware* y *software* necesaria para el establecimiento de redes de comunicación y mecanismos que permitan: interoperar, buscar, consultar, integrar, acceder, proveer y utilizar los datos y metadatos geoespaciales. En definitiva, hace posible el mantenimiento, procesamiento, difusión y acceso a la información.

Borba [20] indica que estos componentes deben ser considerados en su conjunto, dado que una propuesta en uno de ellos repercute en los demás.

2.2.3. Evolución de las IDEs

Borba [20] desarrolla que en las últimas décadas, puede argumentarse que el concepto de IDE ha evolucionado a lo largo de tres generaciones, con cierta superposición entre ellas, dado que no existe una fecha exacta que marque el inicio o el fin de cada generación, sino más bien indicios.

La Figura 2.3 ilustra la tres generaciones de las IDEs.

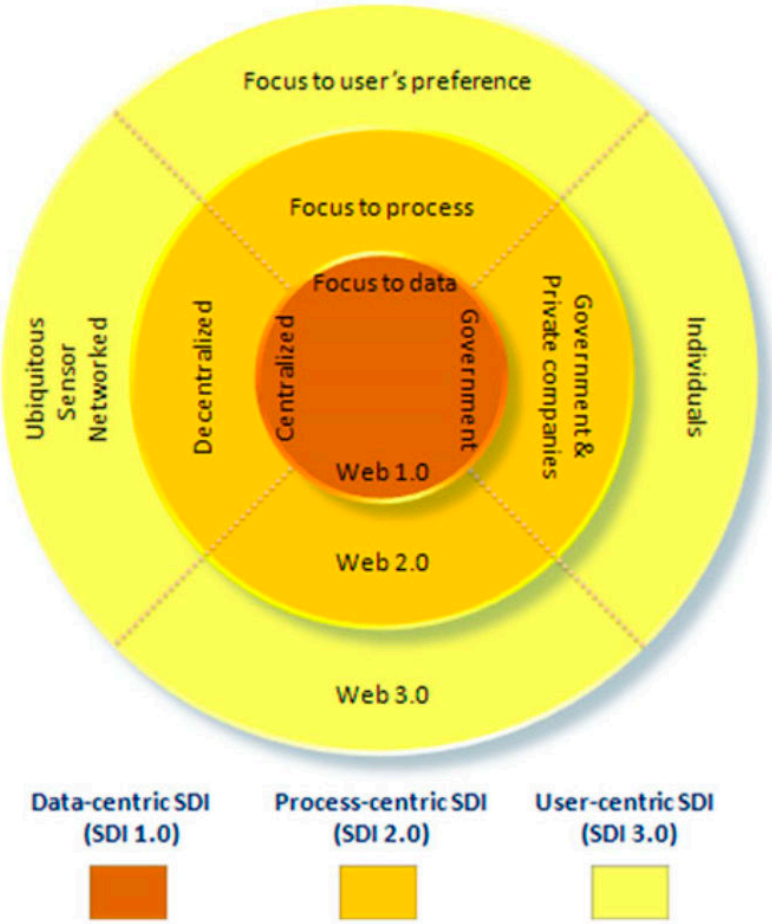


Figura 2.3: Las tres generaciones de las IDEs.

Fuente: Basaraner [23].

2.2.3.1. Primera Generación: IDE Centrada en Datos (década de 1990)

Según Williamson et al. [24], los primeros trabajos relacionados con el desarrollo de IDEs surgieron hacia la segunda mitad de los años ochenta, aunque el tema cobró notoriedad a comienzos de la década de 1990. Sin embargo, como señalan Williamson et al. [15], el conocimiento sobre los distintos conceptos, aspectos y cuestiones vinculados al tema era limitado.

Basaraner [23] explica que la IDE centrada en datos (primera generación) se enfocaba principalmente en la adquisición, modelado y compilación de datos espaciales, ya que los conjuntos de datos espaciales que cumplían con los estándares nacionales e internacionales eran escasos.

En la década de 1990, Calzati y Van Loenen [25] argumentan que las IDEs eran impulsadas por los productores, con un enfoque exclusivo en el suministro de datos georreferenciados públicos por parte de organismos nacionales. Asimismo, Basaraner [23] destaca que los conjuntos de datos producidos eran descentralizados y controlados por agencias gubernamentales individuales.

En esta primera generación, acorde a Crompvoets et al. [26], cada país diseñó y desarrolló su propia IDE en función de sus necesidades, prioridades y características específicas.

De manera general, según Masset [27], los objetivos principales de las IDEs de esta etapa fueron:

1. Promover el desarrollo económico.
2. Estimular un mejor gobierno acorde a sus características y necesidades.
3. Fomentar la sostenibilidad ambiental.

En esta generación, Crompvoets et al. [26] señalan que los datos constituían el elemento central, siendo el foco principal de las iniciativas de desarrollo de IDEs.

Coetzee y Wolff-Piggott [28] señalan que las instituciones nacionales de cartografía tenían un papel dominante en la producción de datos geoespaciales y en la creación de bases centralizadas, lo cual otorgaba al sector público un rol controlador y de mayor interés, aunque con diferencias según el país.

Entre las principales dificultades de esta generación, Rajabifard et al. [29] señalan que existieron limitaciones técnicas e institucionales. Asimismo, Borba [20] añade que destacó la ausencia de ejemplos prácticos que sirvieran como referencia para otros países. Como consecuencia, Williamson et al. [15] señalan que uno de los principales resultados fue la producción de documentación y experiencias sobre iniciativas IDE, en especial, una aproximación orientada a datos.

Si bien las IDEs de primera generación fueron concebidas con el objetivo de diseminar datos, Coleman y McLaughlin [30] argumentan que se descuidaron cuestiones de interoperabilidad, así como las preferencias de los usuarios y la interacción entre proveedores y consumidores de datos geoespaciales.

Craglia y Annoni [31] sostienen que el modelo se caracterizó por la difusión de datos y metadatos a través de servicios adaptados a formatos específicos, lo que derivó en críticas: eran proyectos basados en la creación de bases de datos. Masser [32] agrega que no siempre contemplaban cuestiones de acceso debido a limitaciones técnicas e institucionales. Asimismo, este modelo se enfocaba en la interacción humana, en esta línea Cömert [33] destaca que esto provocó un descuido en la interacción entre sistemas computacionales. Otro aspecto señalado por Masser [27] fueron las diferencias específicas de cada país: extensión, alcance, políticas, sistemas de gobierno, grado de participación estatal y privada, objetivos territoriales, aspectos culturales y condiciones socioeconómicas.

Finalmente, Borba[20] destaca que la dimensión tecnológica (*hardware*, *software* e infraestructura de telecomunicaciones) no estaba suficientemente madura. Las primeras implementaciones asumían la existencia de una infraestructura de telecomunicaciones capaz de soportar grandes volúmenes de datos. Sin embargo, tanto la madurez como la disponibilidad de software geográfico eran limitadas y dependían de plataformas específicas,

lo que dificultaba la interoperabilidad entre instituciones.

2.2.3.2. Segunda Generación: IDE Centrada en Procesos (principios de 2000)

Williamson et al. [15] indican que se inició alrededor del año 2000, aprovechando la experiencia, los aprendizajes y la documentación de la primera generación (ver 2.2.3.1).

Grus et al. [34] señalan que la segunda generación se caracterizó por el establecimiento de vínculos más activos entre las personas, los datos y la coordinación institucional, con una visión de desarrollo a largo plazo.

En esta etapa, Crompvoets et al. [26] destacan que el foco se trasladó hacia el uso de los datos y las necesidades de los usuarios. Masser[35] desarrolla que el desarrollo de IDEs comenzó a integrarse con nuevas tecnologías *web* y Vandenbroucke y Janssen [36] señalan la integración con iniciativas de gobierno electrónico. De este modo, Layne y Lee [37] y Warnest et al. [38] argumentan que las IDEs pasaron a apoyar iniciativas de gobierno en línea, facilitando la provisión de información y servicios a través de Internet.

Rajabifard et al. [29] refieren que el modelo orientado a procesos se apoyó en la creación de arreglos institucionales que mejoraran la comunicación con la comunidad usuaria y fomentaran el intercambio, la reutilización y el compartir recursos.

Asimismo, Craglia y Annoni [31] destacan que fue influenciado por una mejor comprensión de la naturaleza de las IDEs, el surgimiento de arquitecturas descentralizadas y el creciente papel del sector privado.

2.2.3.3. Tercera Generación: IDE Centrada en el Usuario (actualidad)

Emergiendo gradualmente desde mediados de la década de 2000, Budhathoki et al. [39] argumentan que la tercera generación se fundamenta en los aprendizajes previos, pero otorga un papel protagónico a los usuarios.

Williamson et al. [40] destacan que la información geoespacial se concibe como un bien común disponible para toda la sociedad, fomentando la creatividad, la colaboración y el desarrollo de productos.

En este modelo, según Coleman et al. [41], y Hennig y Belgui [42] los usuarios adoptan roles tanto de consumidores como de productores de información. Este enfoque se vincula al concepto de sociedades habilitadas espacialmente, en las que, como señalan Enemark y Rajabifard [43], la información geoespacial influye en la gobernanza y en la organización social.

Hennig et al. [44] señalan que el gobierno pasa de ser el principal productor a actuar como facilitador, promoviendo la participación de actores privados, comunidades auto-organizadas e iniciativas abiertas.

En síntesis, según Borba[20] la tercera generación se caracteriza por ser multiescalar, orientada al usuario, integradora de iniciativas públicas, privadas y de colaboración abierta, e interconectada con otras IDEs.

2.3. *Text Mining*

> *"The information age has made it easy to store large amounts of data. The proliferation of documents available on the Web, on corporate intranets, on news wires, and elsewhere is overwhelming. However, although the amount of data available to us is constantly increasing, our ability to absorb and process this information remains constant...Text mining is a new and exciting research area that tries to solve the information overload problem..."*
>
> – Feldman y Sanger [45]

El creciente volumen de datos de texto no estructurado fruto del creciente información hace que *Text Mining* tome relevancia.

En esta sección, se definen nociones esenciales de *Text Mining*. En el apartado 2.3.1 se presenta la definición de *Text Mining*, en 2.3.2 se presenta el proceso de *Text Mining*, en 2.3.4

se presenta *Clustering*.

2.3.1. Definición de *Text Mining*

Según Abdusalomovna et al. [46], *Text Mining* "es el proceso de examinar grandes conjuntos de documentos para descubrir nueva información o ayudar a responder preguntas de investigación específicas. Identifica hechos, relaciones y afirmaciones restantes".

En términos de Altman y Krzywinski [47], *Text Mining* representa "el descubrimiento por computadora de información nueva y previamente desconocida mediante la extracción automática de información de diferentes recursos escritos".

2.3.2. Proceso de *Text Mining*

Según Chen et al. [48] el proceso general de *Text Mining* incluye :

1. **Adquisición de información de texto:** El texto puede ser obtenido a través de lectura de archivos, *Web Scraping* (ver sección 2.1) o cualquier otro medio que permita obtener una colección de datos.

2. **Preprocesamiento de texto:** Cada documento es preprocesado. Más específicamente, los datos se convierten al formato deseado y se procesan para eliminar contenido no útil para la tarea en cuestión (por ejemplo, hipervínculos, formas no estándar de palabras, redundancias y *stopwords*). Según Medhat et al. [49], las *stopwords* se definen como palabras comunes que, por lo general, no aportan al significado de una oración, específicamente para fines de recuperación de información. Rivas et al. [50] mencionan que estas palabras se caracterizan por su alta frecuencia de aparición en los documentos sin aportar información significativa. Ejemplos de *stopwords* incluyen '*and*', '*the*' y '*a*', pero no existe una lista definitiva, ya que varía según el idioma y la aplicación.

3. **Representación de texto:** Los documentos se transforman de su versión en texto completo a un modelo de espacio vectorial, el cual representa los diferentes conjuntos de características lingüísticas presentes en cada documento.

4. **Análisis:** Se analiza utilizando técnicas de clasificación automática o *clustering* (ver subsección 2.3.4), que también son empleadas en *Data Mining*.

5. **Evaluación del resultado:** El *output* del paso anterior se evalúa y puede almacenarse o utilizarse en una serie de experimentos posteriores de minería de texto.

En la Figura 2.4 se ilustra el proceso de *Text Mining*.

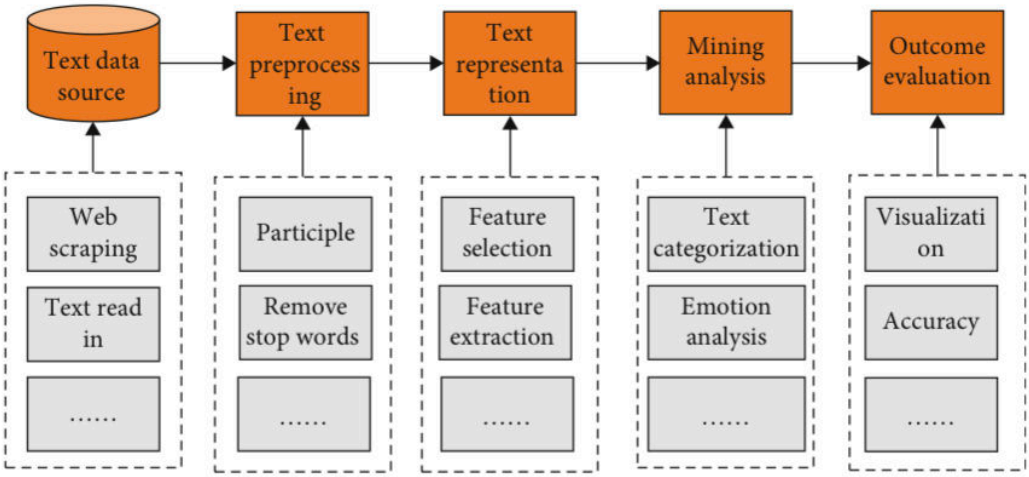


Figura 2.4: Proceso de *Text Mining*.

Fuente: Chen et al. [48].

2.3.3. *World clouds*

Según Baroukh et al. [51] *World clouds*, también conocidas como *tag clouds*, son representaciones visuales de datos textuales en las que las palabras individuales se muestran con diferentes tamaños, colores y posiciones en función de su frecuencia de aparición dentro de un corpus determinado. Baroukh et al. [51] agregan que estas visualizaciones surgieron como una solución basada en la web para resumir rápidamente un texto, maximizando la representación de los términos más relevantes en un espacio reducido.

Schoier et al. [52] destacan que en las visualizaciones de *World clouds*, las palabras que ocurren con mayor frecuencia suelen ocupar posiciones más destacadas y mostrarse con tamaños

de fuente más grandes, mientras que los términos menos frecuentes aparecen más pequeños y en ubicaciones periféricas.

2.3.4. *Clustering*

En la presente subsección se definen conceptos relevantes de *Clustering*. En 2.3.4.1 se presenta la definición de *clustering*, en 2.3.4.2 se presentan métodos de *Clustering*, en 2.3.4.3 se presentan distintas métricas de validación interna relevantes a la investigación.

2.3.4.1. Definición de *Clustering*

Según Jain et al. [53], Liao [54], Bose y Chen [55], Grant y Cheo [56], Samoilenko y Osei-Bryson [57], Xie et al. [58] *clustering* se define como la agrupación de objetos donde existe poca o ninguna información sobre la relación entre ellos en el conjunto de datos.

Kaufman y Rousseeuw [59] y Xu y Wunsch [60] destacan que el *clustering* puede ser una herramienta independiente para analizar la distribución de los datos, observar características de cada *cluster* y centrarse en *clusters* específicos para análisis detallados.

Oyewole y Thopil [61] señalan que el *clustering* busca determinar las clases presentes en la información, agrupando datos no etiquetados con poca o ninguna supervisión. Los objetos dentro de una clase tienen características similares y se diferencian de los objetos de otras clases.

Schwenker y Trentin [62] describen el *clustering* como un aspecto de *Machine Learning* donde los algoritmos extraen patrones de datos obtenidos de observaciones directas o simuladas.

2.3.4.2. Métodos de *Clustering*

El extenso desarrollo en el área de *clustering* ha dado lugar a diversas formas de categorizar la gran cantidad de algoritmos disponibles.

Según Härdle y Simar [63] et al., la distinción más relevante separa los enfoques de *clustering*

en dos grupos principales: algoritmos jerárquicos y algoritmos de particionamiento.

Algoritmos de particionamiento

Según Braun et al. [64], los métodos de particionamiento parten de una configuración de grupos determinada y proceden intercambiando elementos de datos entre los grupos con el fin de optimizar el *clustering*.

A continuación, se enumeran y describen algunos de los algoritmos de particionamiento:

- **K-means:** K-Means es un algoritmo de *clustering* ampliamente utilizado. Fue propuesto por primera vez por Stuart Lloyd en 1957 [65] como una técnica de cuantización vectorial. Sin embargo, la versión moderna del algoritmo fue desarrollada por MacQueen en 1967 [66].

  Según Shutaywi y Kachouie [67], K-means busca minimizar la distancia euclidiana entre cada punto y el centro de su cluster asignado.

  Nkweteyim [68] subraya que cada centro de clúster está representado por el valor medio de los objetos en el *cluster*.

  Por último, Arthur y Vassilvitskii [69] destacan su simplicidad y velocidad.

- **KMeans++:** Según Arthur et al. [70] KMeans++ se trata de una versión refinada de KMeans que selecciona los centros de *cluster* para asegurar que los *clusters* iniciales estén bien distribuidos.

- **K-medoids:** K-medoids, de acuerdo a Nkweteyim [68] utiliza una técnica basada en objetos representativos, donde cada centro de *cluster* está representado por uno de los objetos reales de los datos en lugar de una media calculada.

Algoritmos jerárquicos

Braun et al. [64] destacan que los métodos jerárquicos construyen los *clusteres* mediante la agregación o división sucesiva de los datos, ya sea siguiendo un procedimiento divisivo que comienza con un único *cluster* que contiene todos los datos, o bien un procedimiento

aglomerativo que inicia con cada punto de datos como un *cluster* independiente y va fusionando iterativamente los pares.

A continuación, se enumeran y describen algunos de los algoritmos jerárquicos:

- ***Agglomerative Clustering*:** Cortés et al. [71] argumentan que *Agglomerative Clustering* utiliza un enfoque ascendente donde cada punto de datos comienza como su propio *cluster* y se van fusionando de manera iterativa los *clusters* más similares.
- ***Divisive Analysis (DIANA)*:** Según Kaufman y Rousseeuw [59], *Divisive Analysis* es un método jerárquico divisivo que divide iterativamente los *clusters* en dos más pequeños hasta que se genera el número deseado de *clusters* o hasta que cada *cluster* contenga una sola observación.

2.3.4.3. Métricas de validación interna

Según Inyang et al., Ullah et al. [72, 73], la validación de *clusters* evalúa la calidad de las soluciones de *clustering*, determinando el número de *clusters* que mejor representa la estructura de los datos sin conocimiento previo de las clases. A diferencia de la evaluación externa, que requiere conocer las asignaciones verdaderas, los criterios internos se basan únicamente en la información de los datos.

Hennig [74], Milligan y Cooper [75] señalan que encontrar el número correcto de *clusters* compactos y bien separados es uno de los problemas más desafiantes del análisis de *clustering*.

Bagirov et al. [76] indican que las funciones objetivo suelen formularse para maximizar la cohesión interna y la separación entre *clusters* distintos. Saputra et al. [77] explican que la cohesión evalúa cuán cercanos están los elementos dentro de un *cluster* y la separación cuán distintos o distantes son los *clusters* entre sí.

Zobaed et al. [78] enfatizan que comprender lo que cada métrica prioriza es esencial para seleccionar el enfoque adecuado según los objetivos y características de los datos.

Se recomienda utilizar varias métricas simultáneamente, ya que distintos criterios pueden dar diferentes números óptimos de *clusters* o evaluaciones de calidad.

En el caso específico de K-means, la elección del valor de k –es decir, el número de *clusters*– resulta crítica, ya que según argumentan Topal y Geçer [79] determina directamente la partición de los datos. Seleccionar un k demasiado bajo puede agrupar datos heterogéneos en un mismo *cluster*, mientras que un k excesivo puede fragmentar innecesariamente la estructura de los datos. Por esta razón, las métricas de validación interna no solo evalúan la calidad de la partición, sino que también ayudan a determinar el número óptimo de *clusters* que mejor representa la estructura subyacente.

Algunos métodos para optimizar clustering incluyen el *Elbow Method*, *Silhouette Score*, *Davies-Bouldin Index* (DBI) y Calinski–Harabasz.

Elbow Method

Muharram et al. [80] señalan que *Elbow Method* determina el número óptimo de *clusters* mediante la disminución de la inercia .

Saputra et al. [77] explican que se calcula la suma de cuadrados *Within-Cluster Sum of Squares*(WSS) para distintos valores de k y se grafica. A medida que k aumenta, WSS disminuye, pero la mejora se vuelve marginal a partir de cierto punto; el *elbow* indica el número óptimo de *clusters* como se muestra en la Figura 2.5 (a) con un *elbow* explícito. Mientras tanto, el número óptimo de *clusters* correspondiente al *elbow* depende de la selección realizada manualmente. Sin embargo, existe un problema con el *Elbow Method*, y es que el *elbow* no puede ser distinguido de manera inequívoca por los analistas experimentados cuando la curva representada es bastante suave, como se muestra en la Figura 2.5 (b) con un *elbow* ambiguo.

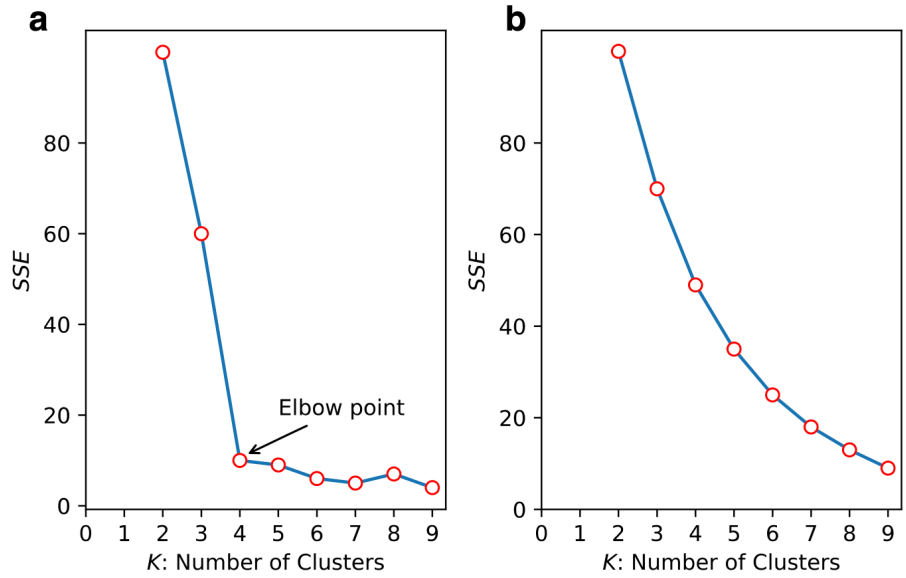


Figura 2.5: (a) Una curva visual con un *elbow* explícito.(b) Una curva visual bastante suave con un *elbow* ambiguo.

Fuente: Shi et al. [81].

En términos generales, *Elbow Method* encuentra el balance entre complejidad del modelo y rendimiento del *clustering*, determinando el número adecuado de *clusters*. Se utiliza como complemento de las otras métricas.

Silhouette Score

Según Rousseeuw [82], el *Silhouette Score* es un método ampliamente usado para interpretar y validar la consistencia de los resultados de *clustering*.

Silhouette Score mide qué tan similar es un punto a su propio *cluster* frente a otros *clusters*.

Villatoro-Tello et al. [83] y Zobaed [78] señalan que evalúa la cohesión intra-cluster y la separación inter-cluster, midiendo qué tan bien se ajusta cada punto a su *cluster*.

Awong y Zielinska [84] destacan que permite identificar puntos mal *clusterizados* y evaluar la calidad a nivel de *cluster* y punto. Destaca por su enfoque particular y precisión en la evaluación individual de los *clusters*.

Según Muharram et al. [80], un valor alto de *Silhouette* indica *clusters* bien definidos y

permite identificar *clusters* óptimos.

De acuerdo con Struyf et al. [85], a partir de la experiencia adquirida, el *Silhouette Score* puede organizarse en las categorías que se presentan en la Tabla 2.1.

Rango de *Silhouette Score*	**Interpretación de la estructura de los *clusters***
0.71 - 1.00	Estructura de *cluster* fuerte
0.51 - 0.70	Estructura de *cluster* razonable
0.26 - 0.50	Estructura de *cluster* débil
$\leq$ 0.25	Datos sin estructura

Tabla 2.1: Interpretación práctica del *Silhouette Score*

Fuente: Elaboración propia basado en Struyf et al. [85]

Davies-Bouldin

El índice Davies-Bouldin fue desarrollado por David L. Davies y Donald W. Bouldin en 1979 [86] como una métrica para evaluar algoritmos de *clustering*.

Punhani et al. [87] consideran que este índice funciona como una medida de validez de *clusters* mediante una evaluación interna, utilizando únicamente las cantidades y características inherentes de la base de datos sin requerir información externa.

El fundamento teórico del índice se basa, según Ansari et al. [88] y Yao et al. [89], en la comparación entre dos componentes clave: la dispersión intra-cluster y la separación inter-cluster.

Chen et al. [90], Davies et al. [86] y Flann et al. [91] señalan que la métrica evalúa la compacidad o cohesión dentro de cada *cluster*, es decir, qué tan cerca están los puntos de datos del centroide de su grupo, frente a qué tan bien separados están los diferentes *clusters* entre sí.

Yao et al. [89] entienden que geométricamente, el objetivo del índice es minimizar la dispersión intra-*cluster* (numerador de la fórmula) mientras se maximiza la separación entre clases (denominador).

Según Davies et al. [86] la interpretación del índice Davies-Bouldin se basa en un principio fundamental: valores más bajos indican mejor calidad de *clustering*.

Febrita et al. [92], Raptis et al. [93], Oliveira y Marçal [94]; Thomas et al. [95]; Davies et al. [86] y Suh [96] explican que el índice varía desde cero hasta infinito positivo, donde un valor de cero representa el óptimo, indicando un *clustering* perfecto .

Según Nave et al. [97], Renjith et al. [98], Aljeri [99] y Davies et al. [86] cuando el índice Davies-Bouldin presenta valores cercanos a cero, esto indica que los *clusters* son compactos internamente (los puntos están cerca de sus centroides) y bien separados entre sí (las distancias entre centroides son grandes),

Suraya et al. [100], Lorencin et al. [101] Hatamikia et al. [102] mencionan que por el contrario, valores altos del índice señalan problemas en la estructura de *clustering*, como *clusters* con superposición significativa, poca separación entre grupos o alta dispersión interna. Febrita et al. [92] especifican que un valor de uno o superior generalmente indica una distribución de datos muy deficiente en los *clusters*.

En base a la literatura revisada y a modo de guía para interpretar los resultados obtenidos, se propone el siguiente esquema ilustrado en la Tabla 2.2.

Valor DBI	**Interpretación práctica**
< 1	Bueno: *clusters* compactos y bien separados
≥ 1	Distribución de datos deficiente: *clusters* dispersos o con superposición significativa

Tabla 2.2: Interpretación cualitativa orientativa del índice Davies-Bouldin

Fuente: Elaboración propia basada en Febrita et al. [92], Raptis et al. [93], Oliveira y Marçal [94], Thomas et al. [95], Davies et al. [86] y Suh [96].

Según Liu [103], You y Rumbe [104] para la selección del número óptimo de *clusters*, se debe elegir aquel valor de k que produzca el menor índice Davies-Bouldin. Sin embargo, tal como plantean Sublime et al. [105] es importante considerar que el índice tiende a favorecer soluciones con menor número de *clusters*, por lo que su interpretación debe complementarse con otros criterios de validación y el conocimiento del dominio específico.

Según Sublime et al. [105] y Yin et al. [106], el índice Davies-Bouldin presenta un sesgo hacia un menor número de *clusters*, lo que puede llevar a soluciones subóptimas cuando el número real de grupos es mayor. Según Duan y Zou [107], este índice es adecuado para evaluar

conjuntos de datos caracterizados por alta compacidad intra-*cluster* y gran separación inter-*cluster*; sin embargo, cuando el grado de superposición inter-*cluster* es elevado resulta muy difícil realizar una evaluación precisa del *clustering* utilizando esta métrica.

Calinski–Harabasz

Según Chikumbo y Granville [108], Bianco et al. [109] y Lorencin et al. [101], el índice Calinski–Harabasz (CH), también conocido como Criterio de Razón de Varianza (VRC), es una herramienta fundamental para evaluar la calidad del *clustering* en aprendizaje no supervisado.

Villatoro-Tello et al. [83] y Han et al. [110] argumentan que maximiza la varianza entre *clusters* respecto a la varianza interna, evaluando compacidad y separación simultáneamente.

Según Zhai et al. [111] y Kallel et al . [112] funciona mejor con *clusters* compactos y bien separados, útil para detectar separación clara y evaluar *clusters* con densidades variables .

Karim et al. [113] y Zouinina et al. [114] indican que el índice Calinski-Harabasz (CH) presenta desafíos particulares de interpretación debido a su naturaleza no acotada, con valores que van desde cero hasta infinito positivo dependiendo de las características del conjunto de datos. Zouinina et al. [114] señalan además que el índice depende fuertemente del tamaño del *dataset* (N) y escala linealmente con el número de puntos de datos, lo que significa que su orden de magnitud puede variar considerablemente entre conjuntos de datos. Esto hace que la comparación directa de valores de CH entre diferentes conjuntos de datos sea problemática, aunque como señalan algunos investigadores como es el caso de la investigación relizada por Karim et al. [113] normalizan los resultados a rangos como [0,1] para facilitar análisis comparativos.

Resumen de métricas de validación interna

En la Tabla 2.3 se presenta un resumen de las métricas de validación interna.

Métrica	Descripción / Criterio
Silhouette Score	Evalúa la precisión en la ubicación de puntos individuales dentro de su *cluster*, considerando cohesión y separación.
Davies–Bouldin Index	Minimizar solapamiento y dispersión; valores más bajos indican menor solapamiento y mejor separación.
Calinski–Harabasz Index	Busca maximizar la separación entre *clusters* en relación a su compacidad interna. Valores más altos indican mejor definición.
Elbow Method	Método gráfico que ayuda a determinar el número óptimo de *clusters* observando el punto de inflexión en la curva de inercia.

Tabla 2.3: Resumen de métricas de validación interna

Fuente: Elaboración propia.

2.4. *Principal Component Analysis(PCA)*

En la presente sección se desarrollan los conceptos básicos de *Principal Component Analysis(PCA)* (ver 2.4.1) y algunas de sus aplicaciones principales (ver 2.4.2).

2.4.1. Conceptos básicos de *Principal Component Analysis(PCA)*

Principal Component Analysis(PCA) fue introducido originalmente por Pearson [115] en 1901. Shen et al. [116] subrayan que fue introducida para simplificar la complejidad de los datos de alta dimensión, manteniendo al mismo tiempo su varianza.

Según V. et al. [117], *Principal Component Analysis(PCA)* es un procedimiento matemático que realiza una reducción de dimensionalidad mediante la extracción de los componentes principales de los datos multidimensionales.

Peng et al. [118] sostienen que esta técnica busca extraer la información más relevante de los datos y expresarla como un conjunto de nuevas variables ortogonales llamadas componentes principales, especialmente cuando las variables independientes están correlacionadas.

De acuerdo con V. et al. [117], el método funciona aplicando una transformación ortogonal que convierte un conjunto de observaciones de variables posiblemente correlacionadas en un

conjunto de valores de variables no correlacionadas, llamadas componentes principales.

Según Uzga y Rebrovs [119], cada componente principal es una combinación lineal de las variables originales. Asimismo, V. et al. [117] y Eslami et al. [120] explican que el primer componente presenta la mayor variabilidad, mientras que cada componente sucesivo alcanza la mayor varianza posible bajo la restricción de ser ortogonal a los anteriores.

2.4.2. Aplicaciones principales de *Principal Component Analysis (PCA)*

PCA se ha consolidado como una técnica versátil con aplicaciones en numerosos ámbitos:

- **Computer Vision and Image Processing:** Kambo et al. [121] señalan que PCA se utiliza ampliamente en reconocimiento y compresión de imágenes, reduciendo la alta dimensionalidad de los datos de imagen mientras se preservan las características visuales esenciales. Kshirsagar et al. [122] destacan aplicaciones específicas como la detección de rostros en sistemas de computer vision. Además, Xuan et al. [123] y Schneider et al. [124] señalan que esta técnica se aplica ampliamente en procesamiento de imágenes y señales, reconocimiento de patrones y compresión de datos.
- ***Bioinformatics and Life Sciences*:** PCA es muy útil en el análisis de datos de expresión génica, ayudando a los investigadores a identificar patrones en conjuntos de datos biológicos complejos, como es el caso de la investigación desarrollada por Yeung et al. [125]. Sadhukhan y Yadav [126] destacan su valor particular en bioinformática para el manejo de datos genómicos de alta dimensionalidad.
- ***Financial Analysis*:** PCA se aplica en finanzas para el análisis de carteras, evaluación de riesgos y predicción de precios de acciones. PCA se aplica en finanzas para el análisis de carteras, evaluación de riesgos y predicción de precios de acciones. Por ejemplo, Ghorbani y Chong [127] utilizaron PCA para la predicción de precios de acciones, Zhong y Enke [128] aplicaron reducción de dimensionalidad para pronosticar retornos diarios del mercado, Yu et al. [129] implementaron un modelo de selección de acciones basado en SVM dentro de PCA, y Pasini [130] empleó PCA para la gestión de carteras de inversión.

- **Data Mining and Machine Learning:** Nagpal et al. [131] señalan que PCA es una técnica poderosa empleada en *Data Mining*, utilizada como preprocesamiento de conjuntos de datos de alta dimensionalidad antes de aplicar algoritmos de *Machine Learning*. También se emplea como método de extracción de características y es común en el análisis de series temporales (Sadhukhan et al., 2023).

- ***Exploratory Data Analysis and Visualization*:** Sadhukhan y Yadav [126] destacan que PCA es especialmente útil para la visualización de datos, permitiendo reducir datos de alta dimensionalidad a dos o tres dimensiones para facilitar su interpretación. Además, Xuan et al. [123] y Omogbai [132] agregan que funciona como herramienta de análisis exploratorio de datos, ayudando a comprender la verdadera dimensionalidad de los datos y la contribución de cada variable.

- **General Statistical Analysis:** Ezekiel et al. [133] mencionan que PCA se utiliza para la extracción de características y reducción de dimensionalidad de grandes matrices en diversas aplicaciones de análisis multivariante. Nagpal et al. [131] agregan que la técnica ayuda a filtrar ruido de fondo y resaltar patrones esenciales, siendo efectiva para simplificar análisis multivariantes complejos.

3 Definición del problema

El presente trabajo se centra en el análisis comparativo de los portales de las Infraestructuras de Datos Espaciales (IDE), con un enfoque en la aplicación de técnicas de *Web Scraping* y *Text Mining* para sistematizar el proceso comparativo y generar nuevos estudios en el área.

Las Infraestructuras de Datos Espaciales (IDE) comprenden el conjunto de tecnologías, políticas, estándares y recursos humanos necesarios para adquirir, procesar, almacenar, distribuir y optimizar el uso de datos geoespaciales. Según Harvey, Iwaniak, Coetzee et al. [134], una IDE es un concepto evolutivo orientado a facilitar y coordinar el intercambio y compartición de datos espaciales entre diversas partes interesadas a diferentes niveles dentro de la comunidad de datos espaciales.

Diversos estudios han propuesto enfoques para comparar IDEs, como los trabajos de Mulder, Wiersma y Van Loenen [135], quienes comparan IDEs a nivel continental considerando dimensiones de descubrimiento, acceso y propiedades de los datos; y Trystuła, Dudzińska y Źróbek [136], quienes analizan la completitud de las IDEs en el contexto de datos catastrales.

Los datos textuales, por su naturaleza no estructurada, presentan mayores desafíos de procesamiento que los datos cuantitativos. Con el objetivo de identificar patrones relevantes en dichos datos, el *Text Mining* se entiende como la aplicación de técnicas de *Data Mining* a datos textuales. Negi [137] compara *Text Mining* y *Data Mining*, reseñando sus principales áreas de aplicación. Profundizando en estas técnicas, Samiyeva y Madyarova [138] distinguen entre *Text Mining*, centrado en identificar información relevante dentro de un texto, y *Text Analytics*, orientado a detectar patrones y tendencias en datos textuales.

Algunas investigaciones han aplicado *Text Mining* al análisis de IDE, como Kaczmarek, Iwaniak y Świetlicka et al. [139], quienes emplean procesamiento de lenguaje natural para estudiar planes de uso del suelo, y Baptista y Figueiras [140], que comparan sitios *web* con

información geográfica basada en narrativas textuales.

En cuanto a la obtención de datos desde sitios *web*, *Web Scraping* se define como la extracción automatizada de información mediante software especializado, como señalan Khder[141] y Sirisuriyab [11], destacando su relevancia como fuente de datos para algoritmos de identificación de patrones.

A pesar de los avances existentes, la investigación centrada en la sistematización del análisis y comparación de IDEs mediante técnicas de *Text Mining* sigue siendo limitada, lo que constituye la motivación principal de este estudio. En este contexto, se identifica como problemática la dificultad de realizar comparaciones exhaustivas de manera manual, especialmente considerando que la información geoespacial ha crecido exponencialmente desde la aparición de las IDEs y que cada infraestructura contiene datos valiosos y específicos. Dado que las IDEs son un concepto en constante evolución, es fundamental desarrollar estrategias de análisis que permitan analizar y comparar sus contenidos de manera eficiente, garantizando que el estudio se mantenga actualizado y refleje el estado del arte de la información geoespacial.

Para abordar esta problemática, se propone la aplicación combinada de *Text Mining* y *Web Scraping*, lo que permite la obtención y el procesamiento semiautomático de los datos provenientes de las IDEs. Mientras que el *Text Mining* facilita la detección de patrones y relaciones relevantes en los datos, el *Web Scraping* asegura la recolección sistemática y actualizada de la información, evitando la dependencia de procedimientos manuales y estructuras rígidas.

El estudio busca identificar y extraer datos clave mediante *Web Scraping* y *Text Mining*. En particular, se aplican herramientas de *clustering* junto con métricas de validación interna para analizar la naturaleza de los datos, su grado de similitud y las relaciones existentes entre ellos con el fin de sistematizar el análisis en un contexto comparativo. Este enfoque permite generar conocimiento significativo en un contexto de información en constante expansión, donde mantenerse a la vanguardia y optimizar los procesos resulta fundamental para impulsar el desarrollo tecnológico.

4 Diseño metodológico

En el presente capítulo se presentan los objetivos del proyecto (ver 4.1), la metodología propuesta (ver 4.2) y la arquitectura en términos generales del prototipo (ver 4.3).

4.1. Objetivos

En esta sección se define el objetivo general (ver 4.1.1) y los objetivos específicos (ver 4.1.2) del proyecto.

4.1.1. Objetivos generales

El objetivo principal del proyecto es desarrollar un marco conceptual para comparar portales de Infraestructuras de Datos Espaciales en base a técnicas de *Text Mining*.

4.1.2. Objetivos específicos

OE1: Diseñar un marco conceptual, especificar los módulos que lo integran y sus principales características.

OE2: Proponer una metodología semiautomática de relevamiento de datos textuales de portales de Infraestructuras de Datos Espaciales.

OE3: Proponer una metodología para la comparación de portales de Infraestructuras en base a técnicas de *Text Mining*.

OE4: Desarrollar una prueba de concepto y construir un prototipo para validar la propuesta.

OE5: Validar el prototipo con un caso de estudio.

4.2. Metodología propuesta

El marco metodológico adoptado en este proyecto se basa en el estándar CRISP-DM [1].

Según Bickel et al. [142], CRISP-DM (*Cross-Industry Standard Process for Data Mining*) es una metodología estandarizada que se ha convertido en el *framework* predominante para proyectos de *Data Mining* y *Data Science* a nivel mundial.

Cogburn [143] considera que dado que *Text Mining* es una técnica relativamente nueva, de alguna forma no se encuentra estandarizada, por lo que CRISP-DM se posiciona como metodología a utilizar en proyectos de *Text Mining*, dado que se trata de una metodología flexible, sencilla de adaptar a distintos proyectos.

CRISP-DM comprende distintas fases, ilustradas en la Figura 4.1. Se puede observar que las distintas fases no siguen un esquema lineal, ya que cada fase puede ejecutarse más de una vez, y el proceso no tiene un orden estricto, dado que las diferentes etapas proporcionan *feedback* de cáracter valioso que guían los pasos siguientes.

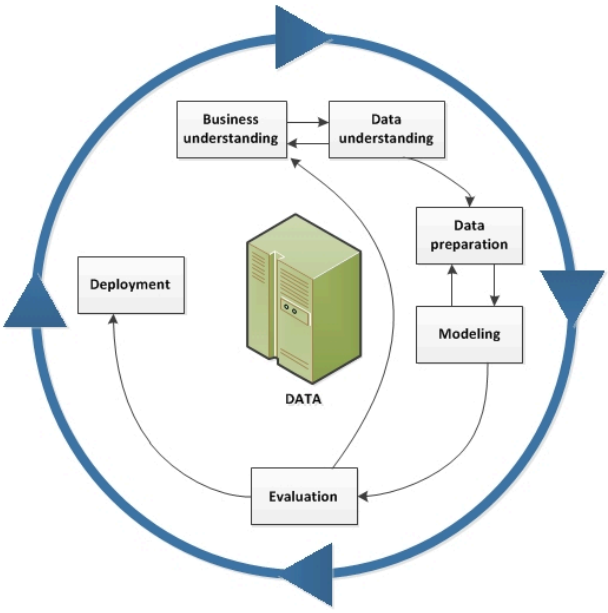


Figura 4.1: CRISP-DM [1] *Methodology*

Fuente: IBM [144]

En esta sección, se desarrollan las fases de la metodología CRISP-DM adaptadas al proyecto

de investigación, identificando actividades clave en cada una de las fases.

Se presentan las fases de *Business Understanding* (ver 4.2.1), *Data Understanding* (ver 4.2.2), *Data Preparation* (ver 4.2.3), *Modeling* (ver 4.2.4), *Evaluation* (ver 4.2.5) y *Deployment* (ver 4.2.6).

4.2.1. *Business Understanding*

Cogburn [143] señala que en los proyectos de *Text Mining*, en la fase de *Business Understanding* se debe lograr un entendimiento del problema que se quiere resolver.

La fase inicial de la investigación tiene sus cimientos en la definición del problema (ver Capítulo 3) y luego la determinación de objetivos (ver 4.1). Asimismo, la revisión de literatura se presenta como actividad clave en esta fase (ver Capítulo 2).

4.2.2. *Data understanding*

Según Cogburn [143], la fase de *Data Understanding* comprende *"identificar y entender los datos disponibles"*.

En la presente subsección, se desarrollan actividades claves orientadas a proporcionar entendimiento de los datos para luego poder aplicar técnicas que se adecúen a los mismos. Se presenta la identificación de los datos disponibles (ver 4.2.2.1) y actividades orientadas a entender los datos enmarcadas en un contexto exploratorio (ver 4.2.2.2).

4.2.2.1. Identificación de los datos disponibles

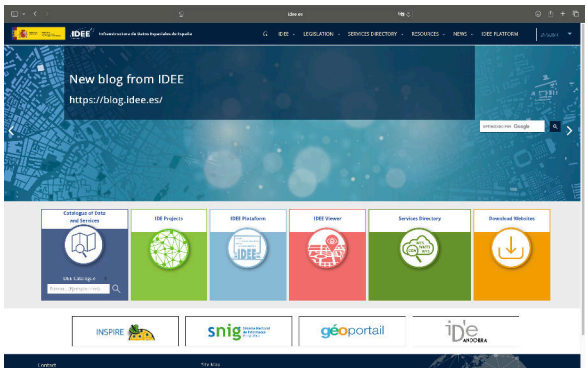
En este caso, el entendimiento de la información se lleva a cabo investigando las IDEs de manera manual con el objetivo de entender la información a tratar.

La elección de tomar las IDEs de España y Uruguay como caso de estudio se basa en:

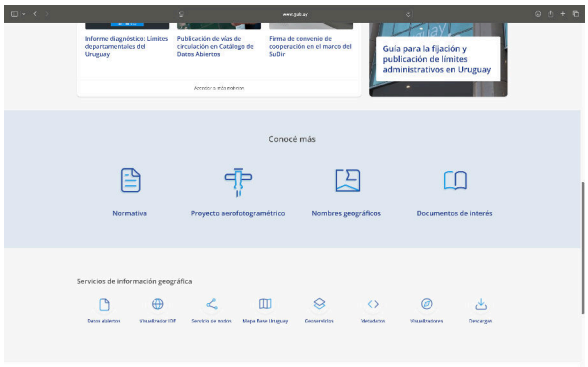
- La similitud en la estructura de sus catálogos de metadatos al compartir un *template* común.

- La riqueza de los datos públicos disponibles en ambas plataformas, lo que permite una actualización automática de los resultados al mantenerse al día con los cambios en los datos publicados. La riqueza está en que los datos son obtenidos automáticamente, los datos son públicos. Esto habilita también a que si los datos son eventualmente modificados o se agregan nuevos datos, la comparación se realiza con datos recientes. A su vez, es extensible para usar en diversas IDEs, convirtiéndolo en un comparador de distintas IDEs.
- La extensibilidad del enfoque utilizado, el cual puede aplicarse para comparar otras IDEs en el futuro.

En la Figura 4.2a se ilustra la página principal de la IDE de España y en la Figura 4.2b se ilustra la página principal de la IDE de Uruguay.



(a) IDE España [145].



(b) IDE Uruguay [146].

Figura 4.2: Páginas principales de las Infraestructuras de Datos Espaciales (IDE) de España y Uruguay

Fuente: Elaboración propia.

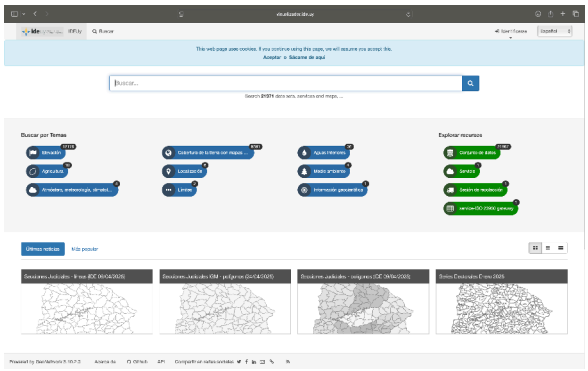
La selección de los elementos a comparar resultó un proceso complejo. Era necesario elegir un aspecto lo suficientemente relevante para el presente proyecto de investigación, pero a la vez manejable, considerando las limitaciones de tiempo y que la investigación se realiza de manera individual. En este contexto, resulta muy útil realizar una comparación manual de ambas IDEs para identificar los puntos en común.

En un principio, se *scrapearon* ambas páginas analizando atributos como el título , los meta tags, los *navs links* , el lenguaje y la documentación existente.

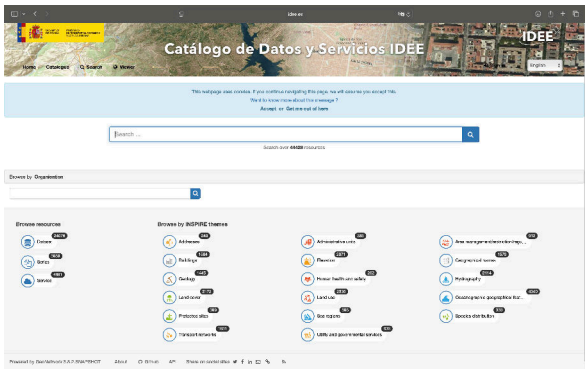
Posteriormente, teniendo en cuenta que las IDEs cuentan con un cátalogo de metadatos disponibles para el usuario final, se comparan las distintas categorías, examinando las categorías de cada IDE, así como las acciones y filtrado de datos disponibles. Finalmente, se *scrapean* ambos sitios de la página del catálogo de metadatos. Se investiga qué categorías tienen en común y cuántos datos tiene cada IDE por categoría , pero no se concibe como un punto fuerte de comparación.

Hasta que finalmente, se opta por comparar metadatos de las categorías.

En la Figura 4.3a se ilustra el cátalogo de metadatos para la IDE de Uruguay y en la Figura 4.3b se ilustra el cátalogo de metadatos para la IDE de España.



(a) Catálogo de metadatos de la IDE de Uruguay [147].

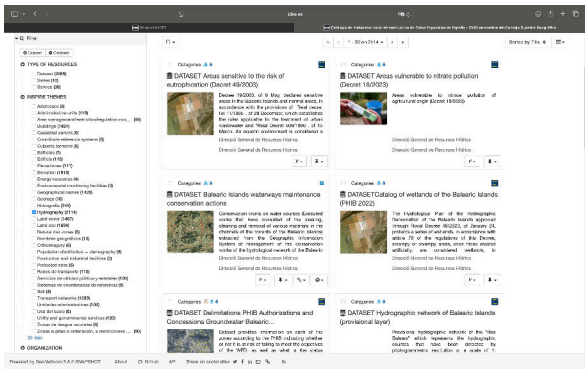


(b) Catálogo de metadatos de la IDE de España [148].

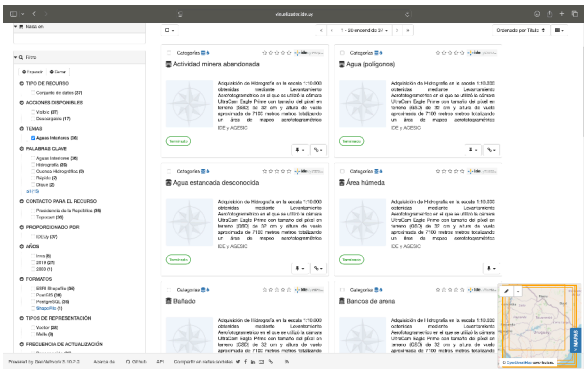
Figura 4.3: Catálogos de metadatos de la IDE de Uruguay y España

Fuente: Elaboración propia.

Dentro de las categorías existentes, se exploraron múltiples categorías, como *Hydrography* ilustrada en la Figura 4.4a y Aguas Interiores ilustrada en la Figura 4.4b, pero se descartaron debido a la poca cantidad de datos en Uruguay y la escasa relevancia de la información encontrada.



(a) IDE España-Categoría *Hydrography* dentro del catálogo de metadatos.



(b) IDE Uruguay—Categoría Aguas Interiores dentro del catálogo de metadatos.

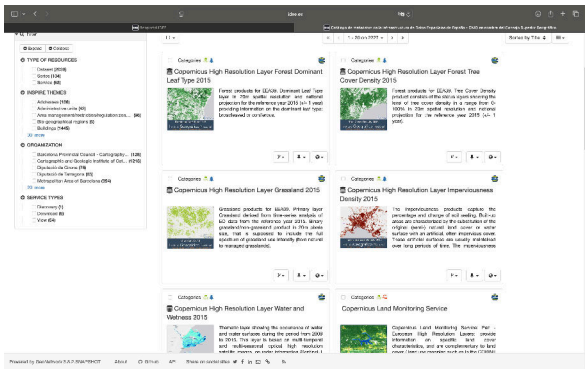
Figura 4.4: Categorías exploradas en los catálogos de metadatos de las IDE de España y Uruguay.

Fuente: Elaboración propia.

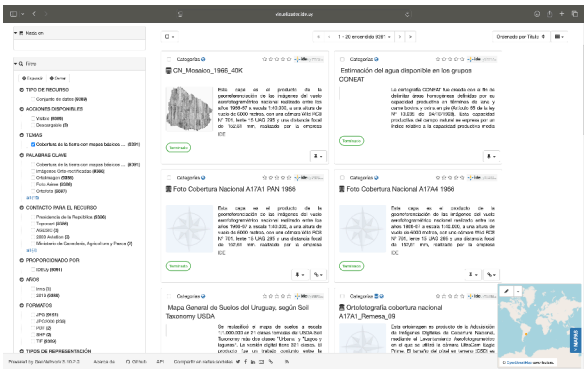
La selección de categorías se realiza con base en la compatibilidad y magnitud de datos disponibles en cada IDE.

Finalmente, se optó por comparar las siguientes categorías:

- ***Land Cover*:** Categoría de metadatos disponible en la IDE de España ilustrada en la Figura 4.5a.
- **Cobertura de la Tierra con mapas básicos e imágenes:** Categoría de metadatos disponible en la IDE de Uruguay ilustrada en la Figura 4.5b.



(a) IDE España — Categoría *Land Cover* dentro del catálogo de metadatos.



(b) IDE Uruguay — Categoría Cobertura de la Tierra con mapas básicos e imágenes dentro del catálogo de metadatos.

Figura 4.5: Categorías de metadatos elegidas para el caso de estudios de la IDE de España y Uruguay.

Fuente: Elaboración propia.

La selección de dichas categorías se fundamenta en las siguientes observaciones:

- Son categorías conceptualmente comparables y relevantes dentro del contexto de análisis de IDEs.
- Tienen una cantidad similar de datos y presentan información comparable en términos de estructura y términos semánticos.

Cabe destacar que resultó útil limitar el análisis de IDEs a España y Uruguay. Este enfoque permite centrar el estudio en categorías comparables, considerando la magnitud de los datos y su relevancia geoespacial. De este modo, se facilita una comparación directa y significativa, asegurando que el *dataset* sea representativo y útil para obtener conclusiones válidas.

4.2.2.2. Análisis exploratorio

El primer acercamiento y por tanto primer entendimiento de los datos disponibles se realiza en un contexto exploratorio.

En este contexto, se estudian distintos aspectos con el fin de comprender qué modelo de espacio vectorial (ver 4.2.4) es el más apropiado para los datos en cuestión.

En líneas posteriores, se describen algunos de los aspectos que se estudian en esta fase :

- **Idioma de los metadatos:** Proporciona información sobre el contexto idiomático de los metadatos.
- **Nube de palabras de los metadatos:** Este enfoque brinda un entendimiento de la semántica del contexto a estudiar.

El análisis exploratorio arroja resultados en términos descriptivos que se pueden ver en la sección 5.2.

4.2.3. *Data preparation*

En la presenta investigación , la preparación de los datos responde al esquema ilustrado en la Figura 4.6:

Figura 4.6: *Pipeline* de la fase *Data preparation*

Fuente: Elaboración propia.

A continuación, se describe de manera breve y concisa los puntos claves identificados en la preparación de datos ilustrados en la Figura 4.6:

- ***Scraping*:** El proceso de *Web Scraping* se consolida como primer resultado obtenido (ver 5.1). En este punto, es relevante tener en cuenta las consideraciones legales de los sitios a *scrapear*. Por este motivo, se revisan ambos archivos robots.txt como parte de metodología del proceso de *scraping* y no se encuentran limitantes.

- ***Structure data*:** La información debe ser estructurada en un formato que luego permita ser analizada. Se decide estructurar los datos guardando los siguientes atributos:
    - id
    - title
    - abstract
    - identifier
    - docLocale
- ***Feature selection***: Se estudia qué atributo o combinación de atributos arroja mayor información y se determina que el *abstract* es el atributo de mayor riqueza semántica.
- ***Clean data***: La limpieza de los datos es clave antes de iniciar el análisis. La estrategia tomada es borrar nulos, quitando atributos que contengan *abstract* (atributo seleccionado) vació.
- ***Text preprocessing:*** Se emplean las siguientes estrategias con el fin de preprocesar el texto:
    - *Lowercasing method*: El texto es pasado a minúscula.
    - *Removing stopwords*: Se remueven *stopwords*.
    - Eliminación de espacios y salto de líneas: Se quitan saltos de líneas innecesarios y se quitan espacios al principio y final de los textos.

    Cabe destacar que mayor preprocesamiento de texto puede ser empleado en futuras líneas de investigación (ver Capítulo 8).

4.2.4. *Modeling*

Según Cogburn [143], en esta fase , los datos deben ser correctamente modelados. En este sentido, esto presenta coherencia con el proceso de *Text Mining* (ver 2.3.2) en donde los datos textuales deben ser transformados de texto a un modelo de espacio vectorial.

El modelo de espacio vectorial debe seleccionarse cuidadosamente para garantizar la obtención de hallazgos relevantes. En este sentido, la elección del modelo depende de la naturaleza intrínseca de los datos identificada en la fase de *Data Understanding* (ver 4.2.2). Los datos preparados (ver 4.2.3) consisten en oraciones donde el contexto semántico resulta fundamental; por lo tanto, no deben tratarse como conjuntos de palabras aisladas. Dos oraciones semánticamente similares deberían reflejar dicha similitud en el modelo, lo cual es esencial para los objetivos de la presente línea de investigación.

En particular, se estudian dos enfoques y se llevan a cabo diversas pruebas con TF-IDF (*Term Frequency-Inverse Document Frequency*) y *Sentence Transformers*.

Según Dang et al. [149], TF-IDF (*Term Frequency-Inverse Document Frequency*) representa a los documentos como vectores, calculando pesos para cada palabra en función de su frecuencia dentro de un documento y y qué tan poco frecuente es en todo el corpus. Westermann et al. [150] señalan que TF-IDF tiene capacidad limitada, ya que se basa únicamente en la frecuencia de palabras y por tanto no es posible capturar las relaciones semánticas entre palabras. Por otro lado, *Sentence Transformers* , según Reimers et al. [151] obtiene representaciones vectoriales de oraciones (*Sentence Embeddings*), cuyo propósito principal es preservar las similitudes entre oraciones, de manera que el espacio vectorial de oraciones similares esté cercano entre sí. Kroll et al. [152] agregan que *Sentence Transformers* entiende el contexto semántico de las oraciones, considerando la oración entera y no la palabra de manera individual.

Una vez analizado la capacidad semántica de TF-IDF y *Sentence Transformers*, se decide utilizar *Sentence Transformers*. En particular, se escoge el modelo paraphrase-multilingual-MiniLM-L12-v2, ya que según señala Mahboub et al. [153], este es un modelo de *embedding* multilingüe que fue entrenado en más de 50 idiomas y genera un vector de *embedding* de 384 dimensiones para la oración dada. Está diseñado principalmente para tareas de *clustering* y búsqueda semántica. En esta línea, dado la naturaleza de los datos que se analiza en la fase de *Data Understanding* (ver 4.2.2), se selecciona este modelo por su capacidad de reconocer oraciones semánticamente equivalentes, independientemente del idioma.

Cabe destacar que, debido al continuo avance de los modelos, es necesario mantener la investigación en esta área. Para este estudio se priorizó la capacidad semántica por sobre la velocidad de procesamiento, la cual no fue analizada en profundidad, ya que el objetivo no era optimizar el rendimiento, sino obtener resultados significativos en un contexto comparativo. El perfeccionamiento del modelo y la mejora de la velocidad de procesamiento constituyen aspectos que pueden explorarse en trabajos futuros (ver Capítulo 8).

4.2.5. *Evaluation*

El objetivo de la evaluación en esta investigación es obtener resultados significativos a partir de la comparación de los metadatos de ambas IDE, con el fin de analizar la naturaleza de los datos y explorar sus similitudes y relaciones.

Estas cuestiones no pueden abordarse de la misma manera que en un contexto típico de *Data Mining*. Por ello, la evaluación, realizada mediante técnicas de *Text Mining*, busca generar resultados interpretables que permitan extraer conclusiones relevantes sobre la relación entre los datos.

Para este propósito se emplean la técnica de Análisis de métricas (ver 4.2.5.1) y el Análisis de *clustering* (ver 4.2.5.2).

4.2.5.1. Análisis de métricas

'Measure what can be measured, and make measurable what cannot be measured.'

– Galileo Galilei. [154]

Siguiendo este principio, es fundamental respaldar el análisis comparativo con resultados cuantitativos. Contar con métricas numéricas permite evaluar de manera objetiva los datos, garantizando que las conclusiones sobre los datos sean sólidas y reproducibles.

En líneas siguientes se listan las métricas que se utilizan:

- *Silhouette Score*
- Davies-Bouldin
- Calinski-Harabasz
- *Elbow Method*

El objetivo es poder estudiar como se relaciona la información entre si según diversos criterios atribuidos por estas métricas, concluyendo con un análisis comparativo de los resultados individuales.

4.2.5.2. Análisis de *Clustering*

En el contexto del estudio comparativo de datos de las IDEs, se aplican técnicas de *clustering* con el objetivo de estudiar la agrupación de datos resultante. Se utiliza el valor de k óptimo hallado en la evaluación anterior (ver 4.2.5.1).

La técnica de *clustering* que se emplea es K-Means dado que el modelo utilizado (ver 4.2.4) genera vectores de *embeddings* que representan cada descripción como un punto en función de su similitud semántica. K-Means resulta un método simple y eficaz para agrupar estos puntos según su semejanza, facilitando la identificación de patrones entre ellos. Como trabajo futuro (ver Capítulo 8), se propone explorar otros algoritmos de *clustering* para analizar cómo varía la distribución de la información en función de diferentes enfoques. Luego, es complementado con PCA para reducir la dimensionalidad y visualizar la distribución de los datos.

4.2.6. *Deployment*

Esta fase requiere que todo el proceso involucrado en esta investigación esté debidamente documentado. Cada parte de los resultados experimentales y del análisis textual comparativo derivados de los hallazgos de esta investigación se discute, presenta y se elabora una prueba conceptual.

4.3. Arquitectura del prototipo

El objetivo de esta sección es presentar una descripción general del prototipo (ver 4.3.1) y presentar sus componentes y funciones (ver 4.3.2).

4.3.1. Descripción general del prototipo

- **IDE *Comparator*:** La aplicación que permite interactuar con el usuario. Esta aplicación realiza peticiones a servicios externos, como son las IDEs y a su vez interactúa con la base de datos realizando operaciones de lectura y escritura.
- **Base de datos (Mongodb):** Mongodb es la base de datos escogida para realizar la prueba de concepto.
- **Infraestructura de datos espaciales (IDE):** Servicio externo de donde se recopila información de los metadatos elegidos para *scraper* y analizar.

4.3.2. Componentes y funciones

En la figura 4.7 se muestra el diagrama de componentes:

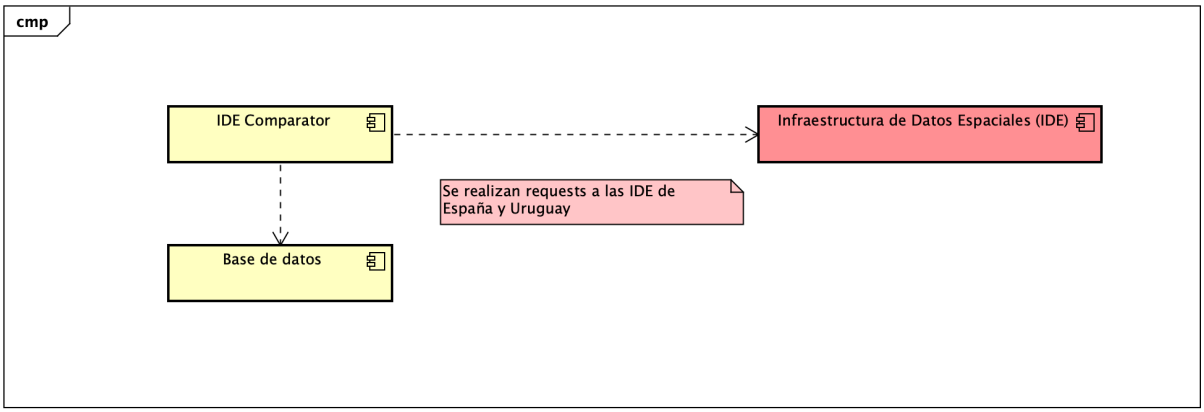


Figura 4.7: Diagrama de componentes

Fuente: Elaboración propia.

En líneas siguientes se detalla una breve descripción de los componentes:

- **IDE Comparator:** Corresponde a la prueba conceptual desarrollada en Streamlit, escogida dado que ofrece una curva de aprendizaje baja y facilita la experimentación con diferentes datos.
- **Base de datos:** Tiene como objetivo almacenar registros de los metadatos.
- **IDE:** Servicio externo del que depende IDE Comparator.

Nota: Aunque IDE Comparator depende de múltiples servicios y recursos externos, se trata de una prueba de concepto diseñada para explorar rápidamente ideas y datos. En trabajos futuros sería recomendable desarrollar la herramienta sobre una arquitectura más desacoplada y robusta, que reduzca estas dependencias (ver Capítulo 8).

4.3.3. Flujos principales

En esta sección se describen los flujos principales de la prueba conceptual a través de diagramas de casos de uso, junto con una breve descripción de cada módulo y las librerías empleadas para su desarrollo.

Se presentan los flujos de Metadatos (ver 4.3.3.1), Scraper (ver 4.3.3.2), Análisis descriptivo (ver 4.3.3.3), Análisis de métricas (ver 4.3.3.4) y Análisis de *clustering* (ver 4.3.3.5).

4.3.3.1. Metadatos

En la Figura 4.8 se ilustra el diagrama de caso de uso del módulo de metadatos. En este módulo, el investigador puede visualizar y descargar metadatos. La librería que se emplea en este proceso es `pymongo` , ya que posibilita la conexión y consulta de la información.

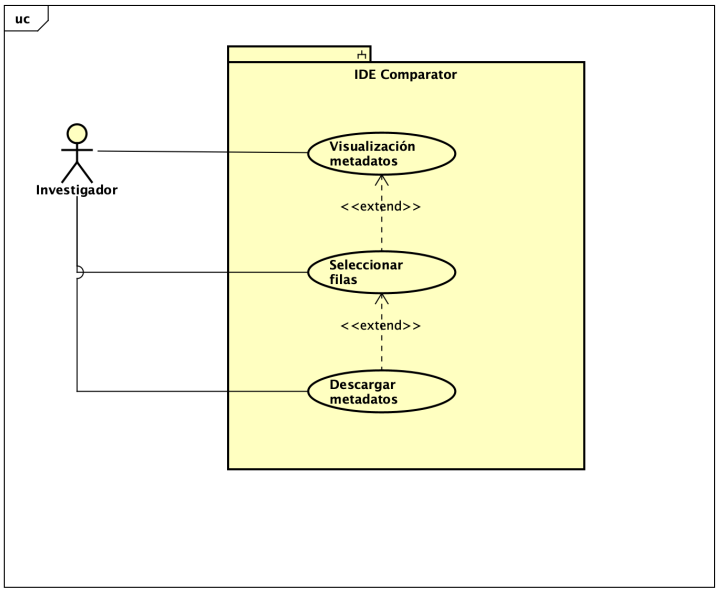


Figura 4.8: Diagrama de caso de uso de metadatos

Fuente: Elaboración propia.

4.3.3.2. Scraper

En la Figura 4.9 se muestra el caso de uso correspondiente al módulo de *scraper*. En este flujo, el investigador selecciona el país del cual desea extraer los datos y luego inicia el proceso de scrapeo. Durante la ejecución, se realizan peticiones a la IDE seleccionada, y los metadatos obtenidos se almacenan en la base de datos.

Para este proceso se utiliza la librería `requests`, implementando tiempos de espera entre solicitudes y peticiones paginadas, con el fin de evitar la sobrecarga del servidor.

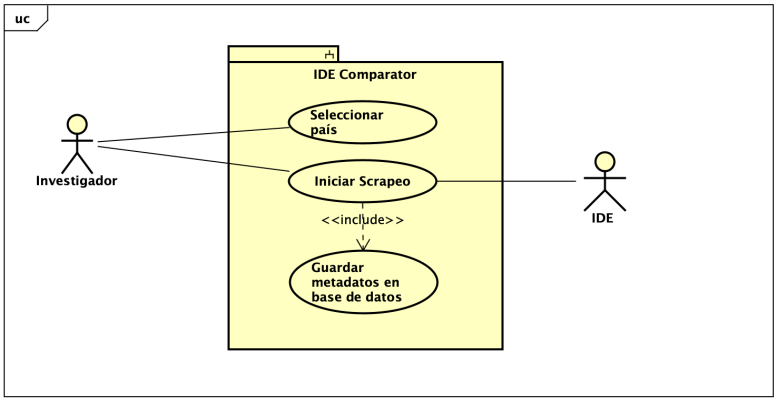


Figura 4.9: Diagrama de caso de uso de scraper

Fuente: Elaboración propia.

4.3.3.3. Análisis descriptivo

La Figura 4.10 muestra el caso de uso correspondiente al análisis descriptivo. El objetivo de este módulo es proporcionar al investigador un primer acercamiento al contenido de los datos. Para ello, se emplea la librería `WordCloud` para la generación de nubes de palabras, junto con `nltk`, utilizada para obtener las *stopwords* en español e inglés. De esta forma, dichas palabras se eliminan antes de construir las nubes de palabras, logrando una representación más significativa del texto.

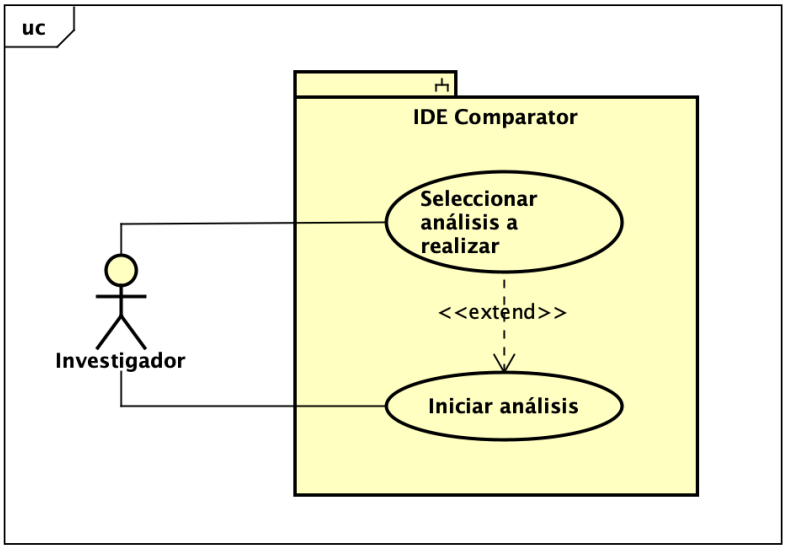


Figura 4.10: Diagrama de caso de uso de análisis descriptivo

Fuente: Elaboración propia.

4.3.3.4. Análisis de métricas

La Figura 4.11 ilustra el caso de uso correspondiente al análisis de métricas. En este módulo, el investigador puede configurar los parámetros necesarios y ejecutar el análisis de los metadatos.

El procesamiento se realizó utilizando la librería `scikit-learn` [155]. En suma, se emplea la herramienta `KneeLocator` para determinar el número óptimo de grupos mediante el método del codo, dada la dificultad de su interpretación gráfica. Por último, se utiliza el modelo `SentenceTransformer` de la librería `sentence-transformers` para la generación de los *embeddings* semánticos.

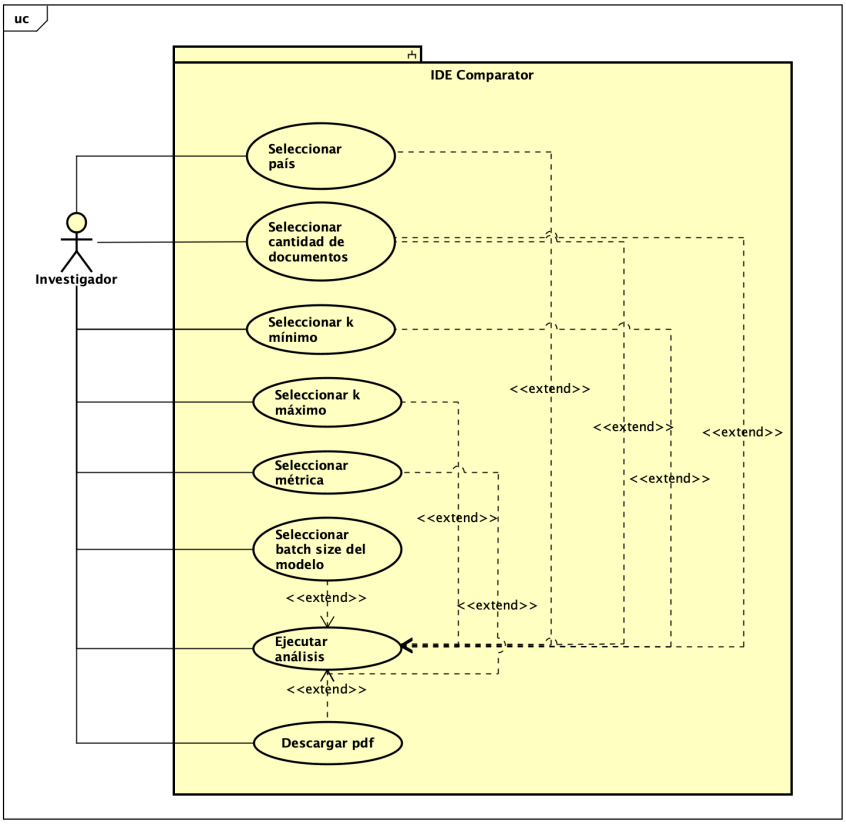


Figura 4.11: Diagrama de caso de uso de análisis de métricas

Fuente: Elaboración propia.

4.3.3.5. Análisis de *clustering*

Como se ilustra en la Figura 4.12, el módulo de Análisis de *Clustering* permite configurar los parámetros necesarios para generar los *clusters*. La generación de los *clusters* se realiza utilizando la clase `KMeans` del paquete `sklearn.cluster`, Por otro lado, los *embeddings* se generaron con el modelo paraphrase-multilingual-MiniLM-L12-v2 de `SentenceTransformer` perteneciente a la librería `sentence-transformers`.

El k óptimo de los *clusters* se evalua mediante la *inercia* y las métricas *Silhouette Score*, *Davies-Bouldin Score* y *Calinski-Harabasz Score*, disponibles en `sklearn.metrics`. Además, se utiliza la herramienta `KneeLocator` para determinar el número óptimo de grupos mediante *Elbow Method* con el fin de automatizar la interpretación visual.

Para la reducción de dimensionalidad se aplicó `PCA` (*Principal Component Analysis*), disponible en `sklearn.decomposition` . Finalmente, la librería `wordcloud` se utilizó para la creación de nubes de palabras que facilitan la interpretación de los temas predominantes en cada *cluster*.

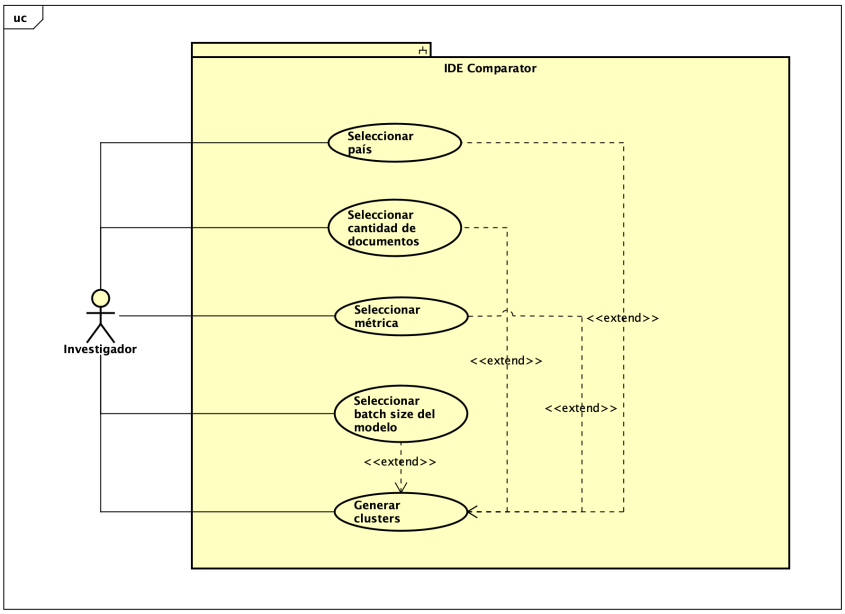


Figura 4.12: Diagrama de caso de uso de análisis de *clustering*

Fuente: Elaboración propia.

4.3.4. Tecnologías utilizadas

Las tecnologías que se utilizan para la construcción de la prueba conceptual son Base de datos (ver 4.3.4.1), Docker [156] (ver 4.3.4.2) y Python [157] como lenguaje de programación (ver 4.3.4.3).

4.3.4.1. Base de datos

Inicialmente, los datos fueron almacenados en archivos CSV con el objetivo de facilitar su interpretación y posterior análisis. En ellos se registraban campos como *title*, *description* y *link* asociados a cada metadato. Este formato resultó adecuado en una primera etapa, ya que permitía mantener la estructura y organización de la información de forma sencilla para futuros análisis comparativos.

No obstante, el incremento progresivo en el volumen de datos, junto con la necesidad de realizar auditorías más exhaustivas, evidenció las limitaciones de este enfoque y sentó las bases para la adopción de una base de datos.

En consecuencia, se optó por utilizar un motor de base de datos NoSQL, específicamente MongoDB [158].

La elección de utilizar MongoDB se fundamenta en los siguientes aspectos:

- **Esquema no definido desde un inicio:** El esquema de las colecciones no se define desde un inicio y debe mantenerse flexible ante posibles modificaciones en la estructura de los datos.
- **Simplicidad en el modelado:** Se prioriza una estructura sencilla, suficiente para satisfacer los requerimientos sin añadir complejidad innecesaria.

4.3.4.2. Docker

Se utiliza Docker para evitar la instalación local de MongoDB , permitiendo ejecutar la base de datos utilizada (ver 4.3.4.1) de forma aislada y gestionar los volúmenes de datos de manera eficiente. Esta estrategia fomenta la portabilidad y facilidad de despliegue en diferentes entornos.

4.3.4.3. Lenguaje de programación

Dado la enorme cantidad y librerías existentes en Python, se ha decidido utilizar Python para implementar la prueba conceptual.

Python es un lenguaje de programación que ofrece una gran cantidad de herramientas y *frameworks* para que *Web Scraping* sea confiable, fácil de utilizar y accesible a programadores de distintos niveles de experiencia[159].

Aprovechando su simplicidad, legibilidad y su amplia gama de bibliotecas especializadas para el procesamiento de datos y *Web Scraping*. Por otro lado, Python se caracteriza por su versatilidad. Se usa en diferentes etapas de desarrollo desde el *scraping* hasta el análisis.

5 Resultados obtenidos

El capítulo de Resultados Obtenidos tiene como propósito principal presentar de manera detallada los hallazgos de la presente investigación, mostrando cómo cada etapa metodológica contribuyó a la obtención de información significativa. Se incluyen los resultados del proceso de *Web Scraping* (ver Sección 5.1), donde se describe el mecanismo consolidad del proceso de *Web Scraping*. Luego, se realiza un análisis descriptivo de la información obtenida (ver Sección 5.2), que permite un primer acercamiento de los datos recopilados. Seguidamente, se exponen los resultados de las métricas de evaluación interna aplicadas, incluyendo *Silhouette Score* (ver Sección 5.3), Davies-Bouldin (ver Sección 5.4) y Calinski-Harabasz (ver Sección 5.5), así como el análisis a través del *Elbow Method* (ver Sección 2.3.4.3), lo que permite comprender indagar en el relacionamiento de los datos. Posteriormente, se presenta un análisis de *clustering* más detallado (ver Sección 5.7), que aporta información sobre la naturaleza de los datos y las relaciones entre ellos. Finalmente, se introduce la herramienta IDE *Comparator* (ver Sección 5.8), la cual integra los hallazgos anteriores.

5.1. Proceso de *Web Scraping*

El primer desafío y resultado obtenido fue el proceso de *Web Scraping*. En este apartado se presenta la evolución del enfoque tomado para llevar a cabo el *scraping* (ver 5.1.1) y el esquema consolidado del proceso de *Web Scraping* (ver 5.1.2).

5.1.1. Evolución del *scraping* de los datos

En esta subsección se describe cómo el proceso de *Web Scraping* evolucionó a lo largo de la investigación, adaptándose a las necesidades y desafíos que surgieron al profundizar en el problema. Se muestran los cambios implementados, desde las primeras aproximaciones hasta la estrategia final optimizada para la obtención de metadatos de las IDEs.

En 5.1.1.1 se presenta el primer acercamiento a la obtención de datos, considerando la página

como contenido dinámico. Posteriormente, en 5.1.1.2 se detalla el enfoque optimizado, basado en tratar a la página como contenido estático, lo que permitió simplificar el proceso y obtener los datos de manera más eficiente y rápida. Por último, en 5.1.1.3 se propone una comparación de ambos enfoques.

5.1.1.1. Enfoque de la etapa 1: *Scraping* mediante Selenium

En un primer momento, se utilizó Selenium para obtener datos de páginas dinámicas.

Durante esta etapa se realizaron las siguientes observaciones:

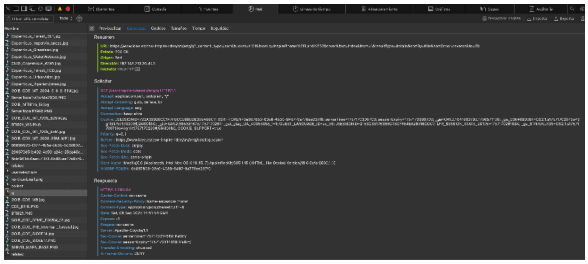
- ***Scraping* de subpáginas:** No solo se requería obtener información de las páginas principales, sino también de las subpáginas de cada metadato. Inicialmente, se implementó un enfoque que identificaba y seguía los *links* hacia estas subpáginas. Posteriormente, se modificó hacia Requests, ya que se logró acceder directamente a los datos mediante la API que las páginas utilizan internamente.
- **Páginas dinámicas:** Al tratarse de contenido cargado dinámicamente, fue necesario recorrer todas las páginas para garantizar la obtención completa y consistente de los datos, lo que implicaba tiempos de respuesta elevados.
- **Integridad de los datos:** Para asegurar que los datos fueran fiables, se realizaron múltiples ejecuciones del *scraper* y pruebas para verificar la correcta recolección y almacenamiento de los metadatos, incluyendo la detección de duplicados.
- **Uniformidad en el *template*:** Ambas IDEs utilizan un *template* similar, lo que permitió reutilizar parte del código y automatizar el proceso de manera más eficiente.

Debido a los altos tiempos de respuesta, este enfoque fue descartado para la etapa final.

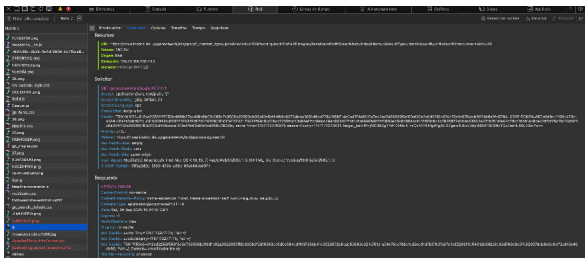
5.1.1.2. Enfoque de la etapa 2: *Scraping* mediante Requests

Tras analizar cuidadosamente las peticiones de red realizadas por las IDEs, se identificó que los datos podían obtenerse directamente a través de la API interna de cada IDE. Esto permitió reemplazar Selenium por Requests, optimizando considerablemente los tiempos de procesamiento.

En las Figuras 5.1a y 5.1b se muestra la inspección de las peticiones de red realizadas por las IDEs de España y Uruguay, respectivamente.



(a) IDE de España



(b) IDE de Uruguay

Figura 5.1: Inspección de las peticiones de red realizadas por las IDEs.

Fuente: Elaboración propia.

5.1.1.3. Comparación de los enfoques tomados

Es sabido que toda solución implica un *trade-off*; por ello, fue necesario analizar ambos enfoques, sin que ello implique que uno sea inherentemente mejor que el otro.

A continuación, se detallan las observaciones estudiadas:

- **Tiempo de respuesta:** Considerando la página como contenido dinámico (ver 5.1.1.1), los tiempos de obtención podían extenderse por horas, especialmente con una conexión inestable. En cambio, con el enfoque 2 (ver 5.1.1.2), la extracción se realizaba en cuestión de minutos, reduciendo drásticamente el tiempo de procesamiento.
- **Fidelidad del contenido:** Con Selenium (ver 5.1.1.1), la extracción era relativamente robusta frente a cambios menores en la página, como modificaciones en el contenido o

etiquetas, aunque seguía existiendo algún riesgo de pérdida de datos. En cambio, con el enfoque final basado en Requests (ver 5.1.1.2), la obtención de datos depende directamente de las APIs internas, por lo que cualquier cambio puede afectar la extracción.

En este caso, dado que los tiempo de respuesta eran demasiado altos en el primer enfoque tomado (ver 5.1.1.1), se optó por continuar con el enfoque tomado al final (ver 5.1.1.2). Sin embargo, la fidelidad del contenido se trata de un aspecto a revisar en trabajos futuros, tal como se detalla en el Capítulo 8.

5.1.2. Esquema consolidado del proceso de *Web Scraping*

El flujo final de extracción de datos se estructura en tres etapas correspondientes a las etapas del proceso de *Web Scraping* (ver 2.1.2).

1. **Obtención:** Se realizan peticiones HTTP a la API de cada IDE.
2. **Extracción:** Se obtiene la respuesta en formato JSON.
3. **Transformación:** Se seleccionan y almacenan los atributos relevantes de cada metadato, tales como *abstract*, *identifier* y *title*.

Con la información de texto obtenida, se sentaron los cimientos para iniciar la etapa de análisis de datos.

5.2. Análisis descriptivo

El análisis descriptivo utilizado consiste en realizar nubes de palabras para ambos países con el objetivo de analizar términos frecuentes.

Al comienzo del análisis con cada conjunto de datos, se contaba con una muestra poco representativa de los metadatos (aproximadamente 100) extraídas en orden alfabético. Esto lleva a la necesidad de recolectar una mayor cantidad de datos para evitar analizar información sesgada. Por este motivo, luego se decide aumentar la cantidad de datos extraídos.

Posteriormente, se elaboran nubes de palabras a partir de las descripciones contenidas en cada metadato, con el objetivo de identificar patrones léxicos y términos de alta frecuencia que pudieran aportar una visión general del contenido.

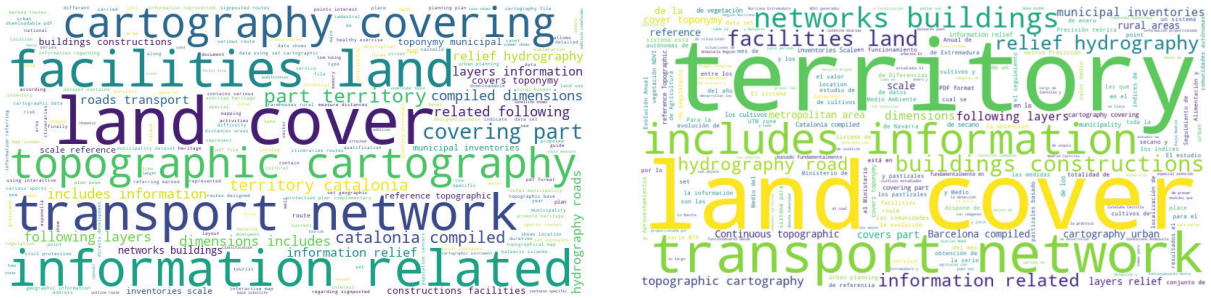
El presente análisis descriptivo se fundamenta en la necesidad de realizar un análisis exploratorio con el fin de entender la semántica que caracteriza a los datos. Este análisis exploratorio corresponde a la fase de *Data Understanding* enmarcada en la metodología propuesta (ver 4.2.2.2). Este análisis exploratorio arroja resultado de índole descriptivo, consolidándose como parte de los resultados obtenidos.

En este apartado, se presentan las principales observaciones obtenidas tras el análisis descriptivo para los metadatos de la IDE de España (ver 5.2.1) y Uruguay (ver 5.2.2). Por último, se presentado un análisis comparativo de ambas IDEs en un contexto descriptivo (ver 5.2.3).

5.2.1. IDE España

A continuación se presentan nubes de palabras extraídas de la IDE de España para la categoría *Land Cover*.

En la Figura 5.2a se presenta una nube de palabras para un conjunto de 100 metadatos extraídos de la IDE de España y en la Figura 5.2b se muestra una nube de palabras para 2200 metadatos.



(a) 100 datos (b) 2200 datos

Figura 5.2: Nube de palabras para la categoría de metadatos de España de *Land Cover* para diferentes tamaños de conjunto de datos.

Fuente: Elaboración propia.

A partir del procesamiento progresivo de los metadatos, se observan las siguientes tendencias:

- **Incremento de términos generales:** Al aumentar la cantidad de metadatos analizados, se evidencia una mayor presencia y relevancia de ciertos términos clave, como *territory*, *land cover*, *transport* y *network*.
- **Pérdida de peso de términos particulares:** Términos como *topographic cartography*, *cartography covering* o *catalonia compiled* pierden relevancia en el conjunto ampliado, reflejando que se trataban de fenómenos locales o específicos de una muestra reducida.
- **Sesgo temático inicial:** En la nube de palabras de 100 metadatos (ver 5.2a) predominaban temáticas particulares, lo que generaba un sesgo hacia registros aislados en lugar de mostrar tendencias generales.
- **Sesgo idiomático:** En esa misma nube inicial se observa una dominancia del inglés, lo que llevó a pensar que la totalidad de los documentos estaban en este idioma. Sin embargo, al ampliar al capturar un tamaño de datos mayor, se observa una distribución más heterogénea en cuanto a idioma (ver 5.2b), revelando que la percepción inicial estaba condicionada por la muestra reducida.

En términos generales, el *scraping* progresivo fue fundamental para reducir sesgos tanto temáticos como idiomáticos. Con una muestra pequeña, las conclusiones se veían distorsionadas por términos particulares y por un aparente predominio del inglés. Al ampliar el número de metadatos, se logra una representación más robusta y generalizable de la categoría *land Cover*, donde destacan los conceptos centrales y transversales de la categoría *Land Cover* de la IDE de España.

En la Tabla 5.1 se presenta la distribución de cantidad de documentos por idioma. Esta distribución muestra que por lo menos en términos de idioma para la categoría *Land Cover* no cumple con un estándar preestablecido. Sin embargo, no se desarrollan pruebas que respalden la razón por la cual existen metadatos en idiomas distintos, ya que está fuera del alcance de la presente investigación.

Idioma	Cantidad de documentos
Inglés	1889
Español	310

Tabla 5.1: Cantidad de documentos por idioma.

Fuente: Elaboración propia.

5.2.2. IDE Uruguay

A continuación se presentan nubes de palabras extraídas de la IDE de Uruguay para la categoría cobertura con mapas básicos e imágenes.

En la Figura 5.3a se presenta una nube de palabras para un conjunto de 100 metadatos extraídos de la IDE de Uruguay y en la Figura 5.3b se muestra una nube de palabras para 9391 metadatos.

(a) 100 datos (b) 9391 datos

Figura 5.3: Nube de palabras para la categoría de metadatos de Uruguay de cobertura con mapas básicos e imágenes para diferentes tamaños de conjunto de datos.

Fuente: Elaboración propia.

En la Tabla 5.2 se presenta la distribución de cantidad de documentos por idioma. Esta distribución muestra que por lo menos en términos de idioma para la categoría cobertura con mapas básicos e imágenes de la IDE de Uruguay cumple con un estándar preestablecido.

Idioma	Cantidad de documentos
es	9391

Tabla 5.2: Cantidad de documentos por idioma

Fuente: Elaboración propia.

5.2.3. Análisis comparativo

El análisis descriptivo realizado sobre los metadatos de las IDEs de España y Uruguay bajo la categoría *Land Cover* para el caso de la IDE de España y cobertura con mapas básicos e imágenes para el caso de la IDE de Uruguay revela diferencias significativas en la forma en que cada país estructura, presenta y orienta los metadatos.

En las nubes de palabras presentadas para la IDE de España (ver 5.2.1) y Uruguay (ver 5.2.2) se presentan términos más frecuentes en cada conjunto de datos, lo que permite inferir la orientación temática de los metadatos.

En la Tabla 5.3 se presenta una comparación que deriva del análisis individual preliminar realizado para la IDE de España (ver 5.2.1) y Uruguay (ver 5.2.2).

Aspecto	**IDE España**	**IDE Uruguay**
Nivel de interpretación	Catálogo más interpretado y orientado a la toma de decisiones.	Datos más técnicos y primarios.
Clasificación	Clasifica en niveles de detalle *(layers, transport).*	Se centra en imágenes de mayor cobertura.
Orientación de datos	Gestión y planificación; capas disponibles para uso inmediato.	Cartografía básica; generación y almacenamiento de información base.
Diversidad lingüística	Cátalogo con metadatos en español e inglés.	Cátalogo con metadatos en español.
Diversidad en los metadatos	La nube de palabras revela una amplia variedad de términos, lo que indica una mayor heterogeneidad temática y lingüística en los metadatos.	La nube de palabras presenta un conjunto más limitado de términos, reflejando una mayor homogeneidad en el contenido y estructura de los metadatos.

Tabla 5.3: Comparación entre los metadatos de la IDE de España y Uruguay

Fuente: Elaboración propia.

Este análisis descriptivo constituye un primer acercamiento que permite identificar tendencias y diferencias significativas entre ambos catálogos. Un aspecto particularmente relevante fue la diversidad de los metadatos: mientras que en el caso de Uruguay predominan registros muy similares, en su mayoría vinculados a ortoimágenes, España presenta un conjunto mucho más heterogéneo en cuanto a temáticas y lenguajes. Esta observación, que

inicialmente surgió de la revisión manual de los metadatos, pudo ser confirmada mediante la construcción de las nubes de palabras.

De este modo, el análisis descriptivo permitió sustentar la hipótesis preliminar sobre la homogeneidad de los metadatos de la IDE de Uruguay de la categoría cobertura de la Tierra con mapas básicos e Imágenes frente a la diversidad de los metadatos de la IDE de España de la categoría *Land Cover*. En este punto, se logra obtener un resultado cualitativo de los datos y un aspecto que se presentaba como un punto fuerte de comparación de los datos. Por este motivo, resulta pertinente avanzar hacia un análisis cuantitativo mediante técnicas de *clustering*, con el objetivo de evaluar métricas de cohesión, separación y agrupamiento de los datos.

En este sentido, las nubes de palabras ofrecen una visión general y descriptiva, pero para obtener conclusiones más precisas sobre la estructura interna de los metadatos es necesario aplicar métricas específicas de *clustering* como *Silhouette Score* (ver 5.3), Davies-Bouldin (ver 5.4), *Elbow Method* (ver 2.3.4.3) y Calinski-Harabasz (ver 5.5). Asimismo, se procede a aplicar técnicas de *clustering* con el objetivo de analizar la estructura interna de los datos y poder seguir generando resultados pertinentes, las cuales se desarrollan en el apartado 5.7.

5.3. Análisis de métrica *Silhouette Score*

En este apartado se calcula la métrica *Silhouette Score*, la cual prioriza qué tan bien ubicado está cada punto dentro de su *cluster*.

Se busca determinar el valor óptimo de k para el conjunto de datos de la IDE de España bajo la categoría *Land Cover* y en el caso de Uruguay bajo la categoría Cobertura de la Tierra con Mapas básicos e imágenes, partiendo de categorías genéricas. En su esencia, se pretende explorar la naturaleza de la estructura de los datos, evaluando si pueden agruparse siguiendo criterios específicos. Esta agrupación nos permite entender la complejidad de los datos, entendiendo a cada uno de los *clusters* como un subtema dentro de las categorías de carácter genérico.

Si el valor óptimo de k es muy alto, indicaría que los datos son muy distintos entre sí y

que se necesitan muchos grupos para *clusterizar* la información. Por el contrario, si k es muy bajo, significaría que los datos son muy similares y que pocos *clusters* son suficientes para representarlos adecuadamente.

En términos generales, se busca dar respuesta a las siguientes preguntas de investigación:

- ¿Cuántos subtemas emergen en las categorías?
- ¿Cuántos grupos son necesarios para capturar la diversidad de los datos?
- ¿Existe variedad en la muestra seleccionada?
- ¿El incremento en la cantidad de datos contribuye a mejorar el valor de la métrica?

Estas preguntas pretenden ser respondidas en la presente sección. Para ello, en la subsección 5.3.1, se expone un análisis evolutivo de la métrica *Silhouette Score* mediante el *scraping* progresivo de los datos. Los hallazgos recabados en 5.3.1 arrojan resultados interesantes en un contexto comparativo y permiten realizar el estudio del análisis de la tasa de cambio de la métrica *Silhouette Score* (ver 5.3.2).

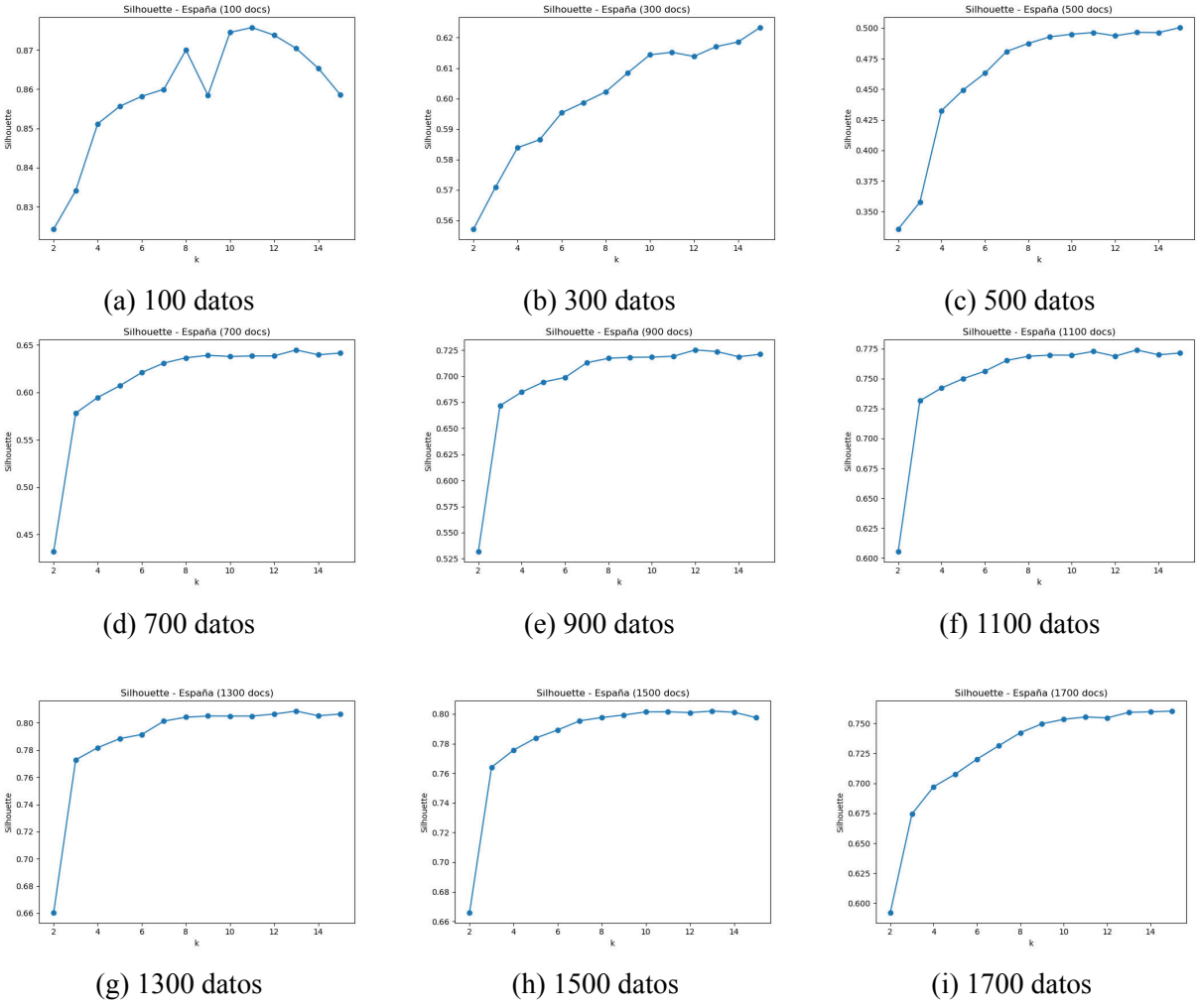
Por último, se realiza un análisis comparativo de los metadatos de ambas IDEs en función de la estructura de los datos (ver 5.3.3) y el k óptimo (ver 5.3.4).

5.3.1. Análisis evolutivo de la métrica *Silhouette Score* mediante el *scraping* progresivo de los datos

A continuación, se muestra la evolución de la métrica *Silhouette Score* mediante el *scraping* progresivo de los datos, así como el k óptimo a medida que se incrementa la muestra. Se realiza un análisis evolutivo para el caso de la IDE de España (ver 5.3.1.1) y para el caso de la IDE de Uruguay (ver 5.3.1.2) para luego proceder a realizar un análisis comparativo de ambos conjuntos de datos (ver 5.3.1.3).

5.3.1.1. IDE España

En las Figuras 5.4 y 5.5 se muestra la evolución del *Silhouette Score* a medida que se incrementa progresivamente el volumen de datos obtenido mediante *scraping*.



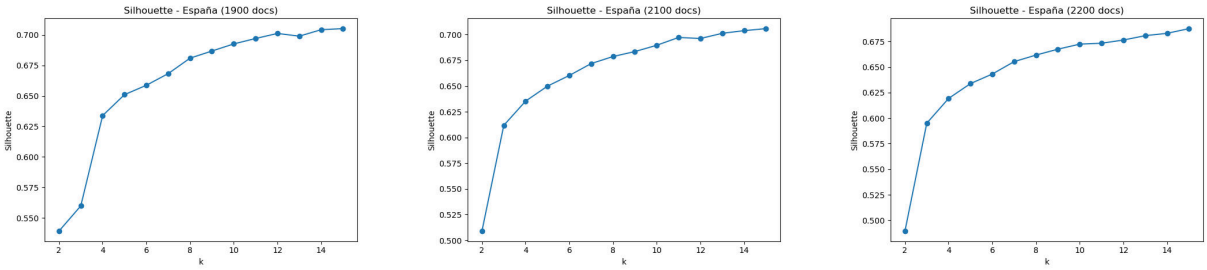
(a) 100 datos (b) 300 datos (c) 500 datos

(d) 700 datos (e) 900 datos (f) 1100 datos

(g) 1300 datos (h) 1500 datos (i) 1700 datos

Figura 5.4: Estimación de la cantidad de *clusters* (k) de la IDE de España con la métrica *Silhouette Score* variando el tamaño de la muestra (100 a 1700 datos).

Fuente: Elaboración propia.



(a) 1900 datos (b) 2100 datos (c) 2200 datos

Figura 5.5: Estimación de la cantidad de *clusters* (k) de la IDE de España con la métrica *Silhouette Score* variando el tamaño de la muestra (1900 a 2200 datos)

Fuente: Elaboración propia.

En la Tabla 5.4 se muestra la evolución del valor *Silhouette Score* en función de la cantidad de datos de España con su respectiva interpretación ilustrada en la Figura 5.6. Los valores se presentan con cinco decimales para conservar la precisión de la métrica, simplificando a la vez los cálculos y su visualización. La Tabla completa se puede visualizar en los Anexos (ver A.2).

País	Cantidad de datos	k óptimo	*Silhouette Score*
España	100	11	0.87573
	300	15 (↑)	0.62332 ↓
	500	11 (↓)	0.50253 ↓
	700	15 (↑)	0.64203 ↑
	900	15 (=)	0.72635 ↑
	1100	11 (↓)	0.77359 ↑
	1300	11 (=)	0.80842 ↑
	1500	13 (↑)	0.80379 ↓
	1700	14 (↑)	0.76355 ↓
	1900	14 (=)	0.71244 ↓
	2100	15 (↑)	0.70882 ↓
	2200	15 (=)	0.69251 ↓

Tabla 5.4: Evolución del valor *Silhouette* y de k en función de la cantidad de datos en España.

Fuente: Elaboración propia.

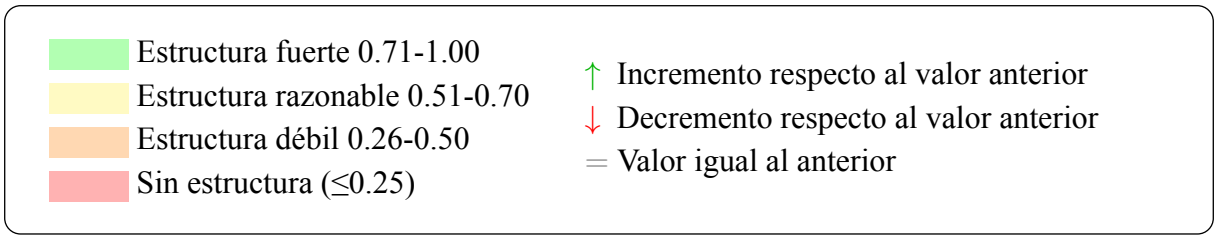


Figura 5.6: Interpretación de colores y símbolos de la Tabla 5.4

Fuente: Elaboración propia.

En líneas posteriores, se detalla una interpretación de los resultados mostrados anteriormente en la Tabla 5.4 y en las Figuras 5.4 y 5.5.

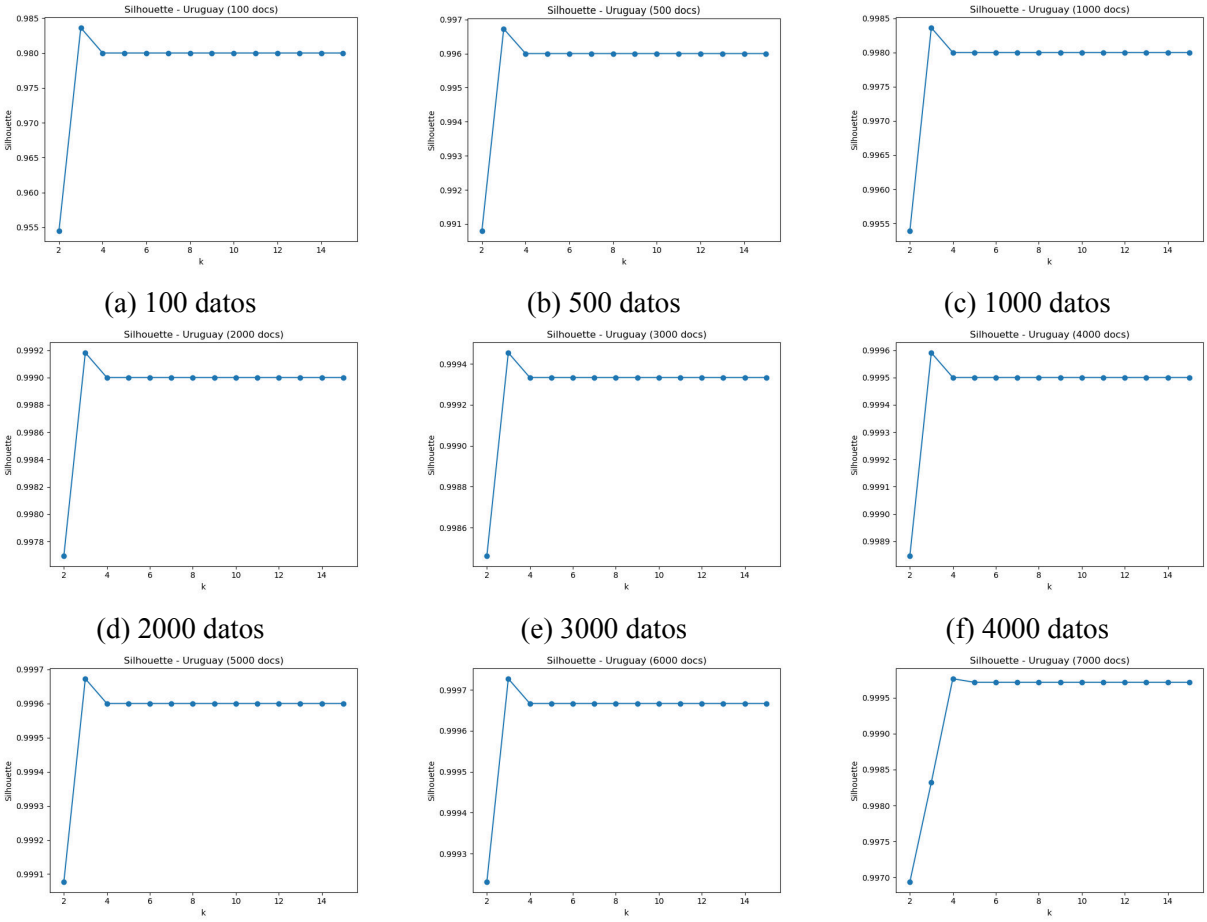
La métrica de *Silhouette Score* muestra un comportamiento interesante: con una muestra reducida de 100 metadatos, se obtiene un valor alto (≈ 0.87) con un número de *clusters* relativamente bajo (k=11), lo que podría sugerir una buena cohesión y separación. Sin embargo, este resultado está condicionado por la limitada diversidad temática de la muestra inicial.

A medida que se incrementa la cantidad de datos mediante *scraping* progresivo, el valor de *Silhouette Score* tiende a disminuir y el número óptimo de *clusters* aumenta (hasta k=15), lo que indica que emergen nuevos subtemas y estructuras internas que no eran capturados con menos datos. Esta evolución demuestra que el *scraping* continuo es fundamental para capturar con mayor precisión la complejidad semántica de los metadatos, evitando conclusiones simplificadas o sesgadas por muestras pequeñas.

Asimismo, se observa que la estructura de agrupamiento se estabiliza recién cuando se alcanza un volumen significativo de datos. Antes de completar el proceso de *scraping*, los resultados del *clustering* muestran variaciones que podrían llevar a una *clusterización* incorrecta o incompleta de los metadatos. Esto evidencia la necesidad de continuar con el *scraping* para capturar adecuadamente la diversidad temática y evitar interpretaciones prematuras basadas en muestras parciales.

5.3.1.2. IDE Uruguay

En las Figuras 5.7 y 5.8 se muestra la evolución del *Silhouette Score* a medida que se incrementa progresivamente el volumen de datos obtenido mediante *scraping*.



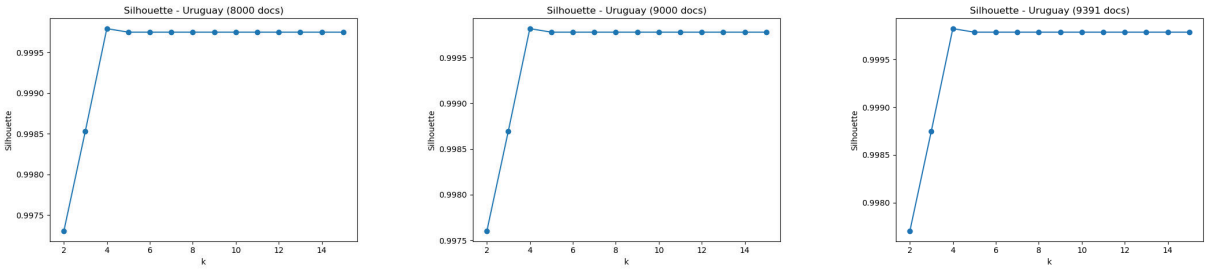
(a) 100 datos (b) 500 datos (c) 1000 datos

(d) 2000 datos (e) 3000 datos (f) 4000 datos

(g) 5000 datos (h) 6000 datos (i) 7000 datos

Figura 5.7: Estimación de la cantidad de *clusters* (k) de la IDE de Uruguay con la métrica *Silhouette Score* variando el tamaño de la muestra (100 a 7000 datos)

Fuente: Elaboración propia.



(a) 8000 datos (b) 9000 datos (c) 9391 datos

Figura 5.8: Estimación de la cantidad de *clusters* (k) de la IDE de Uruguay con la métrica *Silhouette Score* variando el tamaño de la muestra (5000 a 9391 datos)

Fuente: Elaboración propia.

En la Tabla 5.5 se muestra la evolución del valor *Silhouette Score* en función de la cantidad de datos de Uruguay con su respectiva interpretación ilustrada en la Figura 5.9. Los valores se presentan con cinco decimales para conservar la precisión de la métrica, simplificando a la vez los cálculos y su visualización. La Tabla completa se puede visualizar en los Anexos (ver A.1).

País	Cantidad de datos	k óptimo	*Silhouette Score*
Uruguay	100	3	0.98363
	1000	3(=)	0.99836 (↑)
	2000	3(=)	0.99918 (↑)
	3000	3(=)	0.99945 (↑)
	4000	3(=)	0.99959 (↑)
	5000	3(=)	0.99967 (↑)
	6000	3(=)	0.99973 (↑)
	7000	4(↑)	0.99976 (↑)
	8000	4(=)	0.99979 (↑)
	9000	4(=)	0.99982 (↑)
	9391	4(=)	0.99982 (=)

Tabla 5.5: Evolución del valor *Silhouette Score* en función de la cantidad de datos de la IDE de Uruguay

Fuente: Elaboración propia.

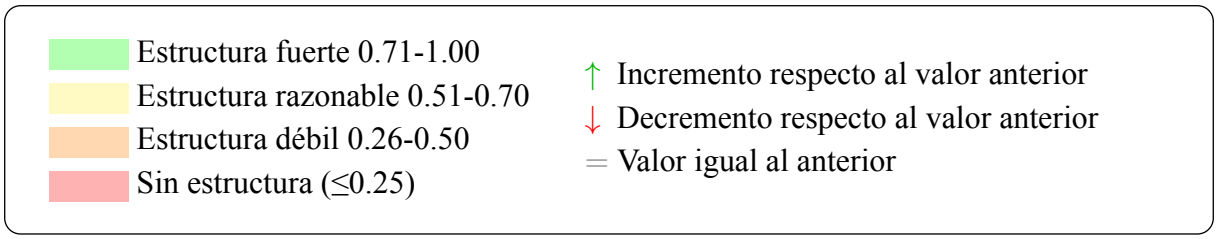


Figura 5.9: Interpretación de colores y símbolos de la Tabla 5.5

Fuente: Elaboración propia.

A continuación, se describe una interpretación de los resultados expuestos en la Figura 5.8 y en la Tabla 5.5.

Esta métrica revela un comportamiento particular: no se observan diferencias significativas al incrementar el volumen de datos. Como se detalla en la Tabla 5.5, el k óptimo para un conjunto de 100 registros es 0.983663, mientras que al analizar el total de 9391 registros se obtiene un k óptimo de 4 con un valor de 0.99982. Este resultado sugiere que, a pesar de aumentar considerablemente la cantidad de datos, la estructura subyacente se mantiene prácticamente inalterada. Sin embargo, este comportamiento no podía conocerse de antemano, ya que no se disponía de información previa sin realizar el *scraping* completo. Esto podría interpretarse como indicio de la existencia de pocos subgrupos bien definidos. No obstante, es necesario continuar evaluando otras métricas para obtener una visión más completa y robusta del panorama general (ver 5.6, 5.5 y 5.4).

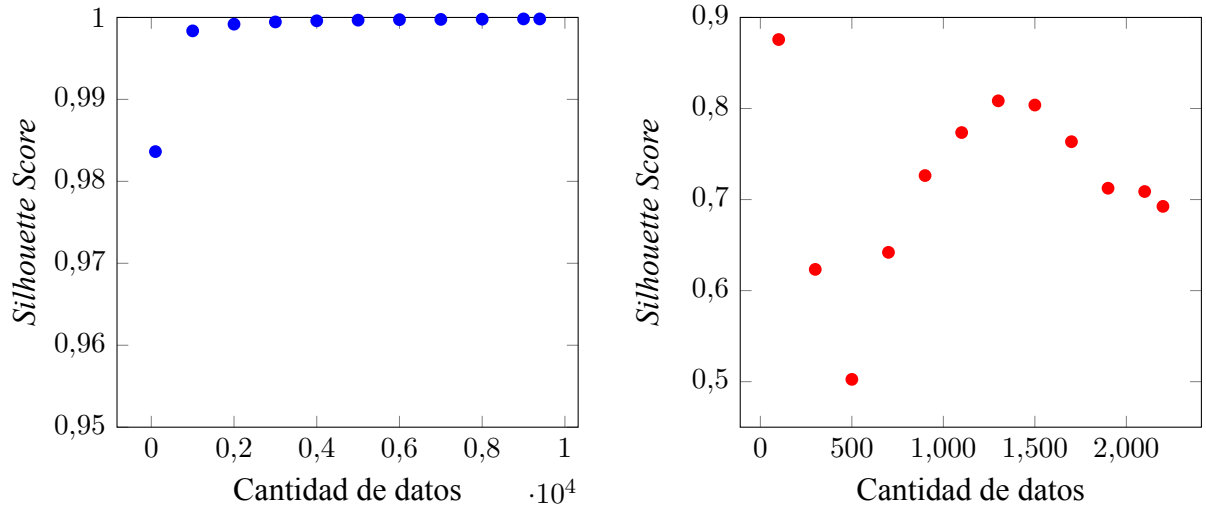
5.3.1.3. Análisis comparativo

En líneas posteriores, se desarrolla un análisis comparativo en el que se establece como línea de argumento que, al aumentar el volumen de datos, la presencia de mayor variabilidad actúa como un indicador de la diversidad intrínseca de la información obtenida.

Como se muestra en la Figura 5.10, la evolución del valor *Silhouette Score* presenta un comportamiento diferente entre Uruguay y España. En Uruguay, los valores son consistentemente altos y muestran mínima variación a medida que aumenta la cantidad de

datos, lo que indica una estructura de *clusters* estable y homogénea. En contraste, los datos de España presentan mayor variabilidad y valores más bajos, reflejando una estructura de *clusters* más heterogénea y sensible al tamaño del conjunto de datos.

De manera complementaria, la Figura 5.11 muestra la evolución del número óptimo de clusters (k). En Uruguay, k se mantiene estable en 3 hasta aproximadamente 6000 datos y aumenta levemente a 4 con volúmenes mayores. En España, k varía entre 11 y 15 incluso con pocas observaciones, evidenciando una mayor complejidad y dependencia del tamaño de la muestra. En conjunto, estos resultados reflejan cómo la diversidad temática y la cantidad de datos afectan tanto la calidad del *clustering* (medida por *Silhouette Score*) como la determinación del número óptimo de *clusters*, justificando la presentación en gráficas separadas para cada país.

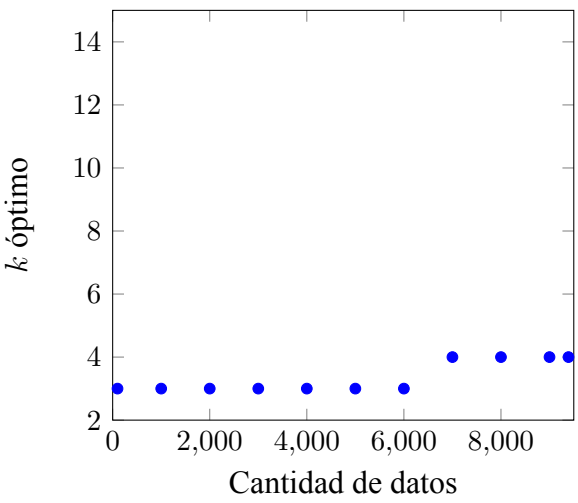


(a) IDE Uruguay.

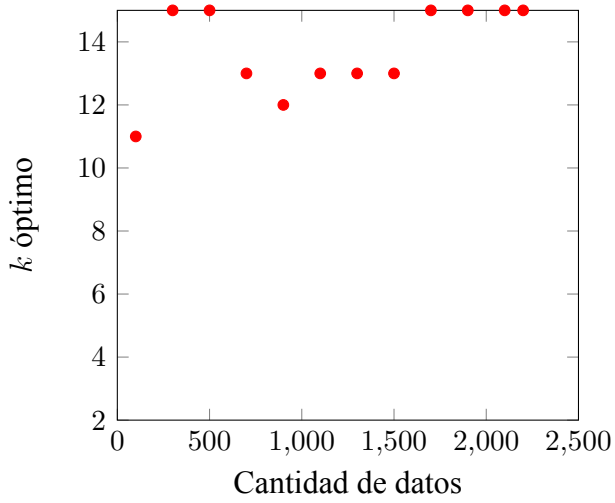
(b) IDE España.

Figura 5.10: Evolución del *Silhouette Score* para la IDE de Uruguay y España en función de la cantidad de datos

Fuente: Elaboración propia.



(a) Evolución del k óptimo en Uruguay



(b) Evolución del k óptimo en España

Figura 5.11: Evolución del k óptimo resultante de calcular la métrica *Silhouette Score* para España y Uruguay en función de la cantidad de datos.

Fuente: Elaboración propia.

5.3.2. Análisis de la tasa de cambio de *Silhouette Score*

En la comparación realizada en el análisis comparativo el el contexto del análisis evolutivo de *Silhouette Score* (ver 5.3.1) se observó que, en el caso de la IDE de España, a medida que se incorporan más datos mediante el proceso de *scraping*, se registran mayores variaciones en el valor del *Silhouette Score* en comparación con el caso de la IDE de Uruguay.

En este sentido, el análisis de la evolución del *Silhouette Score* a medida que aumenta la cantidad de datos extraídos ofrece resultados relevantes y plantea la necesidad de un estudio numérico más profundo. Dicho estudio permitirá sentar las bases para futuras investigaciones y avanzar hacia la automatización de la comparación, reduciendo la dependencia del trabajo manual del investigador (ver capítulo 8).

Se estudiará la tasa de cambio para la IDEs de España (ver 5.3.2.1) y Uruguay (ver 5.3.2.2) con el fin de sustentar de manera numérica los resultados obtenidos en líneas anteriores (ver 5.3.1) y sentar las bases para automatizar procesos manuales de comparación (ver Capítulo 8). Por tanto, el fin es poder capturar variables comparativas en base a aspectos cuantitativos que permitirán en un futuro avanzar hacia estudios automatizados. El análisis comparativo se realiza

con este fin y es expuesto en 5.3.2.3.

5.3.2.1. IDE España

Según la evolución presentada anteriormente (ver Tabla 5.4), la función del *Silhouette Score* en base a la cantidad de datos es:

$$f(N) = \begin{cases} 0{,}87573, & N = 100 \\ 0{,}62332, & N = 300 \\ 0{,}50253, & N = 500 \\ 0{,}64203, & N = 700 \\ 0{,}72635, & N = 900 \\ 0{,}77359, & N = 1100 \\ 0{,}80842, & N = 1300 \\ 0{,}80379, & N = 1500 \\ 0{,}76355, & N = 1700 \\ 0{,}71244, & N = 1900 \\ 0{,}70882, & N = 2100 \\ 0{,}69251, & N = 2200 \\ \text{no definida}, & N \notin \{100, 300, 500, 700, 900, 1100, 1300, 1500, 1700, 1900, 2100, 2200\} \end{cases}$$

Sea $f : \{100, 300, 500, \ldots, 2200\} \to \mathbb{R}$ la función que asigna a cada tamaño de datos N su *Silhouette Score*.

Como f está definida solo en puntos discretos, se aproxima la derivada mediante diferencias finitas hacia adelante:

$$\Delta S_i = f(N_i) - f(N_{i-1}), \quad \Delta N_i = N_i - N_{i-1}, \quad \frac{\Delta S_i}{\Delta N_i} = \frac{f(N_i) - f(N_{i-1})}{N_i - N_{i-1}}, \quad i = 2, \ldots, 12$$

De forma explícita, los cálculos son:

$$\Delta S_2 = f(300) - f(100) = 0{,}62332 - 0{,}87573 = -0{,}25241, \quad \Delta N_2 = 200, \quad \frac{\Delta S_2}{\Delta N_2} = -0{,}00126$$

$$\Delta S_3 = f(500) - f(300) = 0{,}50253 - 0{,}62332 = -0{,}12079, \quad \Delta N_3 = 200, \quad \frac{\Delta S_3}{\Delta N_3} = -0{,}00060$$

$$\Delta S_4 = f(700) - f(500) = 0{,}64203 - 0{,}50253 = 0{,}13950, \quad \Delta N_4 = 200, \quad \frac{\Delta S_4}{\Delta N_4} = 0{,}00070$$

$$\Delta S_5 = f(900) - f(700) = 0{,}72635 - 0{,}64203 = 0{,}08432, \quad \Delta N_5 = 200, \quad \frac{\Delta S_5}{\Delta N_5} = 0{,}00042$$

$$\Delta S_6 = f(1100) - f(900) = 0{,}77359 - 0{,}72635 = 0{,}04724, \quad \Delta N_6 = 200, \quad \frac{\Delta S_6}{\Delta N_6} = 0{,}00024$$

$$\Delta S_7 = f(1300) - f(1100) = 0{,}80842 - 0{,}77359 = 0{,}03483, \quad \Delta N_7 = 200, \quad \frac{\Delta S_7}{\Delta N_7} = 0{,}00017$$

$$\Delta S_8 = f(1500) - f(1300) = 0{,}80379 - 0{,}80842 = -0{,}00463, \quad \Delta N_8 = 200, \quad \frac{\Delta S_8}{\Delta N_8} = -0{,}00002$$

$$\Delta S_9 = f(1700) - f(1500) = 0{,}76355 - 0{,}80379 = -0{,}04024, \quad \Delta N_9 = 200, \quad \frac{\Delta S_9}{\Delta N_9} = -0{,}00020$$

$$\Delta S_{10} = f(1900) - f(1700) = 0{,}71244 - 0{,}76355 = -0{,}05111, \quad \Delta N_{10} = 200, \quad \frac{\Delta S_{10}}{\Delta N_{10}} = -0{,}00026$$

$$\Delta S_{11} = f(2100) - f(1900) = 0{,}70882 - 0{,}71244 = -0{,}00362, \quad \Delta N_{11} = 200, \quad \frac{\Delta S_{11}}{\Delta N_{11}} = -0{,}00002$$

$$\Delta S_{12} = f(2200) - f(2100) = 0{,}69251 - 0{,}70882 = -0{,}01631, \quad \Delta N_{12} = 100, \quad \frac{\Delta S_{12}}{\Delta N_{12}} = -0{,}00016$$

En la Tabla 5.6 se resumen los resultados obtenidos del cálculo de la derivada.

N	*Silhouette Score*	ΔS	ΔN	ΔS/ΔN
100	0.87573	-	-	-
300	0.62332	-0.25241	200	-0.00126
500	0.50253	-0.12079	200	-0.00060
700	0.64203	0.13950	200	0.00070
900	0.72635	0.08432	200	0.00042
1100	0.77359	0.04724	200	0.00024
1300	0.80842	0.03483	200	0.00017
1500	0.80379	-0.00463	200	-0.00002
1700	0.76355	-0.04024	200	-0.00020
1900	0.71244	-0.05111	200	-0.00026
2100	0.70882	-0.00362	200	-0.00002
2200	0.69251	-0.01631	100	-0.00016

Tabla 5.6: Derivada discreta $\Delta S/\Delta N$ para España - *Silhouette Score*

Fuente: Elaboración propia.

En la Figura 5.12 se ilustra la gráfica de la tasa de cambio.

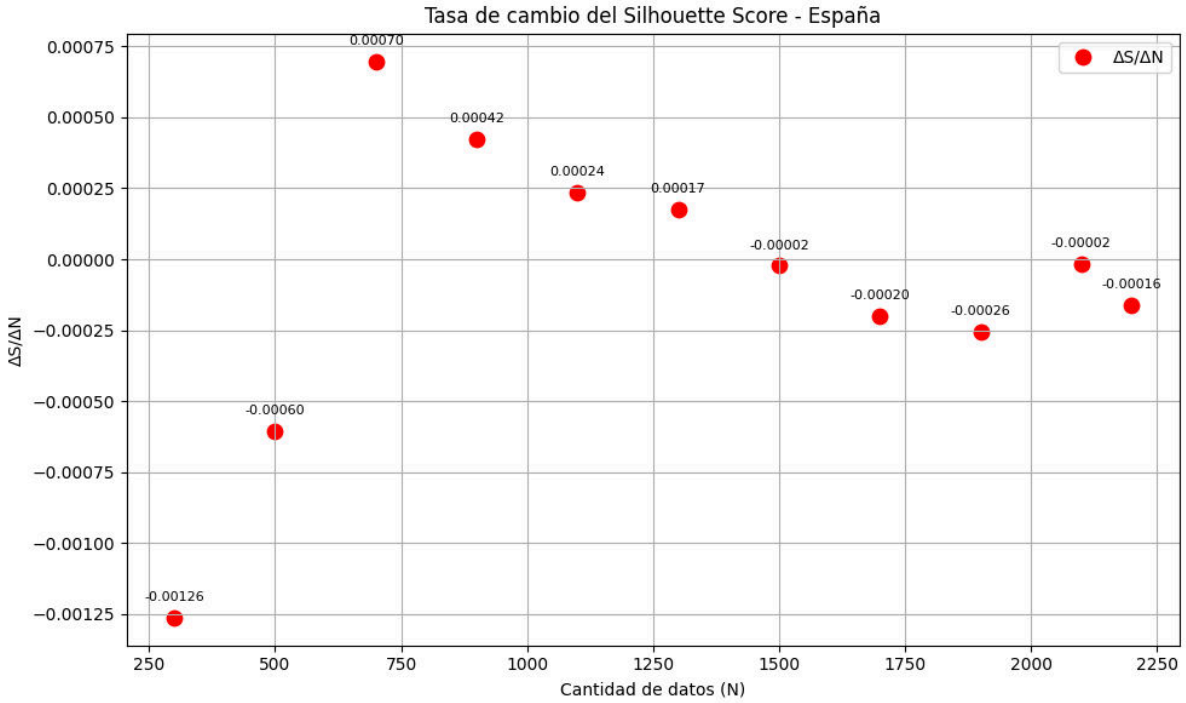


Figura 5.12: Tasa de cambio del valor *Silhouette Score*-IDE España

Fuente: Elaboración propia.

5.3.2.2. IDE Uruguay

Según la evolución presentada anteriormente (ver Tabla 5.5), la función del *Silhouette Score* en base a la cantidad de datos es:

Sea

$$f : \{100, 1000, 2000, \ldots, 9391\} \to \mathbb{R}$$

la función que asigna a cada tamaño de datos N su *Silhouette Score* en Uruguay.

$$f(N) = \begin{cases} 0{,}98363, & N = 100 \\ 0{,}99836, & N = 1000 \\ 0{,}99918, & N = 2000 \\ 0{,}99945, & N = 3000 \\ 0{,}99959, & N = 4000 \\ 0{,}99967, & N = 5000 \\ 0{,}99973, & N = 6000 \\ 0{,}99976, & N = 7000 \\ 0{,}99979, & N = 8000 \\ 0{,}99982, & N = 9000 \\ 0{,}99982, & N = 9391 \\ \text{no definida}, & N \notin \{100, 1000, 2000, \ldots, 9391\} \end{cases}$$

Como f está definida solo en puntos discretos, se aproxima la derivada mediante diferencias finitas hacia adelante:

$$\frac{\Delta S_i}{\Delta N_i} = \frac{f(N_i) - f(N_{i-1})}{N_i - N_{i-1}}, \quad i = 2, \ldots, 11$$

Cálculos explícitos:

$$\Delta S_2 = f(1000) - f(100) = 0{,}99836 - 0{,}98363 = 0{,}01473, \quad \Delta N_2 = 900, \quad \frac{\Delta S_2}{\Delta N_2} = 0{,}00002$$

$$\Delta S_3 = f(2000) - f(1000) = 0{,}99918 - 0{,}99836 = 0{,}00082, \quad \Delta N_3 = 1000, \quad \frac{\Delta S_3}{\Delta N_3} = 0{,}00000$$

$$\Delta S_4 = f(3000) - f(2000) = 0{,}99945 - 0{,}99918 = 0{,}00027, \quad \Delta N_4 = 1000, \quad \frac{\Delta S_4}{\Delta N_4} = 0{,}00000$$

$$\Delta S_5 = f(4000) - f(3000) = 0{,}99959 - 0{,}99945 = 0{,}00014, \quad \Delta N_5 = 1000, \quad \frac{\Delta S_5}{\Delta N_5} = 0{,}00000$$

$$\Delta S_6 = f(5000) - f(4000) = 0{,}99967 - 0{,}99959 = 0{,}00008, \quad \Delta N_6 = 1000, \quad \frac{\Delta S_6}{\Delta N_6} = 0{,}00000$$

$$\Delta S_7 = f(6000) - f(5000) = 0{,}99973 - 0{,}99967 = 0{,}00006, \quad \Delta N_7 = 1000, \quad \frac{\Delta S_7}{\Delta N_7} = 0{,}00000$$

$$\Delta S_8 = f(7000) - f(6000) = 0{,}99976 - 0{,}99973 = 0{,}00003, \quad \Delta N_8 = 1000, \quad \frac{\Delta S_8}{\Delta N_8} = 0{,}00000$$

$$\Delta S_9 = f(8000) - f(7000) = 0{,}99979 - 0{,}99976 = 0{,}00003, \quad \Delta N_9 = 1000, \quad \frac{\Delta S_9}{\Delta N_9} = 0{,}00000$$

$$\Delta S_{10} = f(9000) - f(8000) = 0{,}99982 - 0{,}99979 = 0{,}00003, \quad \Delta N_{10} = 1000, \quad \frac{\Delta S_{10}}{\Delta N_{10}} = 0{,}00000$$

$$\Delta S_{11} = f(9391) - f(9000) = 0{,}99982 - 0{,}99982 = 0{,}00000, \quad \Delta N_{11} = 391, \quad \frac{\Delta S_{11}}{\Delta N_{11}} = 0{,}00000$$

Observación: A partir de los 2000 documentos, la derivada es cero, indicando que el *Silhouette Score* se estabiliza y la cohesión y separación de los *clusters* no mejora con más datos.

En la Tabla 5.7, se resumen las derivada discretas aproximadas de *Silhouette Score* para la IDE de Uruguay:

N	ΔS	ΔN	$\Delta S/\Delta N$
100	-	-	-
1000	0.01473	900	0.00002
2000	0.00082	1000	0.00000
3000	0.00027	1000	0.00000
4000	0.00014	1000	0.00000
5000	0.00008	1000	0.00000
6000	0.00006	1000	0.00000
7000	0.00003	1000	0.00000
8000	0.00003	1000	0.00000
9000	0.00003	1000	0.00000
9391	0.00000	391	0.00000

Tabla 5.7: Derivadas discretas del *Silhouette Score* para los datos de Uruguay

Fuente: Elaboración propia.

En la Figura 5.13 se ilustra la gráfica de la tasa de cambio.

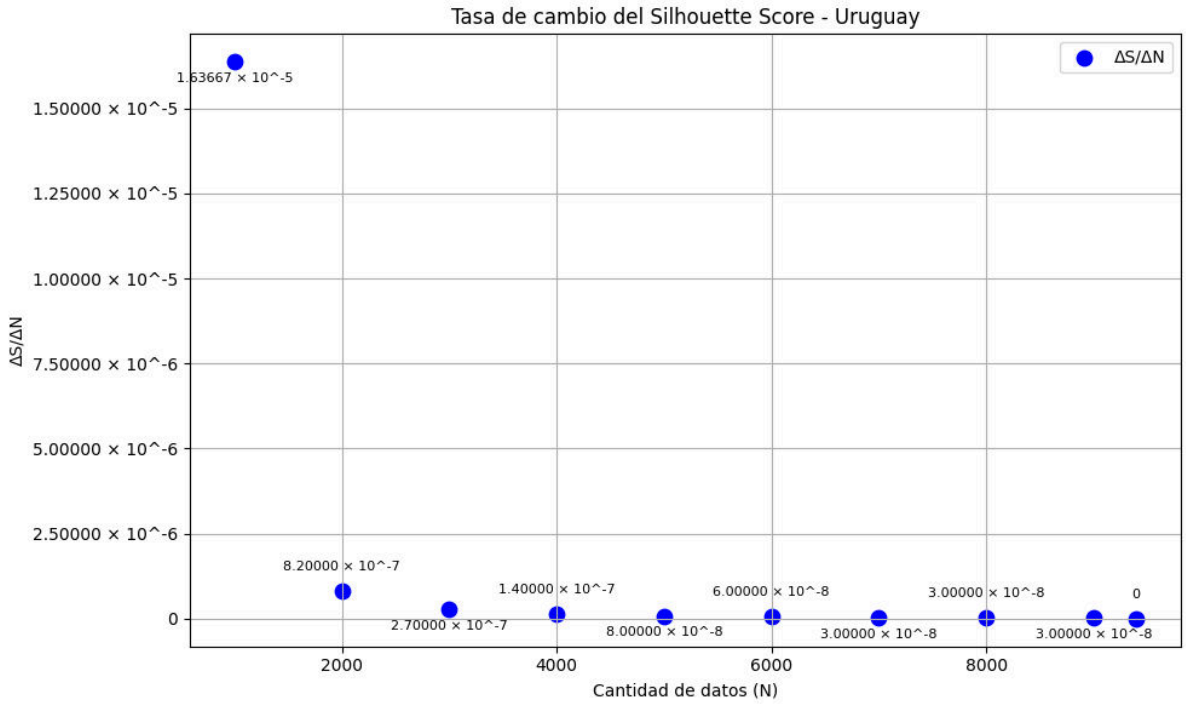


Figura 5.13: Tasa de cambio del valor *Silhouettte Score* -IDE Uruguay

Fuente: Elaboración propia.

5.3.2.3. Análisis comparativo

Una vez realizado el análisis de la tasa de cambio para ambas IDEs, se pueden destacar las siguientes observaciones:

- **Signo de la tasa de cambio:** En el caso de la IDE de Uruguay, como se muestra en la Tabla 5.13, las tasas de cambio son positivas. Esto indica que, a medida que se obtienen más datos mediante *scraping*, el *Silhouette Score* mejora. Se puede interpretar que los nuevos datos mantienen coherencia con los datos anteriores, fortaleciendo la cohesión interna de los *clusters*. Por el contrario, en la IDE de España, ilustrada en la Tabla 5.6, la tasa de cambio presenta valores tanto positivos como negativos, lo que sugiere que la incorporación de nuevos datos puede aumentar la variabilidad de la muestra.
- **Estabilidad:** En Uruguay, a partir de N = 2000, la tasa de cambio se estabiliza prácticamente en cero, lo que indica que la inclusión de más datos no aporta mejoras

significativas en el *Silhouette Score*. Este comportamiento no se observa en España, donde la tasa de cambio sigue variando con la incorporación de nuevos datos.

- **Fluctuaciones en la tasa de cambio:** Como se puede observar en las Tablas 5.13 y 5.6, la IDE de Uruguay presenta menores fluctuaciones en la derivada, reflejando una evolución más estable del *Silhouette Score*. En cambio, en la IDE de España, las fluctuaciones son mayores, indicando variaciones significativas en la cohesión interna a medida que se agregan datos.

5.3.3. Análisis comparativo de la estructura

En la Tabla 5.8 se muestra el valor de *Silhouette Score* con la totalidad de datos *scrapeados*.

País	Cantidad de datos	*Silhouette Score*	Datos utilizados
España	2200	0.6876	100 % de datos *scrapeados*
Uruguay	9391	0.9998	100 % de datos *scrapeados*

Tabla 5.8: Comparación de *Silhouette Score* para España y Uruguay usando todos los datos *scrapeados*

Fuente: Elaboración propia.

Según la métrica *Silhouette Score* y basado en la experiencia (ver Tabla 2.1), la estructura resultante de calcular el *Silhouette Score* para el caso de España es razonable, mientras que para el caso de Uruguay es fuerte.

5.3.4. Análisis comparativo del *k* óptimo

En la Tabla 5.9 se presenta una comparación del valor de k óptimo para España y Uruguay usando todos los datos *scrapeados*.

País	Cantidad de datos	k óptimo	Datos utilizados
España	2200	15	100 % de datos *scrapeados*
Uruguay	9391	4	100 % de datos *scrapeados*

Tabla 5.9: Comparación del valor de k óptimo para España y Uruguay usando todos los datos *scrapeados*

Fuente: Elaboración propia.

Según el resultado numérico arrojado fruto de calcular el k óptimo con la métrica *Silhouette Score*, podemos interpretar que España tiene más complejidad en sus datos, se necesita un k mayor para capturar la diversidad de patrones. Uruguay tiene menos diversidad, los datos se agrupan en pocas categorías muy claras.

5.4. Análisis de métrica Davies-Bouldin

La métrica Davies-Bouldin (ver2.3.4.3) minimiza el solapamiento entre *clusters*; es decir, el k óptimo corresponde al valor que permite agrupar la información de manera que los *clusters* sean lo más distintos posible.

En esta sección se calcula el índice Davies-Bouldin con el objetivo de determinar el valor óptimo de k para el conjunto de datos de la IDE de España bajo la categoría *Land Cover* y para el caso de Uruguay bajo la categoría Cobertura de la Tierra con mapas básicos e imágenes.

El cálculo del índice Davies-Bouldin se enmarca en un estudio que busca de forma persistente lograr una comprensión de la naturaleza de los datos. En su esencia, se busca alcanzar un nivel de entendimiento que arroje resultados relevantes acerca de la diversidad de los datos. Por lo tanto, si bien se desarrollan análisis para otras métricas (ver 5.5, 5.3 y 5.6), el análisis debe abordarse desde distintos niveles con el fin de obtener resultados significativos. El nivel que brinda el cálculo del índice Davies-Bouldin se fundamenta en la evaluación de la separación entre *clusters*, asegurando que los subtemas dentro de una categoría se distingan de manera clara.

El análisis se enmarca en distintos escenarios. En primer lugar, se presenta un análisis evolutivo de la métrica Davies-Bouldin con el fin de lograr comprender cómo evoluciona la métrica a medida que se obtiene un mayor volumen de datos (ver 5.4.1). El análisis evolutivo arroja resultados de interés en un contexto de índole comparativo e investigativo, justificando el desarrollo del análisis de la tasa de cambio del índice Davies-Bouldin (ver 5.4.2). Por último, se presenta un análisis en términos comparativos del índice Davies-Bouldin articulado en la subsección 5.4.3.

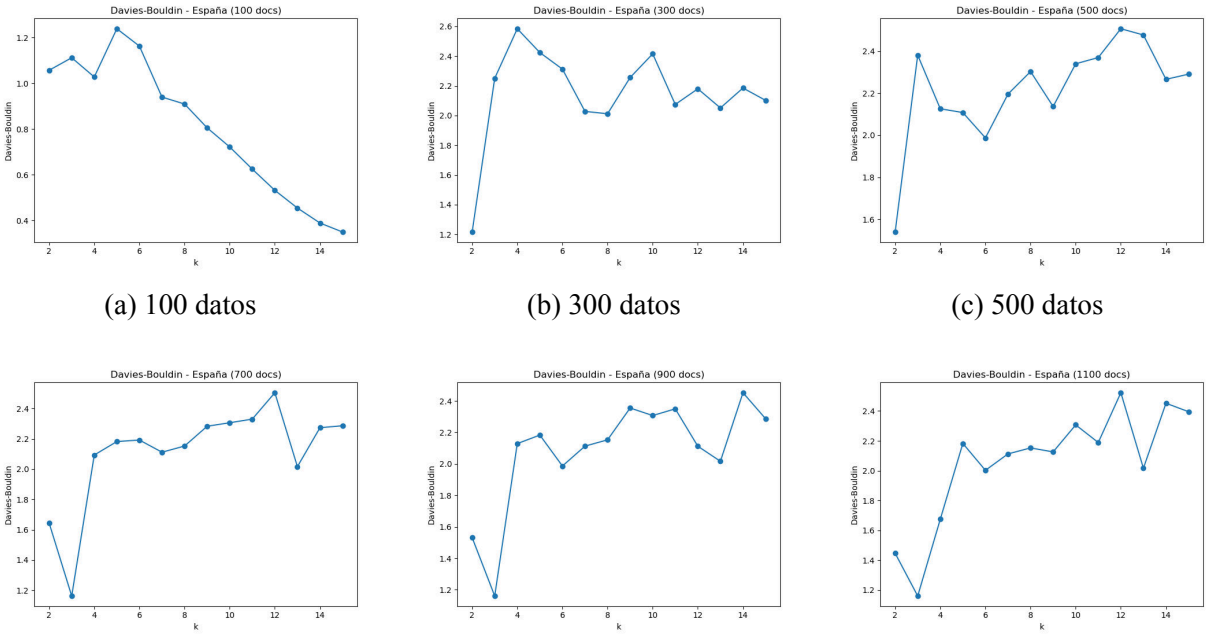
5.4.1. Análisis evolutivo de métrica Davies-Bouldin mediante el *scraping* progresivo de los datos

En la presente subsección se muestra la evolución de la métrica Davies-Bouldin mediante el *scraping* progresivo de los datos, así como el k óptimo a medida que se incrementa la muestra. Esta evolución es presentada para las IDEs de España (ver 5.4.1.1) y Uruguay (ver 5.4.1.2). El objetivo de desarrollar el análisis a nivel individual es comparar la evolución de ambos conjuntos de datos (ver 5.4.1.3).

5.4.1.1. IDE España

En líneas que prosiguen se expone el análisis evolutivo para el caso de la IDE de España.

En las Figuras 5.14 y 5.15 se muestra la evolución del Davies-Bouldin a medida que se incrementa progresivamente el volumen de datos obtenido mediante *scraping*.

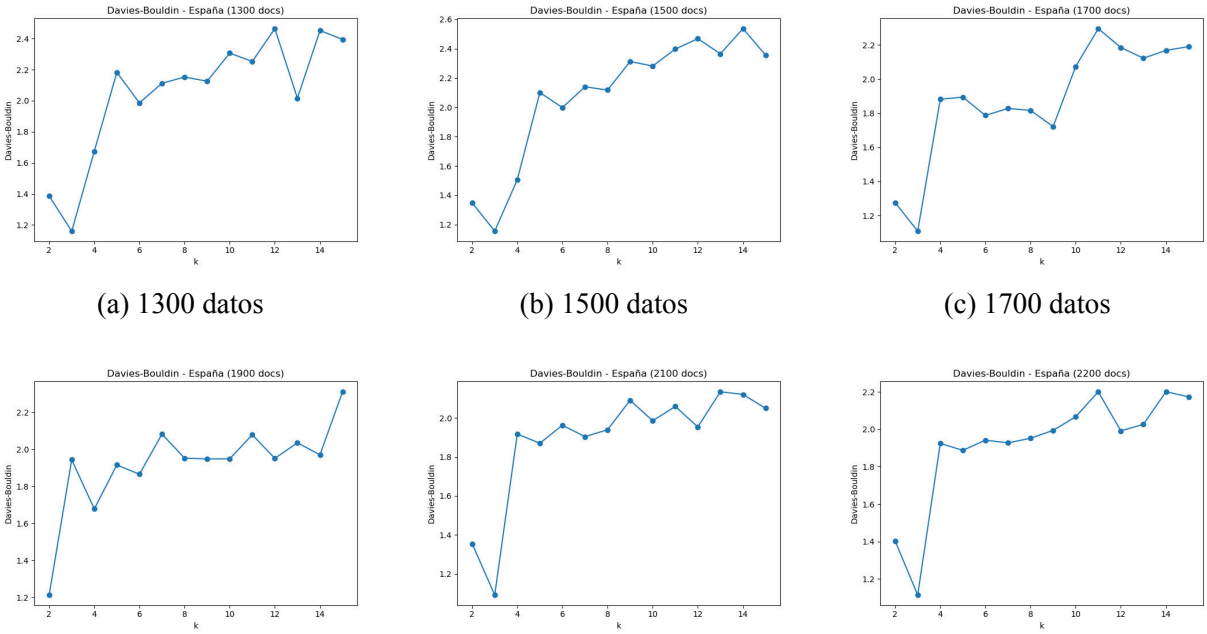


(a) 100 datos (b) 300 datos (c) 500 datos

(d) 700 datos (e) 900 datos (f) 1100 datos

Figura 5.14: Estimación de la cantidad de *clusters* de la IDE de España con la métrica Davies Bouldin variando el tamaño de la muestra(100-1100 datos)

Fuente: Elaboración propia.



(a) 1300 datos (b) 1500 datos (c) 1700 datos

(d) 1900 datos (e) 2100 datos (f) 2200 datos

Figura 5.15: Estimación de la cantidad de *clusters* de la IDE de España con la métrica Davies Bouldin variando el tamaño de la muestra(1300-2200 datos)

Fuente: Elaboración propia.

En la Tabla 5.10 se muestra la evolución del valor Davies-Bouldin en función de la cantidad de datos de España con su respectiva interpretación de su estructura ilustrada en la Figura 5.16 y basada en la Tabla 2.2. En suma, se proporciona una Figura para profundizar en el entendimiento de la evolución que detalla el significado de los símbolos (ver 5.17). Es pertinente destacar que los decimales se encuentran redondeados a cuatro cifras con motivos de simplificar cálculos y visualización del índice. No obstante, esta línea debe analizarse a futuro (ver Capítulo 8). En los Anexos se encuentra disponible la Tabla completa (ver Tabla A.4) para poder visualizarla y poder desarrollar futuros cálculos.

País	Cantidad de datos	*k* óptimo	Davies-Bouldin
España	100	15	0.3495(=)
	300	2(↓)	1.2161(↑)
	500	2(=)	1.5395(↑)
	700	3(↑)	1.1613(↓)
	900	3(=)	1.1605(↓)
	1100	3(=)	1.1602(↓)
	1300	3(=)	1.1601(↓)
	1500	3(=)	1.1552(↓)
	1700	3(=)	1.1085(↓)
	1900	2(↓)	1.2125(↑)
	2100	3(↑)	1.0919(↓)
	2200	3(=)	1.1140(↑)

Tabla 5.10: Evolución del índice Davies-Bouldin en función de la cantidad de documentos de la IDE de España

Fuente : Elaboración propia.

Bueno: *clusters* compactos y bien separados (DBI <1)

Distribución de datos deficiente: *clusters* dispersos o con superposición significativa (DBI $\geq$ 1)

Figura 5.16: Interpretación de colores de la Tabla 5.10

Fuente: Elaboración propia.

↑ Incremento respecto al valor anterior ↓ Decremento respecto al valor anterior = Valor igual al anterior

Figura 5.17: Interpretación de símbolos de la Tabla 5.10

Fuente : Elaboración propia.

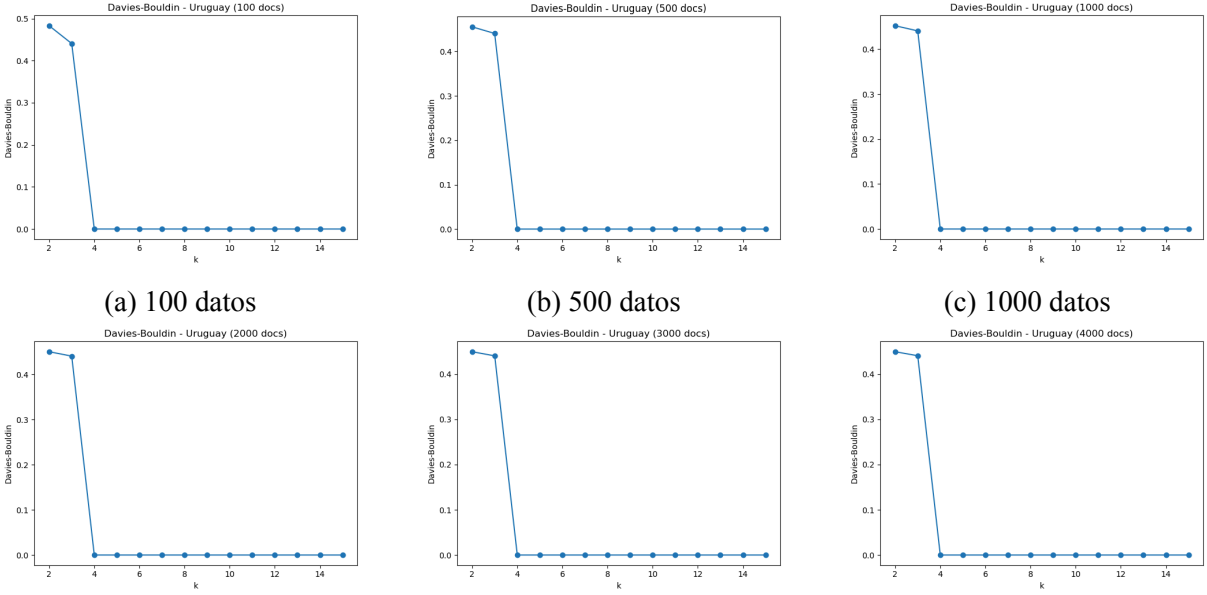
Un aspecto llamativo en la evolución del índice Davies-Bouldin es que, con una muestra reducida de 100 metadatos, se requiere un número elevado de *clusters* (k=15) para lograr una buena separación temática, lo que sugiere que los datos iniciales están agrupados en subtemas bien diferenciados. Sin embargo, al aumentar la cantidad de datos mediante *scraping*, el valor del

índice Davies-Bouldin tiende a subir y el número óptimo de *clusters* disminuye (k≈3), lo que indica que los nuevos datos incorporados presentan una mayor superposición temática o que los grupos principales se consolidan. Esta evolución revela que el *scraping* progresivo permite capturar mejor la estructura global de los datos, mostrando cómo los subtemas se agrupan en categorías más amplias que no son evidentes en muestras pequeñas.

5.4.1.2. IDE Uruguay

En líneas posteriores, se expone e interpreta la evolución del índice Davies-Bouldin a medida que el volumen de datos se acrecienta.

En las Figuras 5.18 y 5.19 se muestra la evolución del Davies-Bouldin a medida que se incrementa progresivamente el volumen de datos obtenido mediante *scraping*.

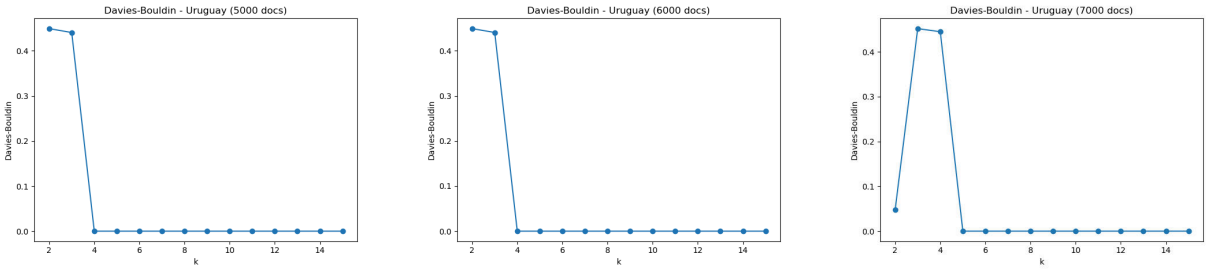


(a) 100 datos (b) 500 datos (c) 1000 datos

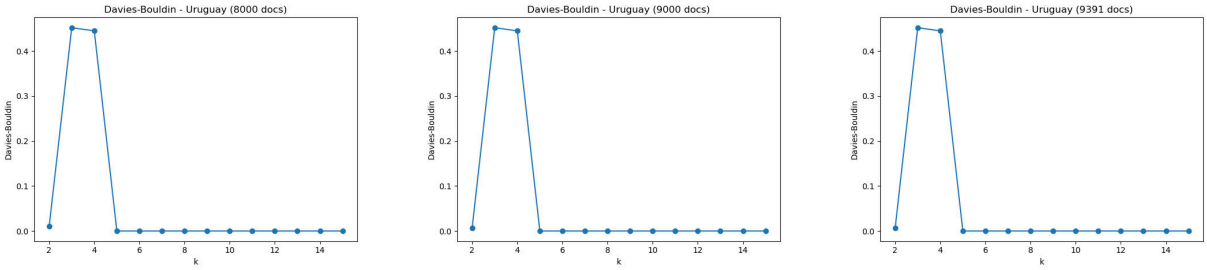
(d) 2000 datos (e) 3000 datos (f) 4000 datos

Figura 5.18: Evolución del Davies-Bouldin Index (DBI) para diferentes volúmenes de datos de la IDE de Uruguay

Fuente: Elaboración propia.



(a) 5000 datos (b) 6000 datos (c) 7000 datos



(d) 8000 datos (e) 9000 datos (f) 9391 datos

Figura 5.19: Evolución del Davies-Bouldin Index (DBI) para diferentes volúmenes de datos de la IDE de Uruguay

Fuente: Elaboración propia.

En la Tabla 5.11 se muestra la evolución del valor Davies-Bouldin en función de la cantidad de datos de Uruguay con su respectiva interpretación (ver Tabla 2.2) basada en literatura especializada (ver 2.2). Los valores se presentan con cuatro decimales para conservar la precisión de la métrica, simplificando a la vez los cálculos y su visualización. En esta línea, cabe aclarar que la precisión debe ser explorada en futuras líneas de investigación (ver Capítulo 8). La Tabla completa se puede visualizar en los Anexos (ver A.3). En suma, en la Figura 5.21 se presenta interpretación de los símbolos

País	Cantidad de datos	*k* óptimo	Valor Davies-Bouldin
Uruguay	100	4	0.0000
	1000	4 (=)	0.0000(=)
	2000	4 (=)	0.0000(=)
	3000	4 (=)	0.0000(=)
	4000	4 (=)	0.0000(=)
	5000	4 (=)	0.0001(↑)
	6000	4 (=)	0.0001(=)
	7000	5 (↑)	0.0001(=)
	8000	5 (=)	0.0001(=)
	9000	5 (=)	0.0002(↑)
	9391	5 (=)	0.0002(=)

Tabla 5.11: Valores de Davies-Bouldin para Uruguay en función de la cantidad de datos

Fuente: Elaboración propia.

Bueno: *clusters* compactos y bien separados (DBI <1)

Distribución de datos deficiente: *clusters* dispersos o con superposición significativa (DBI ≥ 1)

Figura 5.20: Interpretación de colores de la Tabla 5.11

Fuente: Elaboración propia.

↑ Incremento respecto al valor anterior

↓ Decremento respecto al valor anterior

= Valor igual al anterior

Figura 5.21: Interpretación de símbolos de la Tabla 5.11

Fuente: Elaboración propia.

Los resultados obtenidos mediante el índice Davies-Bouldin (DBI) evidencian una estructura de *clustering* notablemente estable para los datos de la IDE de Uruguay. En la mayoría de los escenarios evaluados, desde subconjuntos de 100 hasta 6000 registros, el k óptimo se mantiene constante en cuatro *clusters*, con valores de DBI cercanos a cero, lo que sugiere grupos bien definidos y altamente separados. Al incrementar el volumen de datos a

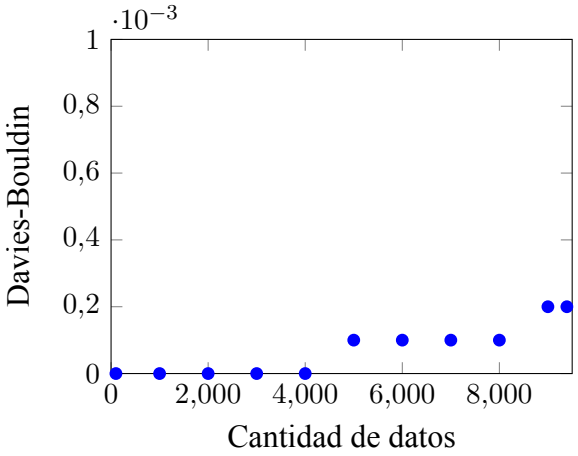
partir de los 7000 registros, se observan ligeras fluctuaciones en el k óptimo, sin embargo, los valores de DBI permanecen en el mismo orden de magnitud, indicando que dichas variaciones no representan cambios sustanciales en la calidad de la partición. Estas oscilaciones puntuales pueden atribuirse a la sensibilidad inherente del algoritmo de *clustering* frente a variaciones en la muestra y a la presencia de subgrupos marginales que se vuelven detectables únicamente cuando el volumen de datos alcanza su totalidad. En conjunto, estos hallazgos refuerzan la hipótesis de que la estructura subyacente del conjunto de datos está compuesta por pocos grupos característicos y que la incorporación progresiva de nuevos registros no revela patrones significativamente distintos, sino ajustes menores en la configuración fina.

5.4.1.3. Análisis comparativo de la evolución

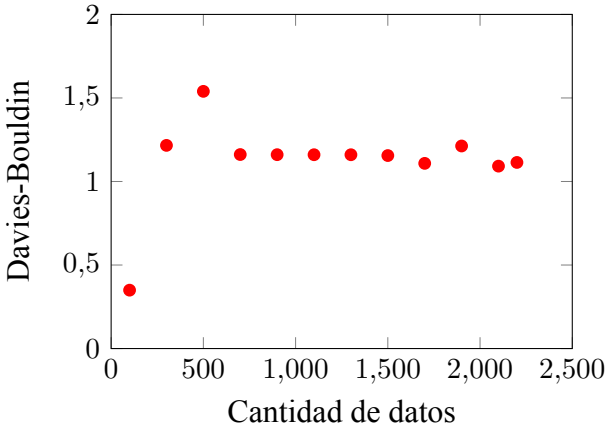
En párrafos posteriores, se expone un análisis comparativo cuyo fundamento argumentativo se basa en que una mayor variabilidad a medida que se *scrapean* más datos indica mayor diversidad en la información.

Como se observa en la Figura 5.22, el índice Davies-Bouldin mantiene valores casi constantes en los datos de la IDE Uruguay a medida que se incrementa la cantidad de documentos, lo que indica una estructura de *clusters* estable y homogénea. En contraste, para los datos de la IDE de España, el índice muestra variaciones más notorias conforme se *scrapean* más datos, lo que refleja una estructura más heterogénea y sensible al tamaño de la muestra.

De manera consistente, la evolución del número óptimo de *clusters* (k) mostrado en la Figura 5.23 evidencia esta diferencia: en Uruguay se mantiene prácticamente estable con leves ajustes, mientras que en España fluctúa con mayor frecuencia, reforzando la idea de que el aumento de datos introduce mayor variabilidad en la organización interna de la información.



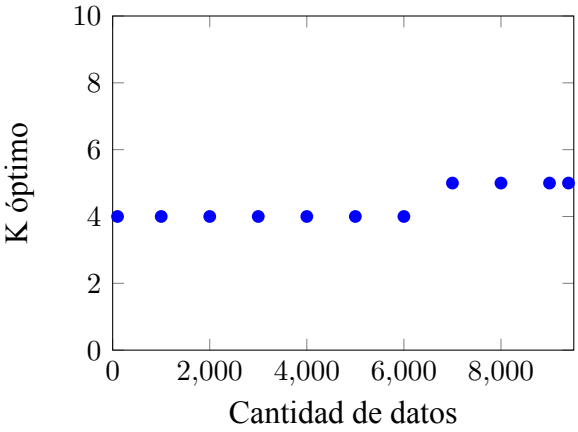
(a) Evolución del índice Davies-Bouldin en Uruguay



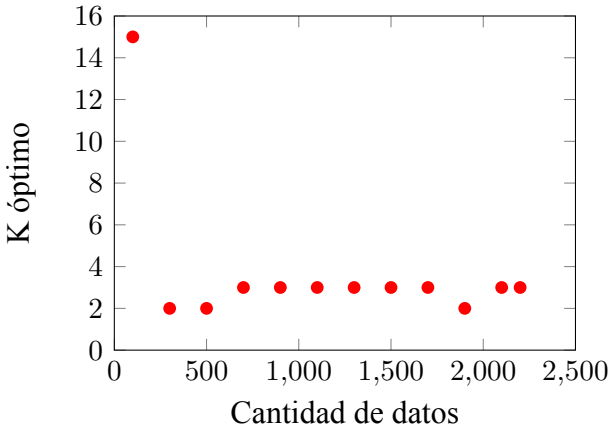
(b) Evolución del índice Davies-Bouldin en España

Figura 5.22: Comparación de la evolución del índice Davies-Bouldin para Uruguay y España

Fuente: Elaboración propia.



(a) Evolución del k óptimo en Uruguay.



(b) Evolución del k óptimo en España.

Figura 5.23: Comparación de la evolución del número de *clusters* óptimo (k) para la IDE de Uruguay y España

Fuente: Elaboración propia.

5.4.2. Análisis de la tasa de cambio de valor Davies-Bouldin

En la comparación realizada en el contexto del análisis evolutivo (ver 5.4.1.3), se observó que, en el caso de la IDE España, a medida que se incorporan más datos mediante el proceso de *scraping*, se registran mayores variaciones en el valor de Davies-Bouldin en comparación con el caso de la IDE de Uruguay.

En este sentido, el análisis de la evolución del Davies-Bouldin, a medida que aumenta la cantidad de datos extraídos ofrece resultados relevantes y plantea la necesidad de un estudio numérico más profundo. Dicho estudio permitirá sentar las bases para futuras investigaciones y avanzar hacia la automatización de la comparación, reduciendo la dependencia del trabajo manual del investigador (ver Capítulo 8). En suma, futuros estudios deben contemplar la precisión de las métricas con el objetivo de obtener resultados significativos (ver Capítulos 8).

Se estudiará la tasa de cambio para las IDEs de España (ver 5.4.2.1) y de Uruguay (ver 5.4.2.2). El objetivo de realizar el análisis de la tasa de cambio es presentar un análisis comparativo (ver 5.4.2.3) en consonancia con la línea de investigación tomada. En adición, otro de los objetivos implica sustentar de manera numérica los resultados obtenidos en líneas anteriores y sentar las bases para automatizar procesos manuales de comparación (ver Capítulo 8).

5.4.2.1. IDE España

Según la evolución presentada anteriormente (ver Tabla 5.10), se puede definir la función del Davies-Bouldin en base a la cantidad de datos, presentada en líneas siguientes.

Sea

$$f : \{100, 300, 500, 700, 900, 1100, 1300, 1500, 1700, 1900, 2100, 2200\} \to \mathbb{R}$$

la función que asigna a cada tamaño de muestra N su valor del índice Davies-Bouldin para los datos de la IDE de España:

$$f(N) = \begin{cases} 0{,}3495, & N = 100 \\ 1{,}2161, & N = 300 \\ 1{,}5395, & N = 500 \\ 1{,}1613, & N = 700 \\ 1{,}1605, & N = 900 \\ 1{,}1602, & N = 1100 \\ 1{,}1601, & N = 1300 \\ 1{,}1552, & N = 1500 \\ 1{,}1085, & N = 1700 \\ 1{,}2125, & N = 1900 \\ 1{,}0919, & N = 2100 \\ 1{,}1140, & N = 2200 \\ \text{no definido}, & \text{si } N \notin \{100, 300, \ldots, 2200\} \end{cases}$$

Como f está definida solo en puntos discretos, aproximamos la derivada mediante diferencias finitas hacia adelante:

$$\frac{\Delta S_i}{\Delta N_i} = \frac{f(N_i) - f(N_{i-1})}{N_i - N_{i-1}}, \qquad i = 2, \ldots, 12.$$

Cálculos explícitos:

$$\Delta S_2 = f(300) - f(100) = 1{,}2161 - 0{,}3495 = 0{,}8666, \quad \frac{\Delta S_2}{\Delta N_2} = \frac{0{,}8666}{200} = 4{,}333 \times 10^{-3}$$

$$\Delta S_3 = f(500) - f(300) = 1{,}5395 - 1{,}2161 = 0{,}3234, \quad \frac{\Delta S_3}{\Delta N_3} = \frac{0{,}3234}{200} = 1{,}617 \times 10^{-3}$$

$$\Delta S_4 = f(700) - f(500) = 1{,}1613 - 1{,}5395 = -0{,}3782, \quad \frac{\Delta S_4}{\Delta N_4} = \frac{-0{,}3782}{200} = -1{,}891 \times 10^{-3}$$

$$\Delta S_6 = f(1100) - f(900) = 1{,}1602 - 1{,}1605 = -0{,}0003, \quad \frac{\Delta S_6}{\Delta N_6} = \frac{-0{,}0003}{200} = -1{,}50 \times 10^{-6}$$

$$\Delta S_7 = f(1300) - f(1100) = 1{,}1601 - 1{,}1602 = -0{,}0001, \quad \frac{\Delta S_7}{\Delta N_7} = \frac{-0{,}0001}{200} = -5{,}00 \times 10^{-7}$$

$$\Delta S_8 = f(1500) - f(1300) = 1{,}1552 - 1{,}1601 = -0{,}0049, \quad \frac{\Delta S_8}{\Delta N_8} = \frac{-0{,}0049}{200} = -2{,}45 \times 10^{-5}$$

$$\Delta S_9 = f(1700) - f(1500) = 1{,}1085 - 1{,}1552 = -0{,}0467, \quad \frac{\Delta S_9}{\Delta N_9} = \frac{-0{,}0467}{200} = -2{,}335 \times 10^{-4}$$

$$\Delta S_{10} = f(1900) - f(1700) = 1{,}2125 - 1{,}1085 = 0{,}1040, \quad \frac{\Delta S_{10}}{\Delta N_{10}} = \frac{0{,}1040}{200} = 5{,}20 \times 10^{-4}$$

$$\Delta S_{11} = f(2100) - f(1900) = 1{,}0919 - 1{,}2125 = -0{,}1206, \quad \frac{\Delta S_{11}}{\Delta N_{11}} = \frac{-0{,}1206}{200} = -6{,}03 \times 10^{-4}$$

$$\Delta S_{12} = f(2200) - f(2100) = 1{,}1140 - 1{,}0919 = 0{,}0221, \quad \frac{\Delta S_{12}}{\Delta N_{12}} = \frac{0{,}0221}{100} = 2{,}21 \times 10^{-4}$$

En la Tabla 5.12 se muestra el resultado de los cálculos obtenidos y en la Figura 5.24 se ilustra la gráfica correspondiente al cálculo de la tasa de cambio.

N	ΔS	ΔN	$\Delta S/\Delta N$
100	-	-	-
300	0.8666	200	$4{,}33 \times 10^{-3}$
500	0.3234	200	$1{,}62 \times 10^{-3}$
700	-0.3782	200	$-1{,}89 \times 10^{-3}$
900	-0.0008	200	$-4{,}00 \times 10^{-6}$
1100	-0.0003	200	$-1{,}50 \times 10^{-6}$
1300	-0.0001	200	$-5{,}00 \times 10^{-7}$
1500	-0.0049	200	$-2{,}45 \times 10^{-5}$
1700	-0.0467	200	$-2{,}335 \times 10^{-4}$
1900	0.1040	200	$5{,}20 \times 10^{-4}$
2100	-0.1206	200	$-6{,}03 \times 10^{-4}$
2200	0.0221	100	$2{,}21 \times 10^{-4}$

Tabla 5.12: Derivadas discretas del índice Davies-Bouldin para los datos de Uruguay

Fuente: Elaboración propia.

Observación. Los intervalos con mayor cambio por unidad de N (en valor absoluto) son 100→300, 1700→1900 y 1900→2100, que indican saltos importantes en la calidad relativa de los *clusters* en esos rangos de tamaño de muestra. Entre 700 y 1300 el índice muestra variaciones muy pequeñas (orden 10^{-6}–10^{-5}), indicando estabilidad local.

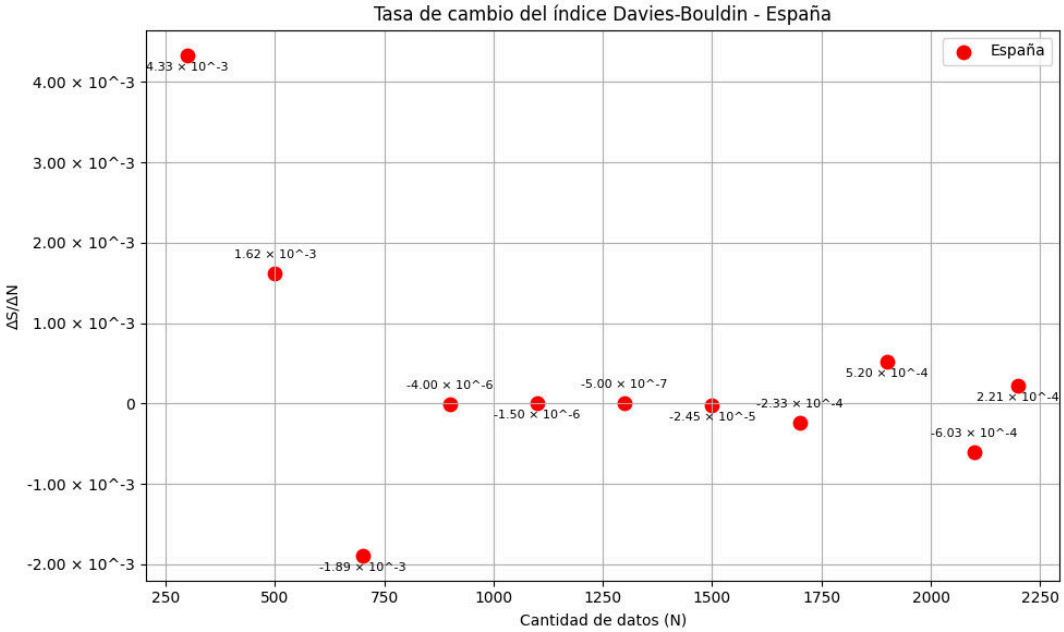


Figura 5.24: Tasa de cambio del valor Davies-Bouldin(España)

Fuente: Elaboración propia.

5.4.2.2. IDE Uruguay

Según la evolución presentada anteriormente (ver Tabla 5.11), se puede definir la función del Davies-Bouldin en base a la cantidad de datos. Esta definición es presentada en líneas siguientes.

Sea $f : \{100, 1000, 2000, \ldots, 9391\} \to \mathbb{R}$ la función que asigna a cada tamaño de datos N su *Davies Bouldin Index* en Uruguay:

$$f(N) = \begin{cases} 0{,}0, & N \in \{100, 1000, 2000, 3000, 4000\} \\ 0{,}0001, & N \in \{5000, 6000, 7000, 8000\} \\ 0{,}0002, & N \in \{9000, 9391\} \\ \text{no definido}, & N \notin \{100, 1000, 2000, \ldots, 9391\} \end{cases}$$

Como f está definida solo en puntos discretos, se aproxima la derivada mediante diferencias finitas hacia adelante:

$$\frac{\Delta S_i}{\Delta N_i} = \frac{f(N_i) - f(N_{i-1})}{N_i - N_{i-1}}, \quad i = 2, \ldots, 11$$

Cálculos explícitos:

$$\Delta S_2 = f(1000) - f(100) = 0{,}0 - 0{,}0 = 0{,}0, \qquad \Delta N_2 = 900, \qquad \frac{\Delta S_2}{\Delta N_2} = 0{,}0$$

$$\Delta S_3 = f(2000) - f(1000) = 0{,}0 - 0{,}0 = 0{,}0, \qquad \Delta N_3 = 1000, \qquad \frac{\Delta S_3}{\Delta N_3} = 0{,}0$$

$$\Delta S_4 = f(3000) - f(2000) = 0{,}0 - 0{,}0 = 0{,}0, \qquad \Delta N_4 = 1000, \qquad \frac{\Delta S_4}{\Delta N_4} = 0{,}0$$

$$\Delta S_5 = f(4000) - f(3000) = 0{,}0 - 0{,}0 = 0{,}0, \qquad \Delta N_5 = 1000, \qquad \frac{\Delta S_5}{\Delta N_5} = 0{,}0$$

$$\Delta S_6 = f(5000) - f(4000) = 0{,}0001 - 0{,}0 = 0{,}0001, \qquad \Delta N_6 = 1000, \qquad \frac{\Delta S_6}{\Delta N_6} = 1{,}0 \cdot 10^{-7}$$

$$\Delta S_7 = f(6000) - f(5000) = 0{,}0001 - 0{,}0001 = 0{,}0, \qquad \Delta N_7 = 1000, \qquad \frac{\Delta S_7}{\Delta N_7} = 0{,}0$$

$$\Delta S_8 = f(7000) - f(6000) = 0{,}0001 - 0{,}0001 = 0{,}0, \qquad \Delta N_8 = 1000, \qquad \frac{\Delta S_8}{\Delta N_8} = 0{,}0$$

$$\Delta S_9 = f(8000) - f(7000) = 0{,}0001 - 0{,}0001 = 0{,}0, \qquad \Delta N_9 = 1000, \qquad \frac{\Delta S_9}{\Delta N_9} = 0{,}0$$

$$\Delta S_{10} = f(9000) - f(8000) = 0{,}0002 - 0{,}0001 = 0{,}0001, \quad \Delta N_{10} = 1000, \quad \frac{\Delta S_{10}}{\Delta N_{10}} = 1{,}0 \cdot 10^{-7}$$

$$\Delta S_{11} = f(9391) - f(9000) = 0{,}0002 - 0{,}0002 = 0{,}0, \qquad \Delta N_{11} = 391, \qquad \frac{\Delta S_{11}}{\Delta N_{11}} = 0{,}0$$

Observación: La derivada discreta muestra que la mayor variación ocurre entre 4000 y 5000 datos y entre 8000 y 9000, indicando que el índice Davies-Bouldin mejora ligeramente con más datos, pero después se estabiliza.

En la Tabla 5.13 se ilustra el resultado de calcular la derivada discreta del valor del índice de Davies-Bouldin para el caso de la IDE de Uruguay y en la la Figura 5.25 se ilustra a nivel gráfico.

N	ΔS	ΔN	$\Delta S/\Delta N$
100	-	-	-
1000	0.0	900	0.0
2000	0.0	1000	0.0
3000	0.0	1000	0.0
4000	0.0	1000	0.0
5000	0.0001	1000	$1{,}0 \cdot 10^{-7}$
6000	0.0	1000	0.0
7000	0.0	1000	0.0
8000	0.0	1000	0.0
9000	0.0001	1000	$1{,}0 \cdot 10^{-7}$
9391	0.0	391	0.0

Tabla 5.13: Derivadas discretas aproximadas del índice Davies-Bouldin para Uruguay según la función $f(N)$

Fuente: Elaboración propia.

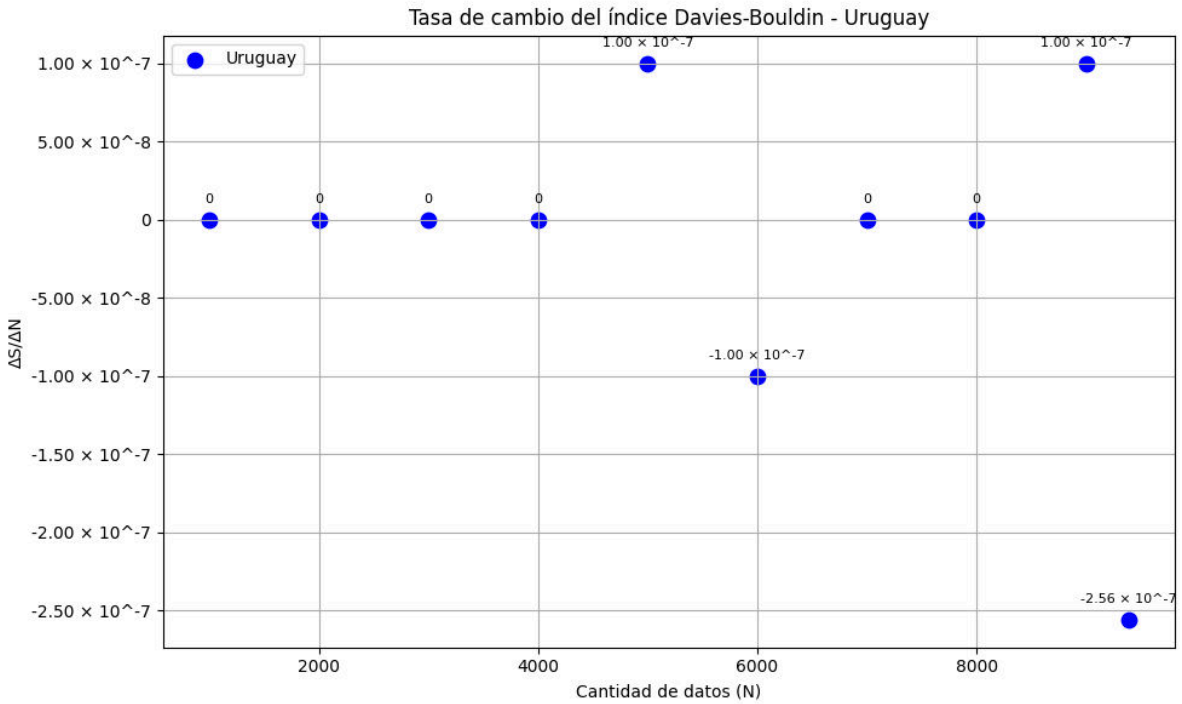


Figura 5.25: Tasa de cambio del valor Davies-Bouldin(Uruguay)

Fuente: Elaboración propia.

5.4.2.3. Análisis comparativo de la tasa de cambio

Al analizar las tasas de cambio del índice Davies-Bouldin (Tablas 5.13 y 5.12), se observa que en Uruguay las variaciones son prácticamente nulas, indicando que los *clusters* se mantienen estables a medida que se incrementa la muestra. Los datos de la IDE de España, en cambio, registran cambios positivos y negativos más marcados, lo que evidencia fluctuaciones en la cohesión y separación de los *clusters* al agregar nuevos datos. Esto sugiere que los *clusters* de los datos de la IDE de Uruguay son más robustos frente a la incorporación de información, mientras que en el caso de España se necesita un mayor volumen de datos para alcanzar una configuración estable.

5.4.3. Análisis comparativo del índice

En la Tabla 5.14 se presenta el k óptimo y el valor de Davies-Bouldin utilizando todos los datos *scrapeados* para ambas IDEs (España y Uruguay). En suma, en la Figura 5.26 se presenta la interpretación de los colores utilizados, seguido de una interpretación textual con las principales observaciones realizadas.

País	Cantidad de datos	K óptimo	Valor Davies-Bouldin	Datos utilizados
España	2200	3	1.1140	100 % de datos *scrapeados*
Uruguay	9391	5	0.0002	100 % de datos *scrapeados*

Tabla 5.14: Comparación del número óptimo de *clusters* (k) y del valor Davies-Bouldin entre España y Uruguay utilizando todos los datos *scrapeados*

Fuente: Elaboración propia.

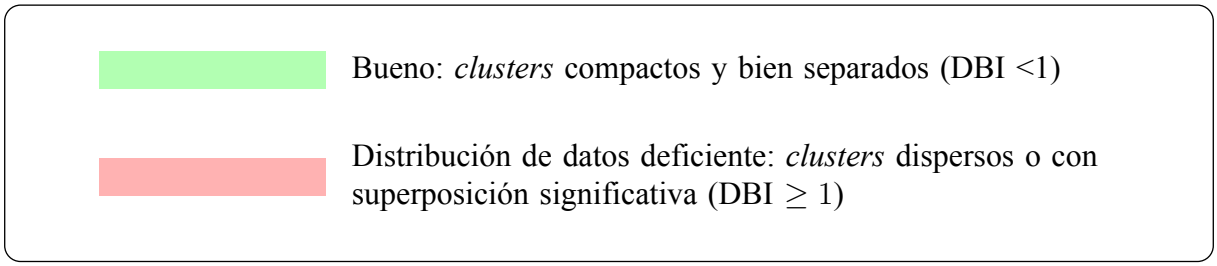


Figura 5.26: Interpretación de colores de la Tabla 5.14

Fuente: Elaboración propia.

Según la métrica Davies-Bouldin y basado en la Tabla 2.2 extraída de literatura especializada, la estructura resultante de calcular el índice Davies-Bouldin para el caso de Uruguay es buena con *clusters* compactos y bien separados. Sin embargo, para el caso de los datos de la IDE de España presentan distribución deficiente, caracterizándose por *clusters* dispersos o con superposición significativa. Esto evidencia diferencias en la estructura de datos y la cohesión de los *clusters* entre ambos países.

5.5. Análisis de métrica Calinski-Harabasz

En esta sección se calcula el índice Calinski-Harabasz con el objetivo de determinar el valor óptimo de k para el conjunto de datos de la IDE de España bajo la categoría *Land Cover* y para el caso de Uruguay bajo la categoría Cobertura de la Tierra con mapas básicos e imágenes.

En un contexto de análisis comparativo, cuyo objeto de estudio es la relación entre los datos, la métrica Calinski-Harabasz sirve como apoyo para responder a la pregunta ¿cuál es la configuración que maximiza la separación entre *clusters* y la compactación interna?. A través de un análisis evolutivo de la métrica Calinski-Harabasz mediante el *scraping* progresivo de los datos (ver 5.5.1) se busca dar respuesta a esta interrogante.

En suma, se pretende estudiar si la métrica Calinski-Harabasz puede utilizarse para extraer conclusiones en un escenario comparativo. La literatura (ver 2.3.4.3) indica que el índice Calinski-Harabasz es relativo y depende de la naturaleza de los datos; por ello, sus valores no permiten conclusiones absolutas. Sin embargo, con fines exploratorios y meramente por curiosidad se procede a su cálculo, ya que los resultados pueden proporcionar hallazgos interesantes. Esto se discute en el análisis y discusión final (ver 5.5.2).

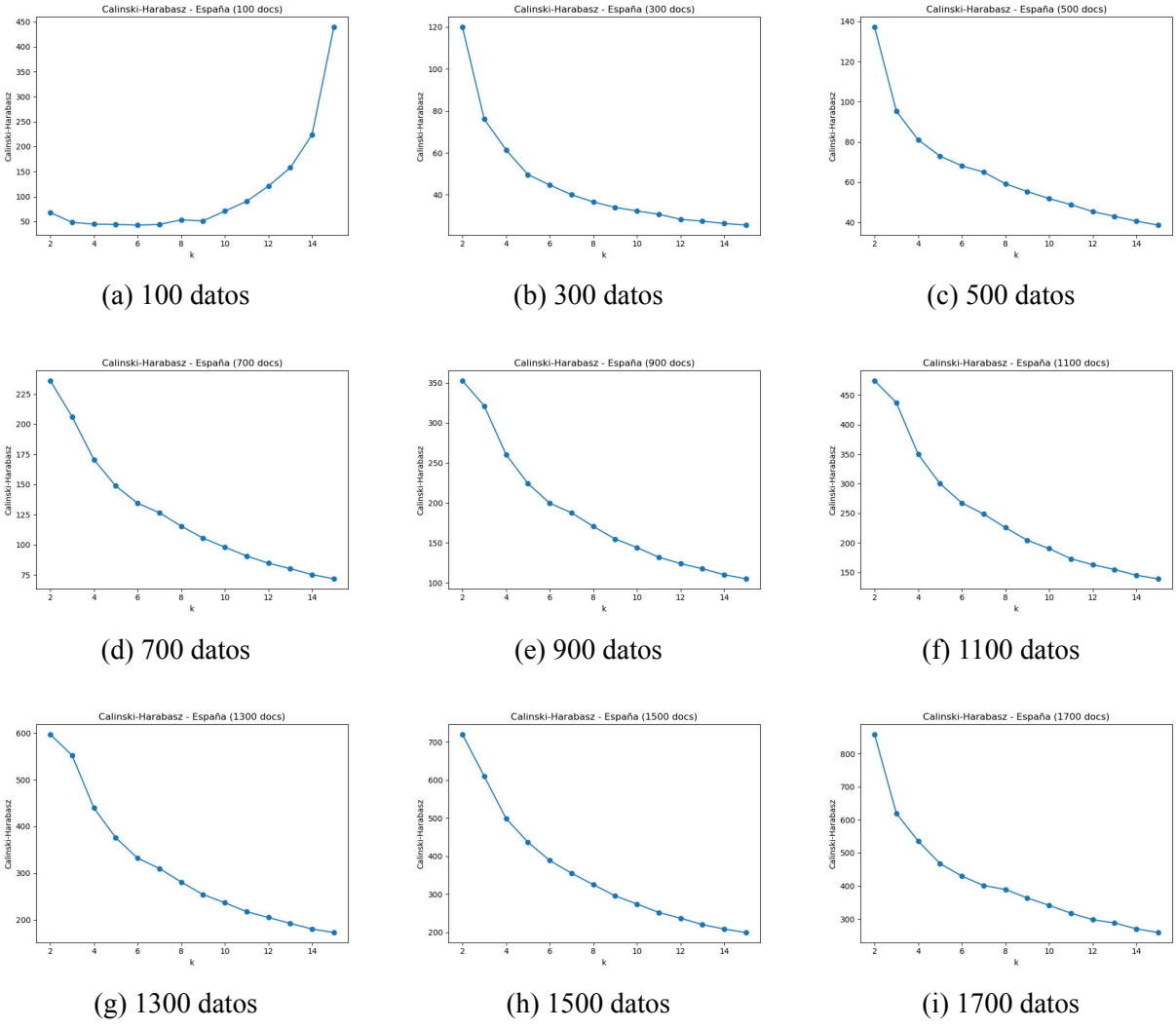
5.5.1. Análisis evolutivo de métrica Calinski-Harabasz mediante el *scraping* progresivo de los datos

A continuación, se muestra la evolución de la métrica Calinski-Harabasz mediante el *scraping* progresivo de los datos, así como el k óptimo a medida que se incrementa la muestra. Este análisis se realiza con el conjunto de datos extraídos de las IDEs de España (ver 5.5.1.1) y

Uruguay (ver 5.5.1.2). Posteriormente, se expone un análisis comparativo (ver 5.5.1.3).

5.5.1.1. IDE España

En la Figuras 5.27 y 5.28 se muestra la evolución de el índice Calinski-Harabasz a medida que se incrementa progresivamente el volumen de datos obtenido mediante *scraping*.



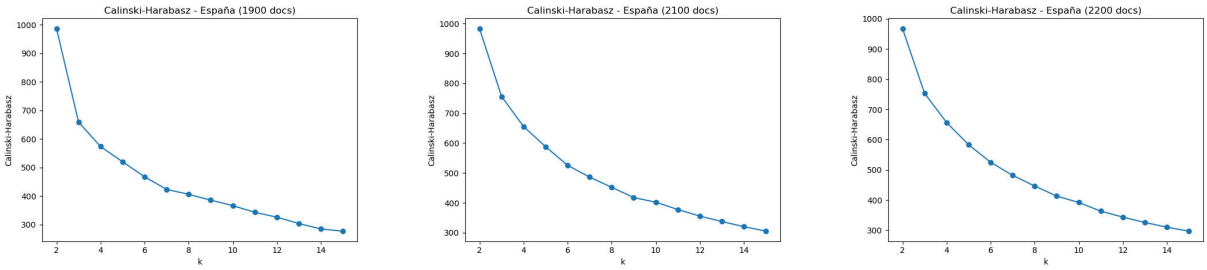
(a) 100 datos (b) 300 datos (c) 500 datos

(d) 700 datos (e) 900 datos (f) 1100 datos

(g) 1300 datos (h) 1500 datos (i) 1700 datos

Figura 5.27: Estimación de la cantidad de *clusters* (k) de la IDE de España con la métrica Calinski-Harabasz variando el tamaño de la muestra(100-1700 datos)

Fuente: Elaboración propia.



(a) 1900 datos (b) 2100 datos (c) 2200 datos

Figura 5.28: Estimación de la cantidad de *clusters* (k) de la IDE de España con la métrica Calinski-Harabasz variando el tamaño de la muestra(1900-2200 datos)

Fuente: Elaboración propia.

En la Tabla 5.15 se muestra la evolución del valor Calinski-Harabasz en función de la cantidad de datos de la IDE de España con su interpretación ilustrada en la Figura 5.29. En este caso, a diferencia del análisis realizado para la métrica *Silhouette Score* (véase sección 5.3), se opta por redondear los valores a cuatro decimales mostrado en notación científica para mejorar la visualización de la información, ya que esta métrica genera valores de alta magnitud. La Tabla completa se puede ver en Anexos (ver Tabla A.6). Esta decisión metodológica se fundamenta en que los valores obtenidos por esta métrica suelen ser de gran magnitud, por lo que una precisión superior no aporta información relevante adicional. Asimismo, esta práctica se consolida como una lección aprendida dentro del proceso de análisis (véase Capítulo 7).

País	**Cantidad de documentos**	**Número de *clusters***	**Calinski-Harabasz**
España	100	15	$4{,}3941 \times 10^2$ ($\uparrow$)
	300	2 ($\downarrow$)	$1{,}2000 \times 10^2$ ($\downarrow$)
	500	2 ($=$)	$1{,}3754 \times 10^2$ ($\uparrow$)
	700	2 ($=$)	$2{,}3660 \times 10^2$ ($\uparrow$)
	900	2 ($=$)	$3{,}5365 \times 10^2$ ($\uparrow$)
	1100	2 ($=$)	$4{,}7501 \times 10^2$ ($\uparrow$)
	1300	2 ($=$)	$5{,}9808 \times 10^2$ ($\uparrow$)
	1500	2 ($=$)	$7{,}2106 \times 10^2$ ($\uparrow$)
	1700	2 ($=$)	$8{,}6271 \times 10^2$ ($\uparrow$)
	1900	2 ($=$)	$9{,}9955 \times 10^2$ ($\uparrow$)
	2100	2 ($=$)	$9{,}9597 \times 10^2$ ($\downarrow$)
	2200	2 ($=$)	$9{,}7861 \times 10^2$ ($\downarrow$)

Tabla 5.15: Evolución del índice de Calinski-Harabasz para los datos de la IDE de España

Fuente: Elaboración propia.

↑ Incremento respecto al valor anterior ↓ Decremento respecto al valor anterior = Valor igual al anterior

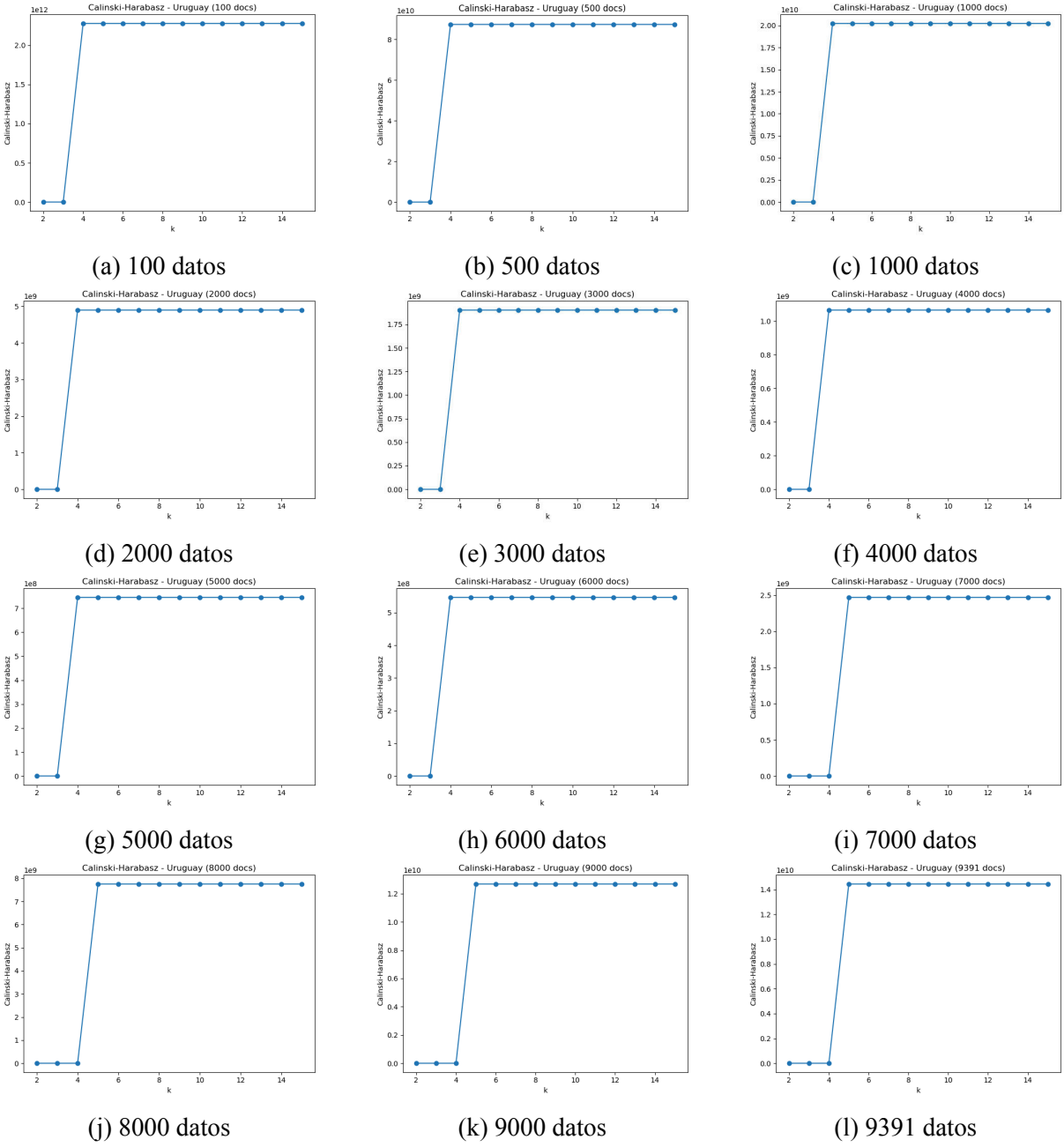
Figura 5.29: Interpretación de símbolos de la Tabla 5.15

Fuente: Elaboración propia.

La evolución del índice Calinski-Harabasz mostrada en la Tabla 5.15 revela un comportamiento similar al observado en otras métricas (ver 5.3, 5.4 y 5.6). Se aprecia que con una muestra reducida de 100 metadatos, se requiere un número elevado de *clusters* (k=15) para lograr una buena separación entre *clusters* y compactación interna. Por esto, la primer gráfica (ver Figura 5.27a) difiere significativamente de las demás Figuras (ver Figura 5.28). Sin embargo, al aumentar la cantidad de datos mediante *scraping*, el número óptimo de *clusters* disminuye (k≈2). Esta tendencia sugiere que los datos adicionales permiten consolidar temáticas dispersas en agrupamientos más definidos, lo que no es posible detectar con muestras pequeñas. Por lo tanto, el *scraping* progresivo es esencial para revelar la estructura global de los metadatos y mejorar la calidad del análisis.

5.5.1.2. IDE Uruguay

En la Figura 5.30 se muestra la evolución de el índice Calinski-Harabasz a medida que se incrementa progresivamente el volumen de datos obtenido mediante *scraping*.



(a) 100 datos (b) 500 datos (c) 1000 datos

(d) 2000 datos (e) 3000 datos (f) 4000 datos

(g) 5000 datos (h) 6000 datos (i) 7000 datos

(j) 8000 datos (k) 9000 datos (l) 9391 datos

Figura 5.30: Evolución de métrica Calinski-Harabasz en función de la cantidad de datos de la IDE de Uruguay

Fuente: Elaboración propia.

En la Tabla 5.16 se muestra la evolución del valor Calinski-Harabasz en función de la cantidad de datos de Uruguay en notación científica redondeado a cuatro decimales, debido a la naturaleza de la métrica que produce valores de alta magnitud. En los anexos (ver Tabla A.5) se muestra la evolución del valor Calinski-Harabasz en función de la cantidad de datos de

Uruguay con los valores sin redondear. Por otro lado, en la Figura 5.31 se puede ver la interpretación de los símbolos de la Tabla 5.16.

País	Cantidad de documentos	Número de *clusters*	Calinski-Harabasz
Uruguay	100	4	$2{,}1435 \times 10^{12}$
	1000	4 (=)	$2{,}04347 \times 10^{10}$ (↓)
	2000	4 (=)	$4{,}78859 \times 10^{9}$ (↓)
	3000	4 (=)	$1{,}89869 \times 10^{9}$ (↓)
	4000	4 (=)	$1{,}065470 \times 10^{9}$ (↓)
	5000	4 (=)	$7{,}4570 \times 10^{8}$ (↓)
	6000	4 (=)	$5{,}47627 \times 10^{8}$ (↓)
	7000	5 (↑)	$2{,}4751 \times 10^{9}$ (↑)
	8000	5 (=)	$7{,}7723 \times 10^{9}$ (↑)
	9000	5 (=)	$1{,}26958 \times 10^{10}$ (↑)
	9391	5 (=)	$1{,}4453 \times 10^{10}$ (↑)

Tabla 5.16: Evolución del índice de Calinski-Harabasz para los datos de la IDE de Uruguay

Fuente: Elaboración propia.

↑ Incremento respecto al valor anterior ↓ Decremento respecto al valor anterior = Valor igual al anterior

Figura 5.31: Interpretación de símbolos de la Tabla 5.16

Fuente: Elaboración propia.

Los valores del índice de Calinski–Harabasz (ver Tabla 5.16) permiten profundizar en el análisis de la estructura del conjunto de datos de Uruguay. Para volúmenes comprendidos entre 100 y 6000 documentos, el número óptimo de *clusters* se mantiene en cuatro, con una tendencia decreciente en los valores del índice, lo que refleja una disminución relativa en la compacidad de los *clusters* al incorporar más registros. A partir de los 7000 documentos, se observa un cambio hacia cinco *clusters*, acompañado de un aumento significativo en los valores del índice. Este comportamiento indica que el incremento en el volumen de datos permite identificar una subdivisión adicional que mejora la relación entre la cohesión interna de los grupos y su separación relativa. Dichos hallazgos, en consonancia con los resultados obtenidos mediante otras métricas (ver 5.3, 5.4 y 5.6) sugieren que la estructura subyacente del conjunto está formada por cuatro grupos principales, con la posibilidad de refinarse a cinco

clusters cuando el análisis incorpora la totalidad de los registros.

5.5.1.3. Análisis comparativo

En líneas que siguen se desarrolla el análisis comparativo de la evolución de la métrica Calinski-Harabasz.

En primer lugar, se puede observar que la evolución de los datos de la IDE de Uruguay (ver 5.5.1.2) presenta menos fluctuaciones que la evolución que presentan los datos extraídos de la IDE de España (ver 5.5.1.1). Por ejemplo, en la IDE de España, los primeros datos recolectados muestran una estructura que difiere notablemente de la que se consolida posteriormente. En cambio, en la IDE de Uruguay, la estructura parece mantenerse relativamente estable a medida que aumenta el volumen de datos. Si bien esta observación es válida, conviene ser cautelosa al interpretar los resultados. La literatura (ver 2.3.4.3) indica que el índice Calinski-Harabasz es relativo y depende de la naturaleza de los datos; por tanto, sus valores no permiten conclusiones absolutas. En términos evolutivos, solo podemos señalar que ambas IDEs presentan patrones distintos, sin profundizar más, consolidando como lección aprendida que es preferible contar con métricas absolutas para facilitar comparaciones (ver Capítulo 7).

5.5.2. Análisis y discusión final

> *'A fact in itself is nothing. It is valuable only for the idea attached to it, or for the proof which it furnishes.'*
>
> – Henri Poincaré. [160]

La metodología propuesta (ver 4.2) incluye una fase de evaluación (ver 4.2.5), cuyo objetivo es determinar si el análisis arroja resultados comparativos. En particular, se busca comprender si la evaluación del modelo (ver 4.2.4) permite extraer conclusiones sobre la estructura de los datos.

La Tabla 5.17 muestra la comparación del número óptimo de *clusters*. Se observa que para España el valor óptimo de k es 2, mientras que para Uruguay es 5. Asimismo, el valor del

índice Calinski-Harabasz es mayor en la IDE de Uruguay, lo que da indicios de la configuración de estructuras distintas. No obstante, como señaló Henri Poincaré [160], los hechos deben interpretarse según la idea detrás. Siguiendo esta línea, la interpretación de la métrica depende del contexto. El índice Calinski-Harabasz fluctua con la cantidad de datos y no produce conclusiones absolutas sobre la calidad de los *clusters*. Su utilidad radica en guiar futuras evaluaciones (ver Capítulo 8) y en la necesidad de considerar métricas comparables que permitan analizar la estructura de los datos, ya sean absolutas o relativas con un modelo de normalización adecuado. Por esta razón, los valores del índice no se comparan directamente entre las IDEs, dado que dependen fuertemente del tamaño de la muestra.

En definitiva, líneas futuras deben trabajar con métricas que produzcan o bien resultados numéricos absolutos o que permitan ser normalizados (ver Capítulo 8). Por tanto, la lección aprendida sirve como *input* para futuras líneas de investigación.

País	Cantidad de datos	k óptimo	Valor óptimo CH	Datos utilizados
España	2200	2	$9{,}7861 \times 10^2$	100 % de datos *scrapeados*
Uruguay	9391	5	$1{,}4453 \times 10^{10}$	100 % de datos *scrapeados*

Tabla 5.17: Comparación del número óptimo de *clusters* (k) y valor máximo del índice Calinski-Harabasz entre España y Uruguay usando todos los datos *scrapeados*

Fuente: Elaboración propia.

5.6. Análisis de métrica *Elbow Method*

En esta sección se calcula la métrica *Elbow Method* con el objetivo de determinar el valor óptimo de k para el conjunto de datos de la IDE de España bajo la categoría *Land Cover* y para el caso de Uruguay bajo la categoría Cobertura de la Tierra con mapas básicos e imágenes.

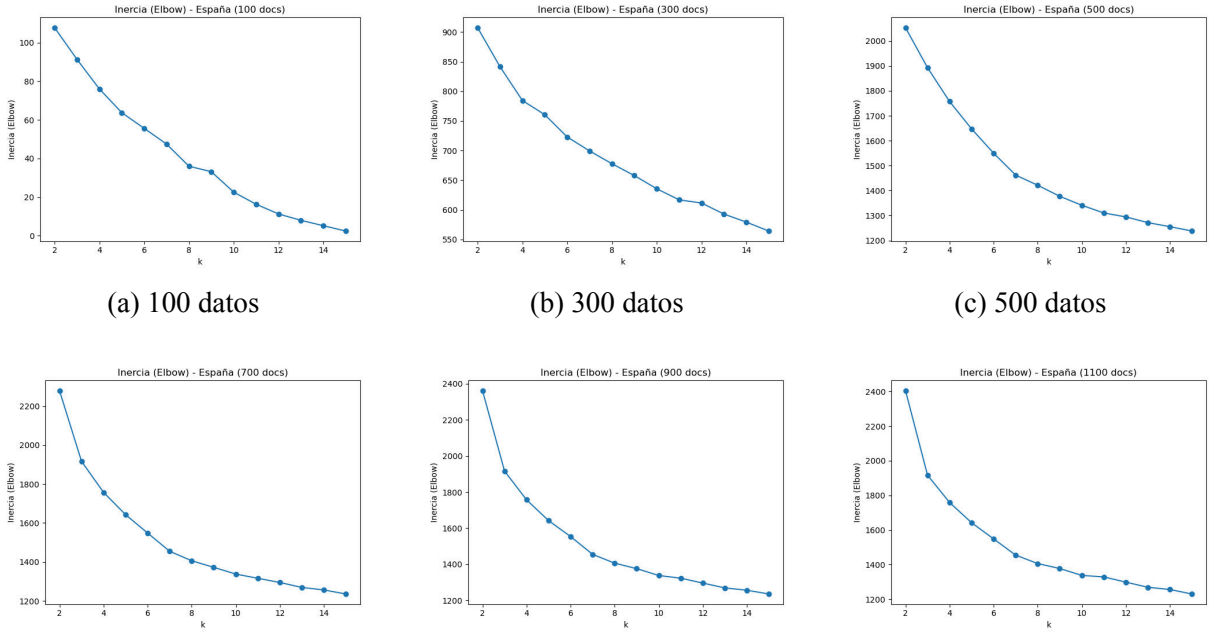
Se muestra el análisis de índole evolutivo de la métrica (ver 5.6.1). Por otro lado, dado que *Elbow Method* se trata de un método de naturaleza principalmente visual, un elemento clave de comparación es el *elbow* (ver 5.6.2). Por último, se presenta un análisis comparativo de la curva de la inercia (ver 5.6.3).

5.6.1. Análisis evolutivo de la inercia (*Elbow Method*) mediante el *scraping* progresivo de los datos

En esta subsección, se muestra la evolución de la inercia utilizando la métrica *Elbow Method* mediante el *scraping* progresivo de los datos, así como el k óptimo a medida que se incrementa la muestra para el caso de los datos de las IDEs de España (ver 5.6.1.1) y Uruguay (ver 5.6.1.2). Finalmente, se presenta un análisis comparativo (ver 5.6.1.3) en términos evolutivos.

5.6.1.1. IDE España

En la Figuras 5.32 y 5.33 se muestra la evolución de *Elbow Method* a medida que se incrementa progresivamente el volumen de datos obtenido mediante *scraping*.

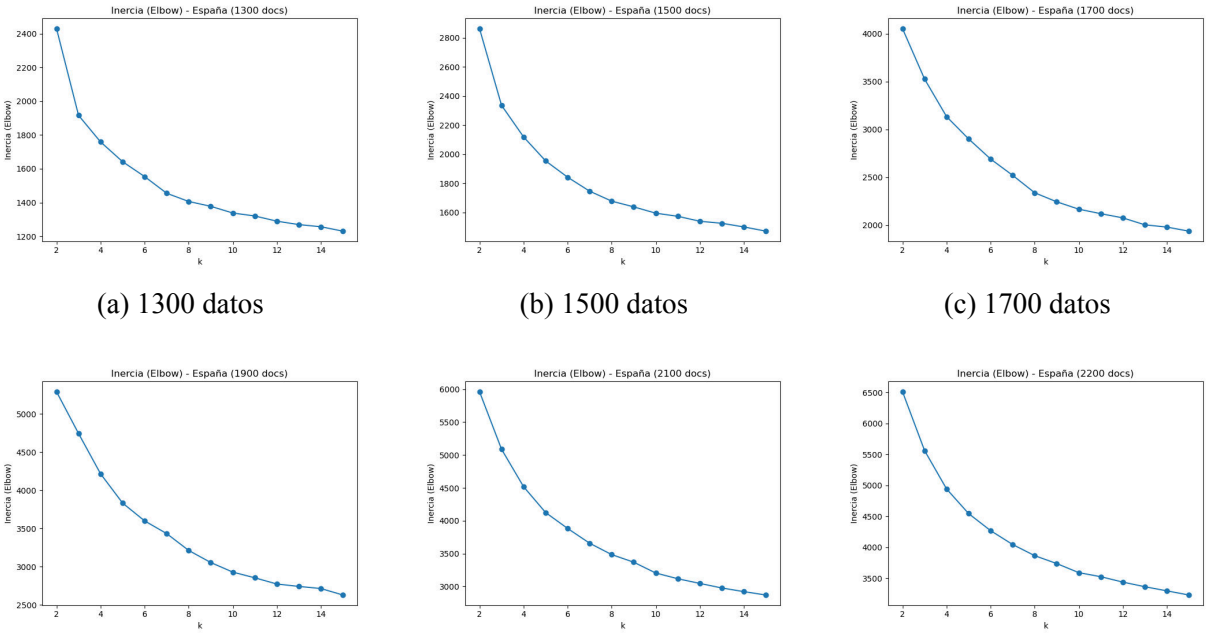


(a) 100 datos (b) 300 datos (c) 500 datos

(d) 700 datos (e) 900 datos (f) 1100 datos

Figura 5.32: Evolución de la inercia (*Elbow Method*) variando el tamaño de la muestra para la IDE de España(100-1100 datos)

Fuente: Elaboración propia.



(a) 1300 datos (b) 1500 datos (c) 1700 datos

(d) 1900 datos (e) 2100 datos (f) 2200 datos

Figura 5.33: Evolución de la inercia (*Elbow Method*) variando el tamaño de la muestra para la IDE de España(1300-2200 datos)

Fuente: Elaboración propia.

En la Tabla 5.18 se muestra la evolución de la inercia para el *Elbow Method* en función de la cantidad de datos de España, mientras que en la Figura 5.18 se presenta su interpretación. Los valores de inercia se expresan en notación científica con dos decimales, de manera que incluso diferencias muy pequeñas, como las observadas en los datos de la IDE de Uruguay (ver 5.6.1.2), puedan capturarse. Para más precisión, véase la Tabla A.9 en los anexos, quedando este aspecto abierto a futuras investigaciones (ver Capítulo 8).

País	Cantidad de documentos	Número de *clusters*	Inercia (WCSS)
España	100	10	$2{,}48 \times 10^{0}$
	300	6 (↓)	$5{,}64 \times 10^{2}$ (↑)
	500	7 (↑)	$1{,}23 \times 10^{3}$ (↑)
	700	7 (=)	$1{,}22 \times 10^{3}$ (↓)
	900	7 (=)	$1{,}23 \times 10^{3}$ (↑)
	1100	6 (↓)	$1{,}22 \times 10^{3}$ (=)
	1300	7 (↑)	$1{,}22 \times 10^{3}$ (=)
	1500	6 (↓)	$1{,}47 \times 10^{3}$ (↑)
	1700	8 (↑)	$1{,}91 \times 10^{3}$ (↑)
	1900	7 (↓)	$2{,}52 \times 10^{3}$ (↑)
	2100	6 (↓)	$2{,}77 \times 10^{3}$ (↑)
	2200	6 (=)	$3{,}16 \times 10^{3}$ (↑)

Tabla 5.18: Valores de inercia y k óptimo para España en función de la cantidad de documentos

Fuente: Elaboración propia.

↑ Incremento respecto al valor anterior ↓ Decremento respecto al valor anterior = Valor igual al anterior

Figura 5.34: Interpretación de símbolos de la Tabla 5.18

Fuente: Elaboración propia.

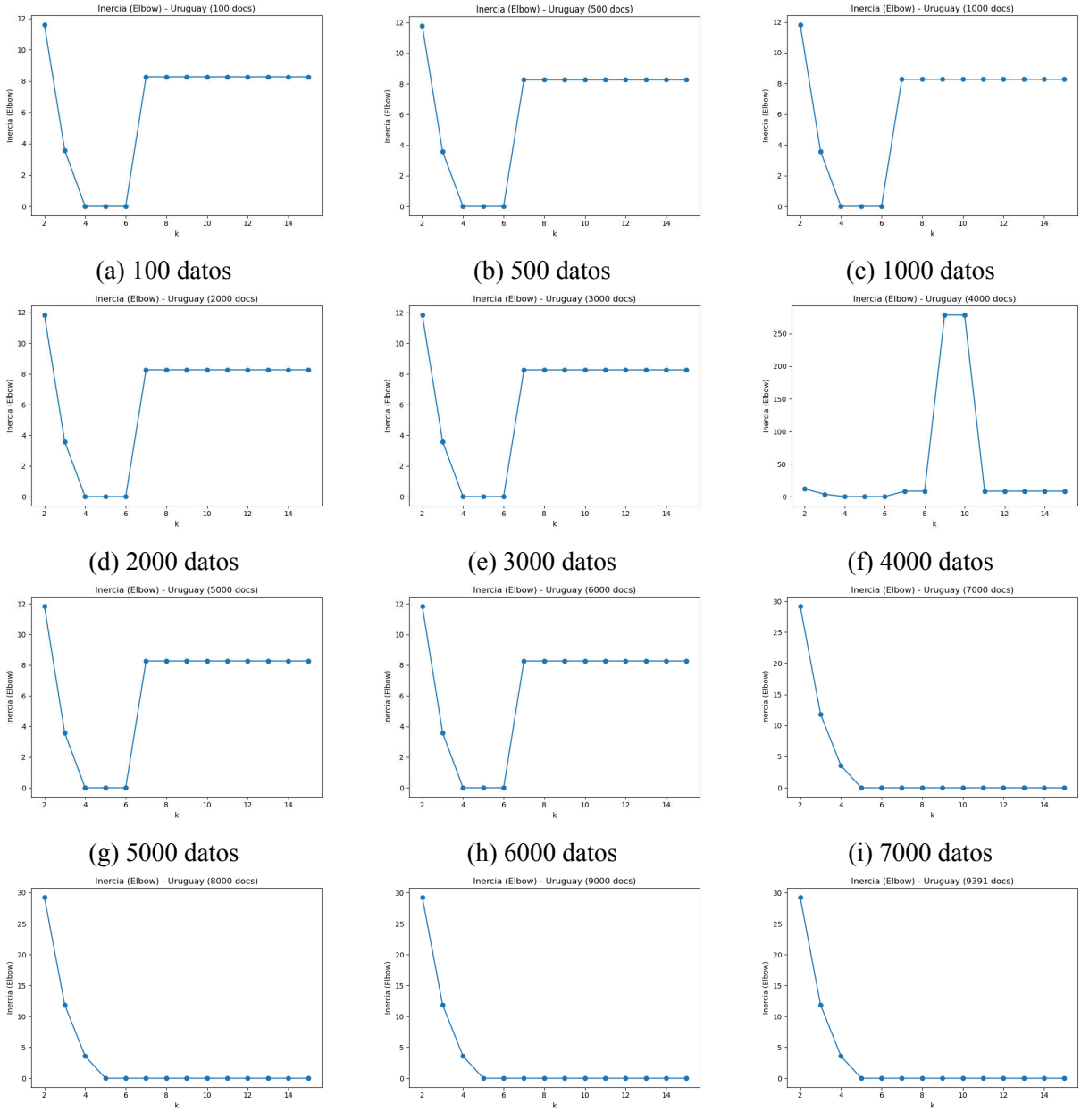
Como se observa en Figura 5.32 (a) y (b), en las primeras etapas del análisis con muestras reducidas no se identifica un *elbow* claro en la curva de inercia. Esto sugiere que, con pocos datos, no se logra capturar una estructura interna definida en los metadatos, lo que dificulta la estimación del número óptimo de *clusters*. Sin embargo, al ampliar progresivamente la muestra mediante *scraping* (ver 5.33f), el *elbow* se vuelve más pronunciado, lo que indica una mejora en la cohesión intra-*cluster* y una mayor claridad en la segmentación temática.

Este comportamiento justifica tanto la necesidad de continuar con el *scraping* como la inclusión de múltiples gráficas en el análisis. Cada gráfico representa un punto clave en la evolución de la estructura de los datos, permitiendo observar cómo se estabilizan las métricas y se define mejor la segmentación a medida que se incorporan más documentos.

En definitiva, se observa que el *scraping* progresivo no solo mejora la calidad del *clustering*, sino que también revela patrones que no son evidentes en muestras pequeñas.

5.6.1.2. IDE Uruguay

En la Figura 5.35 se muestra la evolución de la inercia a medida que se incrementa progresivamente el volumen de datos obtenido mediante *scraping*.



(a) 100 datos (b) 500 datos (c) 1000 datos

(d) 2000 datos (e) 3000 datos (f) 4000 datos

(g) 5000 datos (h) 6000 datos (i) 7000 datos

(j) 8000 datos (k) 9000 datos (l) 9391 datos

Figura 5.35: Evolución de la inercia (*Elbow Method*) para diferentes volúmenes de datos en Uruguay

Fuente: Elaboración propia.

En la Tabla 5.19 se muestra la evolución de la inercia para el *Elbow Method* en función de la cantidad de datos de Uruguay. En los anexos (ver Tabla A.8) se presenta la versión original. Los valores se expresan en notación científica con dos decimales, de manera que incluso diferencias muy pequeñas puedan ser capturadas de forma precisa. Este mismo criterio se aplicó en el caso de España (ver 5.6.1.1) para garantizar la consistencia en la representación de la inercia.

País	**Cantidad de documentos**	**Número de *clusters***	**Inercia**
Uruguay	100	4	$6{,}30 \times 10^{-13}$
	1000	4 (↑)	$3{,}15 \times 10^{-13}$ (↓)
	2000	4 (=)	$4{,}08 \times 10^{-14}$ (↓)
	3000	4 (=)	$3{,}85 \times 10^{-14}$ (↓)
	4000	4 (=)	$2{,}44 \times 10^{-14}$ (↓)
	5000	2 (↓)	$2{,}02 \times 10^{-13}$ (↑)
	6000	4 (↑)	$5{,}50 \times 10^{-14}$ (↓)
	7000	5 (↑)	$9{,}99 \times 10^{-10}$ (↑)
	8000	5 (=)	$5{,}21 \times 10^{-09}$ (↑)
	9000	5 (=)	$5{,}58 \times 10^{-09}$ (↑)
	9391	5 (=)	$1{,}44 \times 10^{-08}$ (↑)

Tabla 5.19: Valores de inercia y k óptimo para España en función de la cantidad de documentos

Fuente: Elaboración propia.

↑ Incremento respecto al valor anterior ↓ Decremento respecto al valor anterior = Valor igual al anterior

Figura 5.36: Interpretación de símbolos de la Tabla 5.19

Fuente: Elaboración propia.

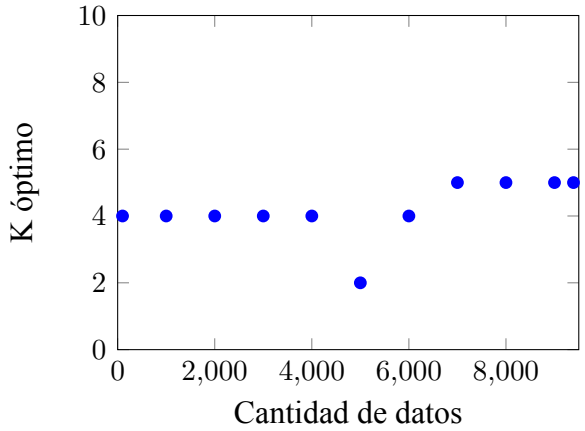
Los resultados obtenidos mediante *Elbow Method* (ver Tabla 5.19) presentan un comportamiento consistente con las métricas de Davies–Bouldin (ver 5.4) y Calinski–Harabasz (ver 5.5). En la mayoría de los casos, hasta 3000 documentos, el número óptimo de *clusters* se mantiene en cuatro. A partir de los 7000 documentos, el análisis identifica un quinto *cluster*, que se conserva hasta la totalidad de los registros. La única excepción se observa con 5000 documentos, donde el método sugiere dos *clusters*, probablemente debido a una ambigüedad puntual en la localización del *elbow* en el gráfico de inercia. Este hallazgo no se repite en los conjuntos con mayor volumen, por lo que puede interpretarse como una variación aislada y no

como una tendencia estructural del conjunto de datos.

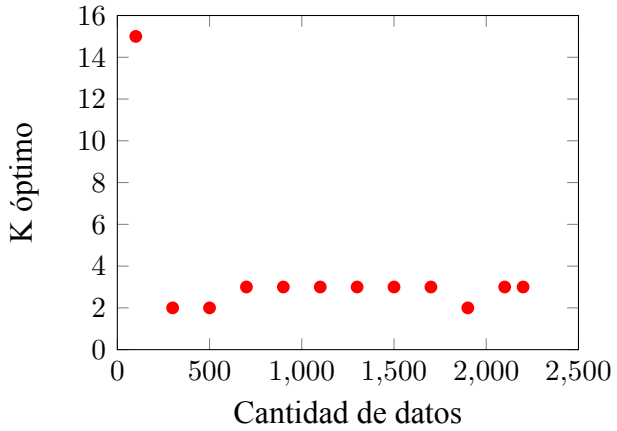
5.6.1.3. Análisis comparativo de la evolución

Se compara la evolución de ambos conjuntos de datos tomando como premisa que una mayor variabilidad a medida que se *scrapean* más datos indica mayor diversidad en la información.

En el caso de los datos de la IDE de Uruguay, tanto la evolución de la inercia (ver Figura 5.38a) como del k óptimo (ver Figura 5.37a) se mantiene relativamente constante a medida que aumenta la cantidad de documentos. Esto sugiere que los *clusters* son estables y bien definidos, mostrando un *elbow* claro que facilita la elección del número óptimo de *clusters*. Por el contrario, en los datos de la IDE de España, tanto la inercia (ver Figura 5.38b) como el k óptimo (ver Figura 5.37b) presentan fluctuaciones significativas con el incremento de documentos. Esto refleja una mayor heterogeneidad de la información y hace que el *elbow* sea menos evidente, dificultando la determinación del número óptimo de *clusters*.



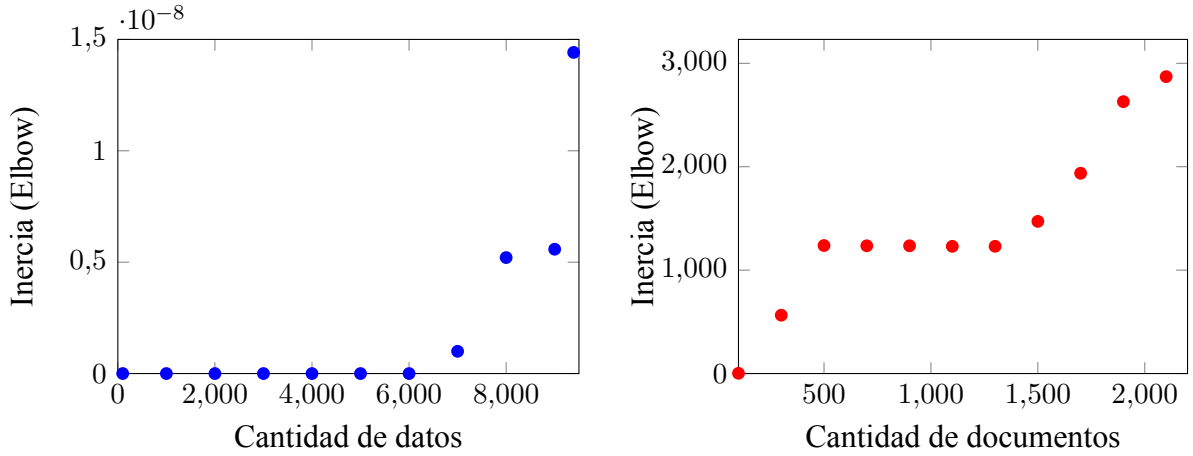
(a) Evolución del k óptimo en Uruguay.



(b) Evolución del k óptimo en España.

Figura 5.37: Comparación de la evolución del k óptimo en Uruguay y España en función de la cantidad de datos

Fuente: Elaboración propia.



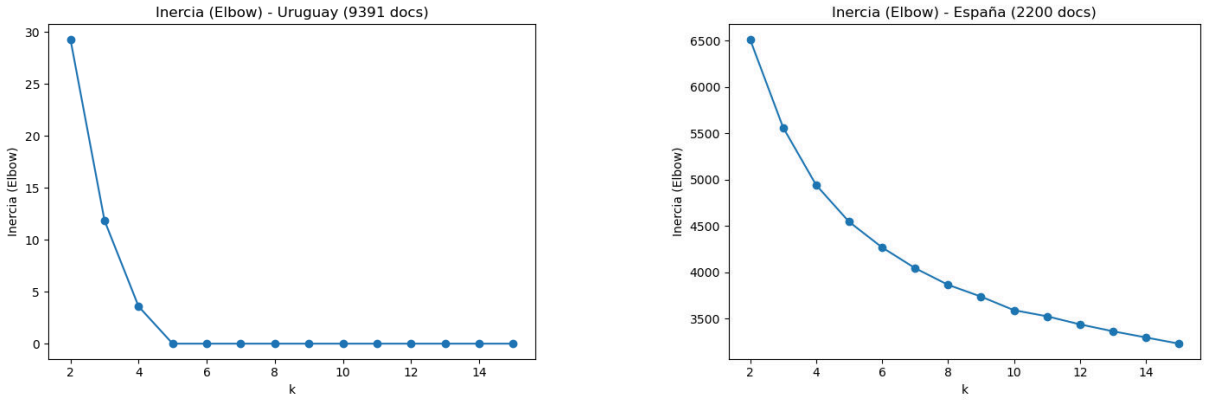
(a) Evolución de la Inercia (*Elbow*) en Uruguay (b) Evolución de la Inercia (*Elbow*) en España

Figura 5.38: Comparación de la evolución de la inercia entre Uruguay y España

Fuente: Elaboración propia.

5.6.2. Análisis comparativo del *Elbow*

En la Figura 5.39a se aprecia que el *elbow* en la gráfica correspondiente a los datos de la IDE de Uruguay es claramente definido. Sin embargo, tal como se ilustra en la Figura 5.39b, los datos de la IDE de España bajo el cálculo del *Elbow Method*, presentan un *elbow* más ambiguo. Esto sugiere que la estructura de los datos en España es más compleja y presenta mayor variabilidad que la de Uruguay.



(a) Uruguay - 9391 documentos (b) España - 2200 documentos

Figura 5.39: Análisis comparativo del *Elbow* para Uruguay y España

Fuente: Elaboración propia.

En definitiva, esta comparación arroja resultados interesantes en términos comparativos y debe ser estudiada a futuro (ver 8).

5.6.3. Análisis comparativo de la curva de la inercia

El análisis comparativo de la curva de la inercia complementa el enfoque visual presentado en la subsección 5.6.2. El *Elbow Method* busca determinar de manera global el número óptimo de *clusters*, identificando el punto a partir del cual agregar más *clusters* produce reducciones mínimas en la inercia del agrupamiento. En esta subsección, nuestro objetivo es comparar la evolución de la inercia y responder a preguntas clave: ¿es claro el *elbow*? ¿cuándo deja de tener variaciones significativas? Para evaluar numéricamente la claridad del *elbow*, se emplean técnicas que analizan la estabilidad de la función de inercia: la curva deja de mejorar cuando su evolución se estabiliza, es decir, cuando sus derivadas tienden a cero. Por esta razón, se procede a comparar las derivadas discretas de la inercia en función del número de *clusters*.

Los valores de la inercia para los distintos números de *clusters* se presentan en los anexos: para la IDE de España en la Tabla A.10 y para la IDE de Uruguay en la Tabla A.8. Con estos datos, es posible analizar la evolución de la inercia en función del número de *clusters* y proceder a la evaluación numérica de la claridad o ambigüedad del *elbow*.

Se define la función de la inercia de la siguiente manera:

$I : \{2, \ldots, 15\} \to \mathbb{R}$:

En la Figura 5.40 se define para el caso de los datos de la IDE de España y Uruguay respectivamente.

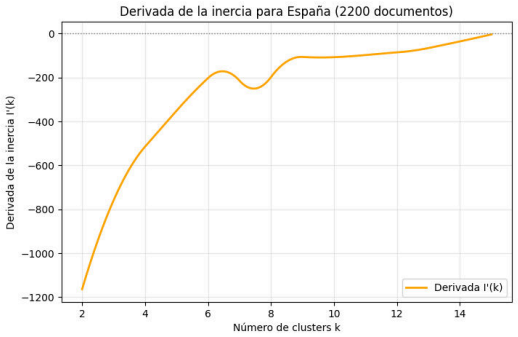
$$I_{\text{España}}(k) = \begin{cases} 6{,}4387 \times 10^3, & k = 2 \\ 5{,}4906 \times 10^3, & k = 3 \\ 4{,}8671 \times 10^3, & k = 4 \\ 4{,}4345 \times 10^3, & k = 5 \\ 4{,}1604 \times 10^3, & k = 6 \\ 3{,}9767 \times 10^3, & k = 7 \\ 3{,}7415 \times 10^3, & k = 8 \\ 3{,}6064 \times 10^3, & k = 9 \\ 3{,}4987 \times 10^3, & k = 10 \\ 3{,}3958 \times 10^3, & k = 11 \\ 3{,}3049 \times 10^3, & k = 12 \\ 3{,}2280 \times 10^3, & k = 13 \\ 3{,}1775 \times 10^3, & k = 14 \\ 3{,}1581 \times 10^3, & k = 15 \end{cases} \qquad I_{\text{Uruguay}}(k) = \begin{cases} 29{,}2593, & k = 2 \\ 11{,}8414, & k = 3 \\ 3{,}5777, & k = 4 \\ 1{,}4444 \times 10^{-8}, & k = 5 \\ 1{,}4461 \times 10^{-8}, & k = 6 \\ 1{,}4434 \times 10^{-8}, & k = 7 \\ 1{,}4464 \times 10^{-8}, & k = 8 \\ 1{,}4473 \times 10^{-8}, & k = 9 \\ 1{,}4399 \times 10^{-8}, & k = 10 \\ 1{,}4460 \times 10^{-8}, & k = 11 \\ 1{,}4480 \times 10^{-8}, & k = 12 \\ 1{,}4488 \times 10^{-8}, & k = 13 \\ 1{,}4535 \times 10^{-8}, & k = 14 \\ 1{,}4529 \times 10^{-8}, & k = 15 \end{cases}$$

Figura 5.40: Inercia para España y Uruguay en función del número de *clusters* (k)

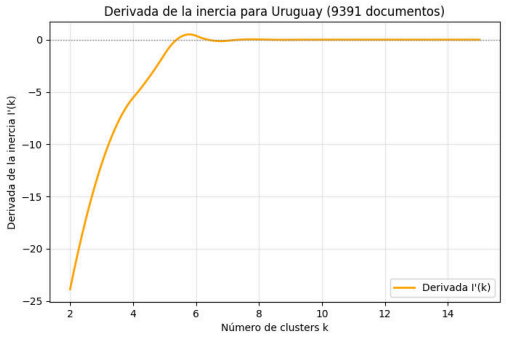
Fuente: Elaboración propia.

Para estimar una derivada continua a partir de valores discretos de inercia, se emplea una interpolación spline cúbica [161], la cual genera una función continua y diferenciable que pasa por todos los puntos de datos originales (ver A.1.5). Esto permite visualizar la tendencia de I(k) y determinar de manera más precisa el punto donde la derivada tiende a cero, es decir, el *elbow* del método. Esto no altera los valores originales, solo proporciona una representación que facilita el análisis y el límite de la derivada.

En la Figura 5.41 se ilustran las derivadas de la inercia. Para la IDE de Uruguay, las derivadas tienden rápidamente a cero y se mantienen estables, lo que evidencia un *elbow* claramente definido. En cambio, para la IDE de España, las derivadas presentan fluctuaciones y no alcanzan una estabilidad similar, indicando un *elbow* más ambiguo y haciendo que la determinación del número óptimo de *clusters* sea menos evidente.



(a) Derivada de la inercia para España



(b) Derivada de la inercia para Uruguay

Figura 5.41: $I'(k)$ de España y Uruguay

Fuente: Elaboración propia.

5.7. Análisis de *clustering*

En términos generales, las métricas de validación interna proporcionan una base sólida para seleccionar los parámetros óptimos con los que llevar a cabo el proceso de *clustering*. Una vez obtenidos estos resultados, se procede a realizar el *clustering* según resultados arrojados en secciones anteriores (ver 5.5, 5.6, 5.3 y 5.4).

En esta etapa buscamos responder preguntas clave de la investigación, tales como: ¿qué tan similares son los grupos entre sí?, ¿es realmente posible agrupar los documentos de forma coherente?, ¿qué tan compleja es la estructura subyacente?, ¿hasta qué punto los documentos comparten características en común? Estas preguntas guían este apartado y pretenden encontrar una respuesta desarrollando un análisis en un contexto meramente exploratorio.

En primer lugar, el análisis exploratorio comienza con la visualización de los *clusters* resultantes en un espacio reducido a través de la técnica de PCA. Esta técnica nos permite representar los datos en un espacio bidimensional simplificado, lo que facilita la visualización de la distribución y proximidad relativa de los documentos. La representación visual complementa el análisis cuantitativo, ya que permite observar la composición de cada *cluster* y evaluar de forma intuitiva si los grupos detectados reflejan patrones reales en los datos o si se solapan entre sí. De esta manera, no solo ponemos en práctica las métricas calculadas

previamente, sino que también verificamos su coherencia en un contexto exploratorio.

Por otro lado, se presenta la distribución de documentos por *cluster* con el fin de lograr una vasta comprensión de la estructura de *clustering* resultante.

Asimismo, para continuar con la línea de análisis exploratorio se realizan nubes de palabras para cada *cluster* con el objetivo de analizar los subtemas subyacantes y determinar la complejidad de la estructura en análisis.

Por último, se presenta una breve descripción de cada *cluster*, así como la asignación de una etiqueta posible. En esta línea, cabe aclarar que la calidad de *clustering* en algunos casos no es buena. En algunos casos, los *clusters* resultantes se solapan o la cohesión no es elevada y por tanto el etiquetado se vuelve más interpretativo y no tan determinante por la estructura en cuestión.

Finalmente, se detallan algunas observaciones proporcionadas tras el análisis exploratorio.

El análisis exploratorio relatado se presenta para la IDE de España (ver 5.7.1) y para la IDE de Uruguay (ver 5.7.2). Luego, en la subsección 5.7.3 se expone una comparación entre ambos, con el objetivo de obtener una visión integral de las diferencias y similitudes en sus respectivas estructuras de datos.

5.7.1. IDE España

En la presenta subsección se presenta el análisis de *clustering* realizado tras el cálculo del k óptimo según la métrica *Silhouette Score* (ver 5.7.1.1), Davies-Bouldin (ver 5.7.1.2), Calinski-Harabasz (ver 5.7.1.3) y según *Elbow Method* (ver 5.7.1.4).

5.7.1.1. *Clustering* según la métrica *Silhouette Score*

En las siguientes líneas, se detalla el análisis de *clustering* realizado tras el cálculo del k óptimo con la métrica *Silhouette Score*. En este sentido, el k óptimo es 15, por tanto, el análisis de *clustering* se realiza conforme a este resultado.

En la Figura 5.42 se observa el *clustering* para k=15, obtenido tras calcular el k óptimo de la métrica *Silhouette Score* y tras reducir la dimensionalidad de los datos mediante *Principal Component Analysis (PCA)*.

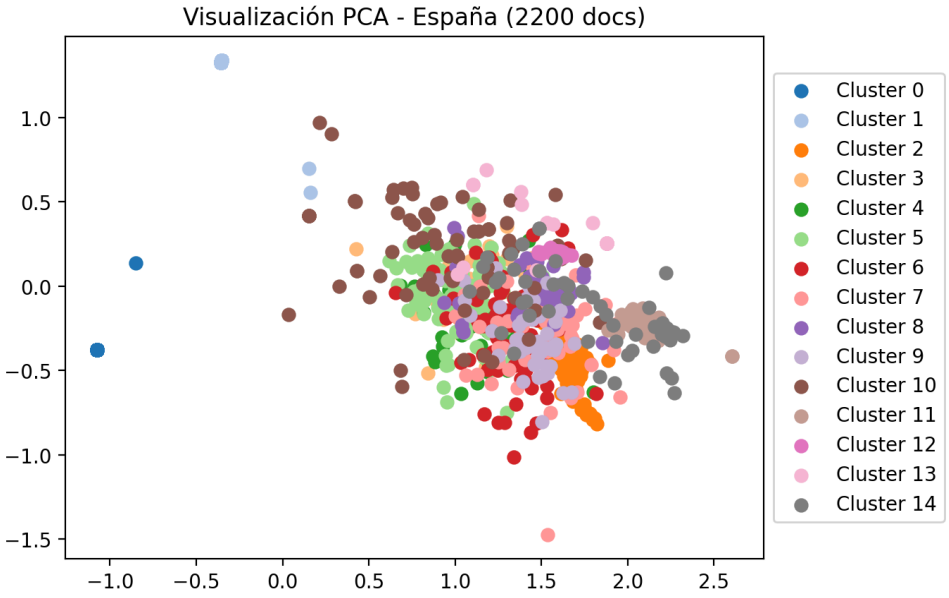


Figura 5.42: Clustering de la IDE de España evaluado con Silhouette Score (k=15) luego de aplicar PCA para visualización bidimensional

Fuente: Elaboración propia.

La Figura 5.42 muestra el siguiente comportamiento:

- **Agrupación temática y separación:** Aunque el *clustering* se realizó con un número elevado de *clusters* ($k = 15$), se observa cierta superposición temática, lo que indica que las fronteras entre *clusters* no son completamente nítidas.
- **Cohesión interna y estructura:** Tal como se analizó previamente mediante el *Silhouette Score* (ver 5.3), la estructura general de los *clusters* es razonable. La cohesión interna no es débil, aunque algunos puntos se encuentran dispersos, sin comprometer significativamente la agrupación.

En la Figura 5.43 se puede observar la distribución de documentos por *cluster* para la IDE de España.

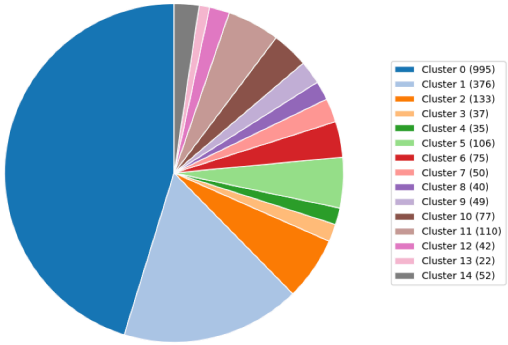
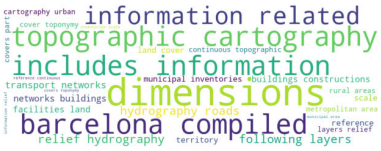


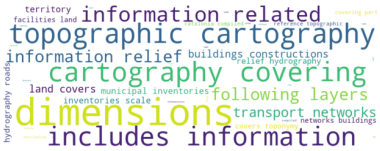
Figura 5.43: Distribución de documentos por *cluster* tras cálculo del *Silhouette Score*

Fuente: Elaboración propia.

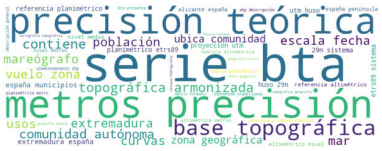
En las Figuras 5.44 y 5.45 se muestran las nubes de palabras de cada *cluster*.



Cluster 0: Gestión y planificación territorial en Barcelona



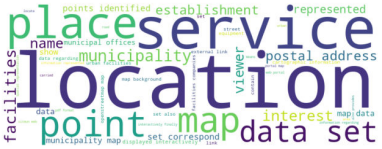
Cluster 1: Gestión y planificación territorial en Cataluña



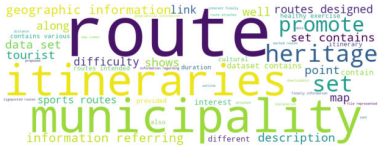
Cluster 2: Zonas costeras y mareografía

Cluster 3: Geografía general y mapas

Cluster 4: Turismo Castelló d'Empúries

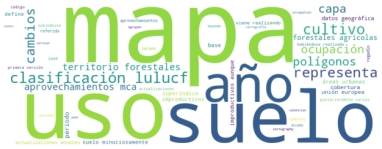


Cluster 5: Información local y oficinas municipales



Cluster 6: Turismo, deporte y cultura

Cluster 7: Incendios y conservación ambiental

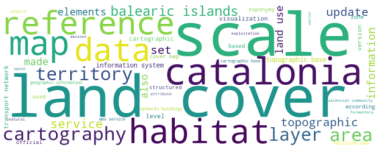


Cluster 8: Uso del suelo y clasificación territorial

Figura 5.44: Nube de palabras para *clusters* (k=15, resultado de calcular el k óptimo con la métrica *Silhouette Score*) - IDE España

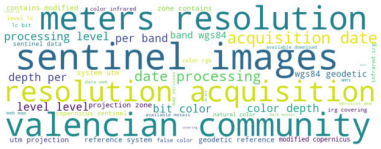
Fuente: Elaboración propia.

Cluster 9: Planificación urbana

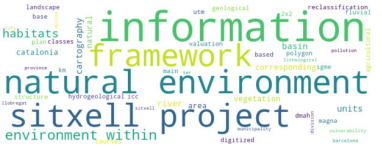


Cluster 10: Cobertura del suelo – Cataluña, Baleares, Valencia

Cluster 11: Monitoreo ambiental y tecnologías



Cluster 12: Estudios de la Tierra y datos geográficos



Cluster 13: Ríos, contaminación y vulnerabilidad ambiental

Cluster 14: Ayudas agropecuarias – Navarra

Figura 5.45: Nube de palabras para *clusters* (k=15, resultado de calcular el k óptimo con la métrica *Silhouette Score*) - IDE España

Fuente: Elaboración propia.

A continuación, se asigna una etiqueta y una breve descripción a lo observado en la Figura 5.45:

- ***Cluster* 0 - Gestión y planificación territorial en Barcelona:** Como se observa en la Figura 5.43, este *cluster* tiene la mayor cantidad de documentos. Se destacan datos de Barcelona y temas abarcativos de gestión y planificación del territorio, como *networks*, *buildings*, *constructions*, coincidiendo con el análisis descriptivo (ver 5.2.3).
- ***Cluster* 1 - Gestión y planificación territorial en Cataluña:** Segundo *cluster* con más documentos, similar al *cluster* 0 pero centrado en Cataluña en términos generales. Se observa que comparte gran parte de la información con el *cluster* 0 relativo a Barcelona, indicando posible subagrupaciones demasiado finas.
- ***Cluster* 2 - Zonas costeras y mareografía:** Agrupa 133 documentos con términos asociados al mar, mareógrafos y regiones costeras como Extremadura y Alicante.
- ***Cluster* 3 - Geografía general y mapas:** Incluye conceptos genéricos sobre lugares, ubicaciones, calles y datos geográficos de interés.

- ***Cluster* 4 - Turismo en Castelló d'Empúries:** Información relacionada con turismo y lugares visitados, útil para la toma de decisiones en el sector turístico.
- ***Cluster* 5 - Información local y oficinas municipales** Agrupa referencias a distintos lugares mediante mapas y recursos disponibles en la web.
- ***Cluster* 6 - Turismo, deporte y cultura:** Similar al *cluster* 4, abarcando deporte, cultura y salud, posiblemente una subagrupación fina.
- ***Cluster* 7 - Incendios y conservación ambiental:** Documentos sobre prevención de incendios y conservación del espacio, incluyendo términos como avia.
- ***Cluster* 8 - Uso del suelo y clasificación de terrenos:** Centrado en la utilización del suelo, clasificación de suelos y organización de información cartográfica.
- ***Cluster* 9 - Planificación urbana y POUM:** Incluye términos de planificación urbana y referencias al Plan de Ordenación Urbanística Municipal(POUM).
- ***Cluster* 10 - Cobertura del suelo en Cataluña, Baleares y Valencia:** Información específica sobre cobertura del suelo en estas regiones.
- ***Cluster* 11 - Tecnologías de monitoreo ambiental (AVHRR):** Datos sobre tecnologías de monitoreo ambiental, vegetación y sequía.
- ***Cluster* 12 - Estudios geográficos y ambientales:** Agrupa estudios de la Tierra para recopilar datos geográficos y ambientales.
- ***Cluster* 13 - Ríos, contaminación y vulnerabilidad ambiental:** Metadatos relacionados con ríos, contaminación y aspectos ambientales vulnerables.
- ***Cluster* 14 - Ayudas agropecuarias en Navarra**: Centrado en ayudas agropecuarias proporcionadas por la localidad de Navarra.

Las nubes de palabras mostradas en la Figura 5.45 permiten complementar el análisis presentado en la Figura 5.42. En ellas, se observa claramente la repetición de ciertos términos,

lo que provoca solapamiento entre algunos conceptos. Por ejemplo, en las subfiguras (a) y (b) se repiten términos como *topographic cartography* y *dimensions*.

Por otro lado, se observa la complejidad de los metadatos de la categoría *Land Cover* de la IDE de España, donde coexisten diversas temáticas y subtemáticas muy específicas, como turismo, estudios geográficos, ríos, entre otras, representadas en los distintos *clusters*. Este patrón refleja características particulares del país y de los datos, que son altamente detallados y especializados. Este comportamiento es consistente con el uso del *Silhouette Score*, ya que esta métrica promueve la cohesión interna dentro de cada *cluster*.

En definitiva, el análisis basado en *Silhouette Score* no solo permite identificar los subtemas que emergen y contribuyen a la variabilidad temática, sino que también enriquece la comprensión de la métrica en sí, mostrando cómo la cohesión interna facilita el estudio de los distintos subgrupos emergentes.

5.7.1.2. *Clustering* según la métrica *Davies-Bouldin*

En líneas posteriores, se realiza el análisis *clustering* según el k óptimo (k=3) obtenido tras realizar el cálculo de la métrica Davies-Bouldin.

En la Figura 5.46 se observa el *clustering* para$k$=3, obtenido tras calcular el k óptimo de la métrica Davies-Bouldin y tras reducir la dimensionalidad de los datos mediante *Principal Component Analysis (PCA)*.

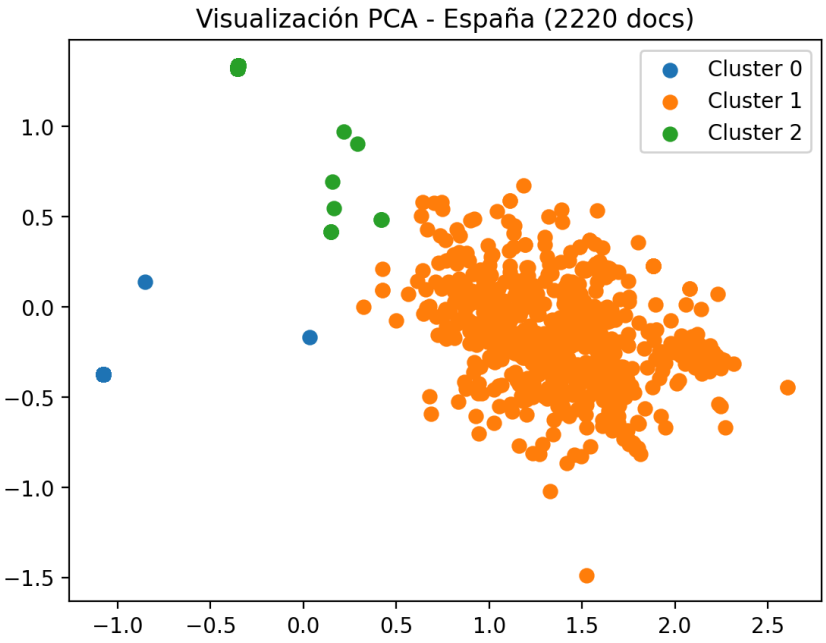


Figura 5.46: *Clustering* de la IDE de España evaluado con Davies-Bouldin Index (k=3)

Fuente: Elaboración propia.

La Figura 5.46 permite realizar un estudio de los siguientes puntos:

- **Cohesión interna:** El *cluster* 1 evidencia poca cohesión interna, los puntos se encuentran sumamente dispersos. Por otro lado, el *cluster* 0 y 2 también presentan cierta dispersión pero no tan pronunciada como el *cluster* 1 que evidencia una cohesión interna muy baja.
- **Pocas temáticas claras y compactas:** Este punto evidencia el comportamiento de la métrica Davies-Bouldin, prefiriendo una estructura global simple y no capturando subtemáticas dentro del grupo mayoritario. Se observa que el *clustering* basado en Davies-Bouldin detecta unas pocas categorías temáticas claras y compactas, sacrificando granularidad y dejando la mayor parte de los documentos en un único gran grupo, lo cual es coherente con la limitación conocida de este índice (tiende a subestimar el número óptimo de *clusters*).

En la Figura 5.47 se puede observar la distribución de documentos por *cluster* para la IDE de España.

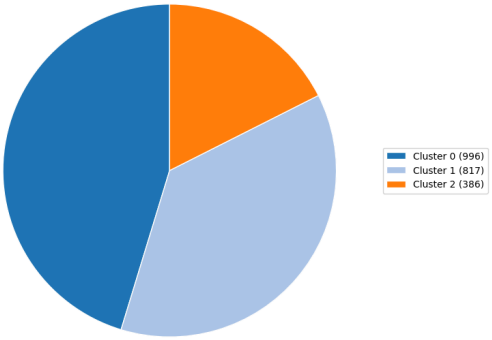
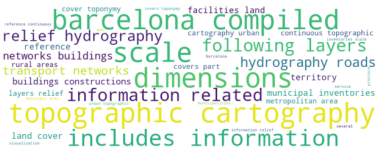


Figura 5.47: Distribución de documentos por *cluster* tras cálculo de Davies-Bouldin

Fuente: Elaboración propia.

En la Figura 5.48 se muestran las nubes de palabras de manera granular por *cluster*.



Cluster 0: Vista general (*Land Cover* e hidrografía) - IDE España

Cluster 1: Aspectos ambientales - IDE España

Cluster 2: Vista general (*Land Cover* e infraestructuras) - IDE España

Figura 5.48: *Clusters* obtenidos según Davies-Bouldin Index - IDE España

Fuente: Elaboración propia.

En líneas posteriores, se detalla el análisis por *cluster* trás realizar el análisis exploratorio ilustrado en la Figura 5.48:

- ***Cluster* 0 - Vista general (*Land Cover* e hidrografía):** Presenta una vista general de la información, incluyendo la categoría *Land Cover* y repeticiones del término *hydrography*. No se observa un enfoque específico en los documentos agrupados.

- ***Cluster* 1 - Aspectos ambientales:** Agrupa información relacionada con vegetación y sequía, destacando aspectos ambientales del área analizada.
- ***Cluster* 2 - Vista general (*Land Cover* e infraestructuras):** *Cluster* genérico que incluye documentos relacionados con la categoría (*Land Cover*) presentados de manera poco granular. Incluye términos tales como *Land Cover*, redes, construcción de edificios y otras infraestructuras, abarcando distintos tipos de información.

En conclusión, el análisis de *clustering* basado en la métrica Davies-Bouldin no produce una separación de *clusters* lo suficientemente significativa como para permitir un estudio detallado de los subtemas emergentes o de la diversidad temática presente en los datos. Esto se ve reflejado en líneas anteriores, en donde, las agrupaciones resultantes resultan poco especificas y no aportan información del detalle de metadatos. No obstante, esta aproximación resulta útil para obtener una visión panorámica del conjunto de datos y para comprender el comportamiento de la métrica Davies-Bouldin en la delimitación de los *clusters*.

5.7.1.3. Clustering según métrica Calinski-Harabasz

En párrafos siguientes se desarrolla el análisis de *clustering* según el k óptimo (k=2) obtenido tras realizar el cálculo de la métrica Calinski-Harabasz.

En la Figura 5.49 se observa el *clustering* para k=2, obtenido tras calcular el k óptimo de la métrica Calinski-Harabasz y tras reducir la dimensionalidad de los datos mediante *Principal Component Analysis* (*PCA*).

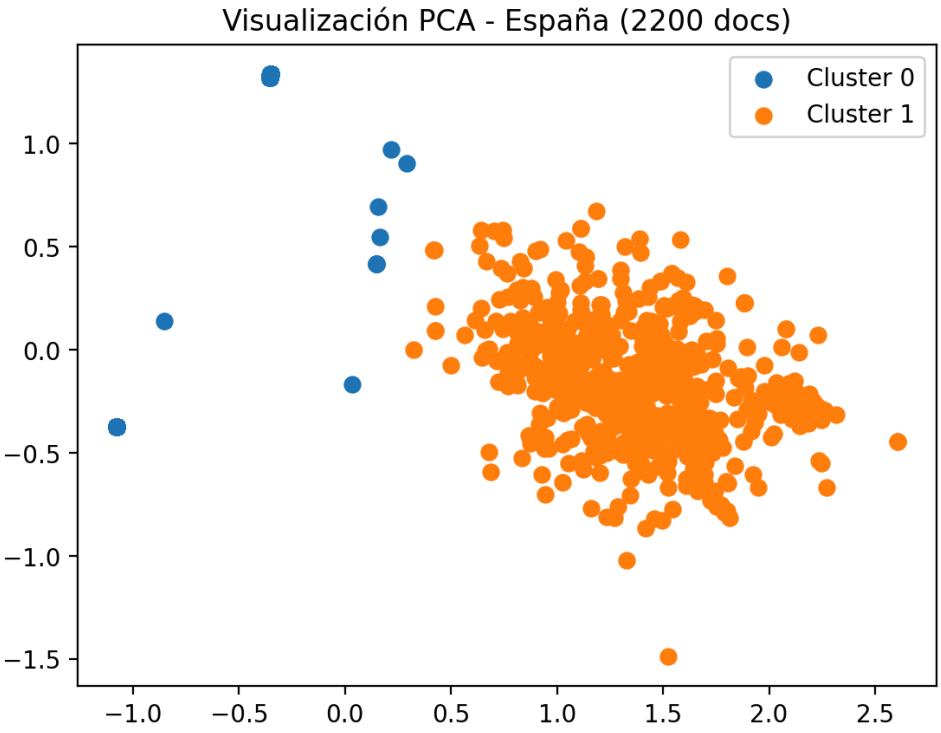


Figura 5.49: *Clustering* de la IDE de España evaluado con Calinsky-Harabasz(k=2)

Fuente: Elaboración propia.

La Figura 5.49 permite la observación de los siguientes puntos:

- **Maximización de la separación de *clusters*:** Se puede observar que ambos *clusters* se encuentran separados entre si. Esto se debe a que la métrica Calinski-Harabasz prioriza la maximización de la separación entre *clusters*.
- **Cohesión interna:** Los puntos se encuentran dispersos dentro del espacio reducido de coordenadas, lo que indica que los documentos no presentan una cohesión interna razonable.

En la Figura 5.50 se puede observar la distribución de documentos por *cluster* para la IDE de España.

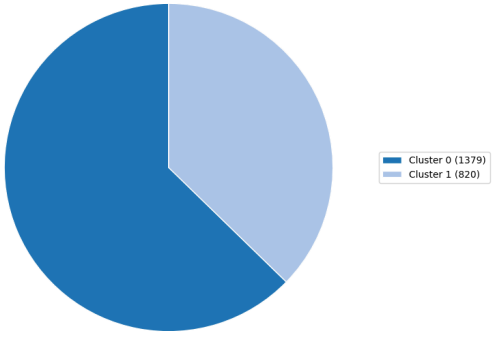
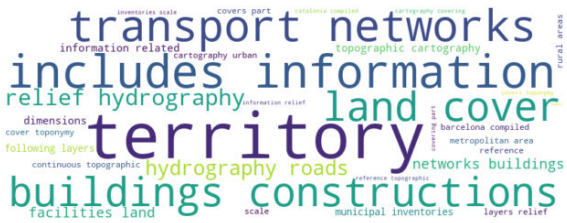


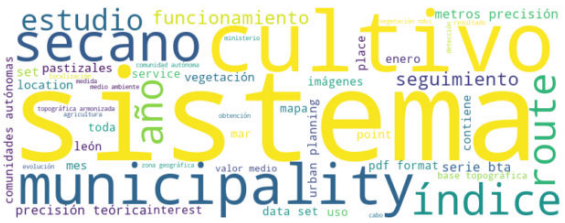
Figura 5.50: Distribución de documentos por *cluster* tras cálculo de Calinski-Harabasz

Fuente: Elaboración propia.

En la Figura 5.49 se observa la nube de palabras tras realizar el *clustering* con la métrica Calinski-Harabasz.



Cluster 0: Gestión y planificación urbana-IDE España



Cluster 1: Uso agrícola del territorio

Figura 5.51: Nube de palabras para *clusters* obtenidos según Calinski-Harabasz Index-IDE España

Fuente: Elaboración propia.

A continuación se describen algunos de los *clusters* identificados en el análisis utilizando la métrica de Calinski-Harabasz en la IDE de España, considerando los términos más representativos de cada *cluster*:

- ***Cluster* 0 – Gestión y planificación urbana:** Se observan términos como *transport networks*, *includes*, *information*, *land cover* (nombre de la categoría), *hydrography*,

building y *constructions*, lo que indica que este *cluster* agrupa información relacionada con infraestructura y planificación urbana.

- **Cluster 1 – Uso agrícola del territorio:** Los términos más predominantes son *secano*, *cultivo* y *sistema*, seguidos de *vegetación*, *enero*, *agricultura*, *pastizal* e *índice*, sugiriendo que este *cluster* se centra en el uso agrícola del territorio y la vegetación.

El *clustering* con Calinski-Harabasz ofrece un panorama general, mostrando cómo se maximiza el separamiento entre *clusters*. Resulta interesante visualizar esto a través de las nubes de palabras, complementando el análisis numérico previo.(5.5).

5.7.1.4. *Clustering* según *Elbow Method*

En la presente subsubsección, se llevó a cabo el *clustering* según el k óptimo(k=6) obtenido tras realizar *Elbow Method*.

En la Figura 5.53 se observa el *clustering* para k=6, obtenido tras calcular el k óptimo de la métrica *Elbow Method* y tras reducir la dimensionalidad de los datos mediante *Principal Component Analysis* (PCA).

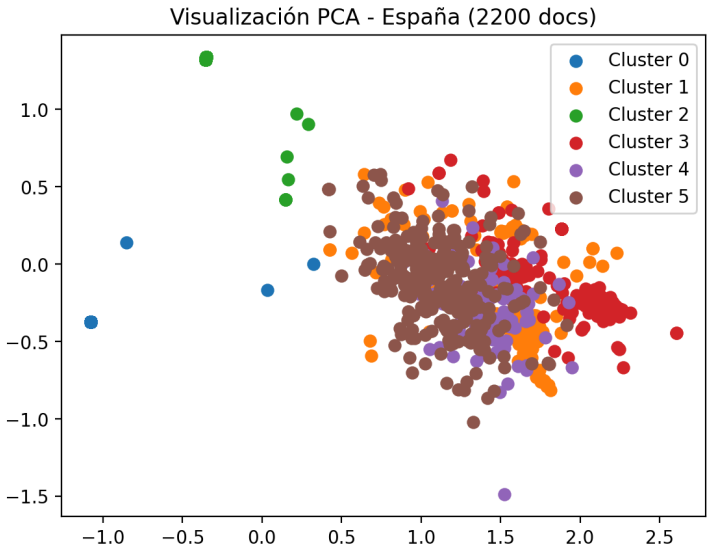


Figura 5.52: *Clustering* de la IDE de España evaluado con *Elbow Method*(k=6)

Fuente: Elaboración propia.

La figura 5.53 permite realizar las siguientes observaciones:

- ***Clusters* solapados:** Se observa que los *clusters* se encuentran parcialmente solapados, lo que indica que un número tan reducido de *clusters* no es suficiente para capturar temáticas independientes.
- **Entendimiento de la métrica:** La visualización permite comprender cómo actúa el *Elbow Method* en la selección de un número de *clusters*, mostrando el equilibrio entre complejidad y rendimiento, es decir, muestra menos complejidad que la estructura resultante trás el análisis de *clustering* para *Silhouette Score* (ver 5.7.1.1), ni tan poco compleja como la estructura resultante tras el análisis de *clustering* para Calinski-Harabasz(ver 5.7.1.3) y Davies-Bouldin (ver 5.7.1.2).

En la Figura 5.50 se puede observar la distribución de documentos por *cluster* para la IDE de España.

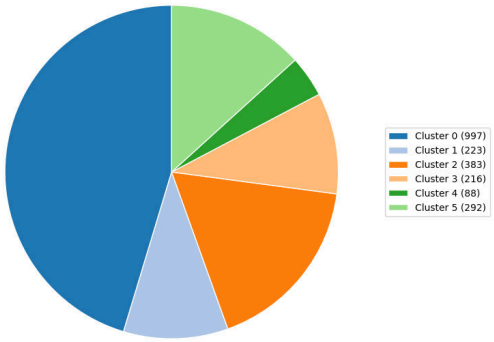
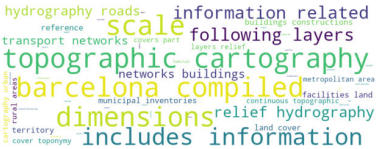


Figura 5.53: Distribución de documentos por *cluster* tras cálculo de *Elbow Method*

Fuente: Elaboración propia.

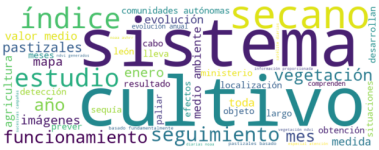
En la Figura 5.49 se ilustra la nube de palabras tras realizar el *clustering* con la métrica *Elbow Method*.



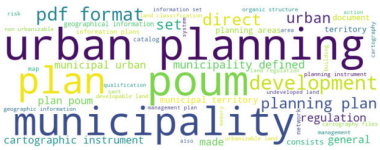
Cluster 0: Gestión y análisis de información-IDE España

Cluster 1: Precisión y referencias geográficas-IDE España

Cluster 2: Transporte e infraestructura-IDE España



Cluster 3: Vegetación y uso del suelo agrícola



Cluster 4: Planeamiento urbano y digitalización

Cluster 5: Localización y servicios municipales

Figura 5.54: *Clusters* obtenidos según el *Elbow Method*

Fuente: Elaboración propia.

A continuación se describen los *clusters* identificados en el análisis utilizando la métrica *Elbow Method* en las IDE de España, considerando los términos más representativos de cada *cluster*:

- ***Cluster* 0 – Gestión y análisis de información:** Se centra en la información y su uso. Los términos frecuentes, como *municipal inventory*, destacan no solo la relevancia de la información en sí, sino también cómo se estudia y aplica. Este *cluster* refleja la gestión y análisis de datos en distintos contextos: municipios, zonas rurales, transporte, hidrografía, y especialmente en la zona metropolitana. En resumen, abarca tanto la información disponible como los estudios y aplicaciones que se realizan con ella.

- ***Cluster* 1 – Precisión y referencias geográficas:** La palabra precisión aparece con frecuencia. Este *cluster* destaca la importancia de los sistemas de medida y referencias geográficas, incluyendo metros de precisión, mareógrafos, referencias altimétricas y planimétricas. Lo relevante aquí es la exactitud de la información y su fiabilidad para análisis y planificación.

- ***Cluster* 2 – Transporte e infraestructura:** Contiene términos relacionados con transporte, redes, construcción de edificios, hidrografía, carreteras y territorio. Se

observa la presencia de términos como *covering*, indicando el análisis de la cobertura territorial y la infraestructura en distintas áreas.

- ***Cluster* 3 – Vegetación y uso del suelo agrícola:** Este *cluster* se enfoca en la vegetación, cultivos, pastizales y secano, reflejando aspectos ambientales y agrícolas de los territorios analizados.

- ***Cluster* 4 – Planeamiento urbano y digitalización:** Contiene términos relacionados con planificación urbana (*planning*, *plan*) y menciona formatos digitales (*pdf format*). Aunque abarca múltiples temáticas, se destaca la gestión de información para planeamiento y digitalización de documentos.

- ***Cluster* 5 – Localización y servicios municipales:** Abarca términos como lugares, mapas, *location*, servicios (*service*, *facilities*) y municipalidades (*municipality*), reflejando la información geoespacial vinculada a localización y servicios locales.

El *clustering* con *Elbow Method* ofrece un panorama ni demasiado genérico ni demasiado granular. La estructura resultante nos otorga cierto conocimiento de los metadatos en cuestión pero no tan granular como la estructura resultante tras el cálculo de *Silhouette Score* (ver 5.7.1.1). Resulta interesante visualizar esto a través de las nubes de palabras, complementando el análisis numérico previo.(5.6).

5.7.2. IDE Uruguay

El objetivo de esta subsección es presentar un análisis exploratorio guiado por la métrica Silhouette Score (ver 5.7.1.1) y según resultados arrojados por métrica Davies-Bouldin, Calinsky-Harabasz y *Elbow Method* (ver 5.7.2.2). Cabe destacar que este último análisis se presenta de forma conjunta, ya que los resultados arrojados por dichas métricas fueron los mismos.

5.7.2.1. *Clustering* según *Silhouette Score*

En las líneas que siguen se propone un análisis exploratorio orientado a comprender la estructura subyacente de los datos a través del análisis de *clustering*, realizado con k=4.

En la Figura 5.55 se observa el *clustering* para k=4, obtenido tras calcular el k óptimo de la métrica *Silhouette Score* y tras reducir la dimensionalidad de los datos mediante *Principal Component Analysis* (*PCA*).

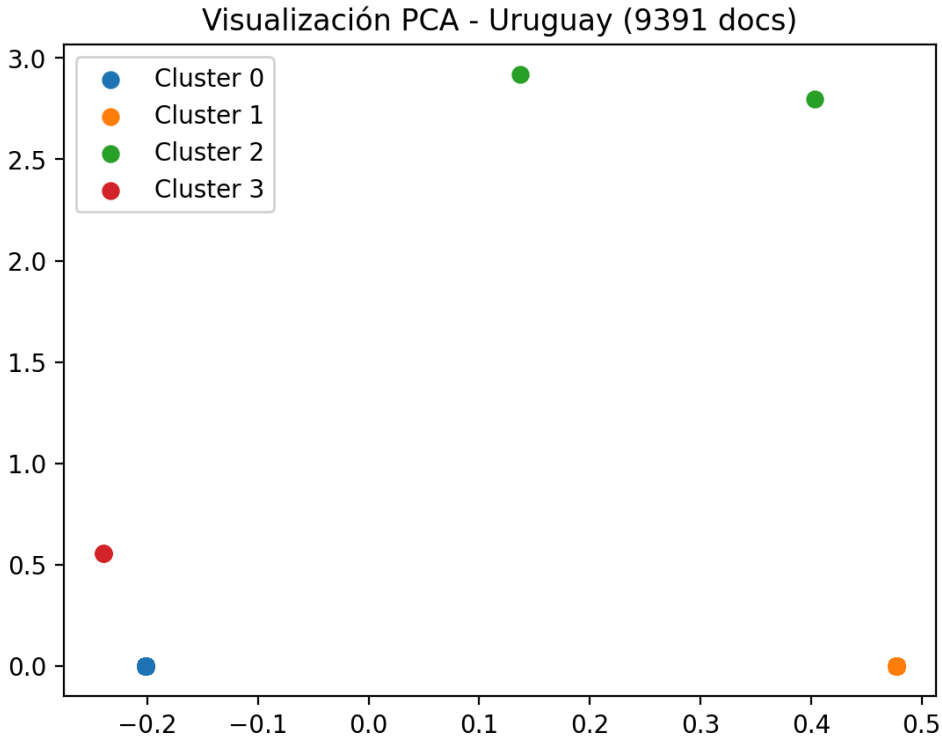


Figura 5.55: *Clustering* de la IDE de Uruguay evaluado con *Silhouette Score* (k=4)

Fuente: Elaboración propia.

En la Figura 5.50 se puede observar la distribución de documentos por *cluster* para la IDE de Uruguay.

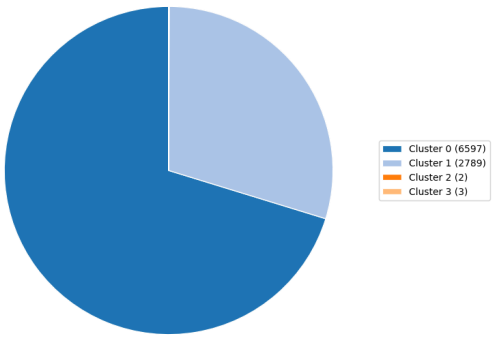


Figura 5.56: Distribución de documentos por *cluster* tras cálculo de *Silhouette Score*

Fuente: Elaboración propia.

Las Figuras 5.55 y 5.56 permiten realizar las siguientes observaciones:

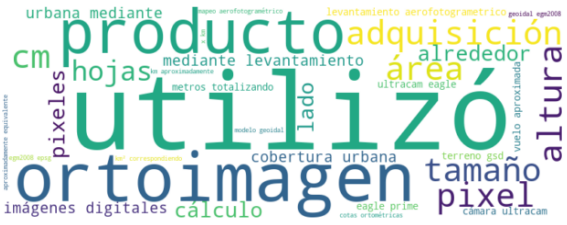
- **Coincidencia con los centroides:** La mayoría de los puntos se ubican casi exactamente en la posición de su centroide dentro de la proyección PCA, lo que refleja una alta cohesión interna. La única excepción corresponde al *cluster* 2, compuesto por solo dos documentos, que presentan una ligera separación respecto a su centroide.
- ***Clusters* con coordenadas compartidas:** Se observa que los *clusters* 0 y 1 comparten la misma coordenada en el eje y del PCA. Esto indica que, en esa dimensión, los documentos de ambos *clusters* tienen patrones muy similares
- **Cohesión interna muy alta:** Coincide con lo observado en el apartado del *Silhouette Score* (ver 5.3), confirmando que la estructura de los grupos es fuerte.
- **Estructura de datos trivial:** La disposición de los puntos en PCA se asemeja a un gráfico formado únicamente por centroides. La proyección asigna a los documentos coordenadas prácticamente idénticas.
- ***Clusters* con pocos documentos:** Los *clusters* 2 y 3 cuentan con únicamente 2 y 3 documentos respectivamente. Este aspecto requiere un análisis más profundo (véase

Capítulo 8).

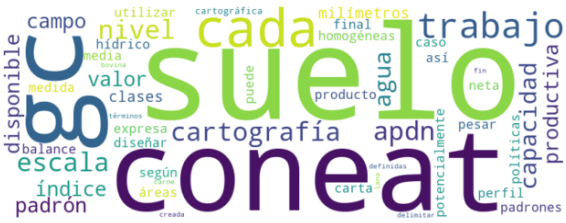
- **Estructura simple de capturar:** La alta cohesión y baja dispersión de los datos evidencian una estructura sencilla de representar mediante *clustering*.

En la Figura 5.49 se observa la nube de palabras tras realizar el *clustering* con la métrica *Silhouette Score*.

Cluster 0: Ortoimágenes y cobertura nacional



Cluster 1: Ortoimágenes y cobertura urbana



Cluster 2: Suelo y cartografía agrícola

Cluster 3: Georreferenciación

Figura 5.57: *Clusters* obtenidos según *Silhouette Score* para la IDE de Uruguay

Fuente: Elaboración propia.

- ***Cluster* 0 – Ortoimágenes y cobertura nacional:** Se centra en el uso de ortoimágenes, su tamaño en pixeles y la cobertura nacional. Refleja la gestión y análisis de imágenes geoespaciales a gran escala.
- ***Cluster* 1 – Ortoimágenes y cobertura urbana:** Contiene términos como producto, área, ortoimagen e imágenes digitales. Este *cluster* se enfoca en la cobertura urbana y el manejo de imágenes geoespaciales para zonas específicas.
- ***Cluster* 2 – Suelo y cartografía agrícola:** Incluye términos relacionados con suelo, campo, CONEAT y APDN (*Agua Potencialmente Disponible Neta*), así como cartografía. Representa la información geoespacial vinculada a análisis agrícolas y del uso del suelo.

- ***Cluster* 3 – Georreferenciación:** Se centra en la georreferenciación y referenciación de datos, destacando la importancia de la ubicación precisa y la integración de información espacial.

El análisis exploratorio mediante nubes de palabras complementa las observaciones previas: los *clusters* 0 y 1 comparten temáticas centradas en ortoimágenes. Esta similitud se refleja en la proyección PCA mostrada en la Figura 5.55, donde ambos *clusters* presentan la misma coordenada en el eje y. Este hallazgo evidencia la coincidencia temática entre los dos grupos y resulta particularmente relevante para la interpretación de los datos.

5.7.2.2. *Clustering* según Davies-Bouldin,Calinski-Harabasz y *Elbow Method*

A continuación, se lleva a cabo un análisis exploratorio orientado a identificar las tendencias y características predominantes mediante análisis de *clustering* con k=5.

En la Figura 5.55 se observa el *clustering* para k=5, obtenido tras calcular el k óptimo de la métrica Davies-Bouldin, Calinski-Harabasz y *Elbow Method* y luego de reducir la dimensionalidad de los datos mediante *Principal Component Analysis (PCA)*.

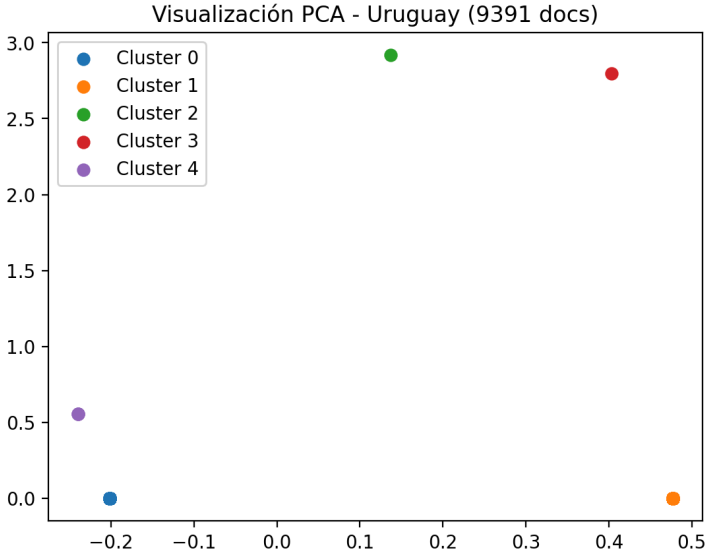


Figura 5.58: *Clustering* de la IDE de Uruguay evaluado con Davies-Bouldin, Calinski-Harabasz y *Elbow Method*

Fuente: Elaboración propia.

En la Figura 5.59 se puede observar la distribución de documentos por *cluster* para la IDE de Uruguay.

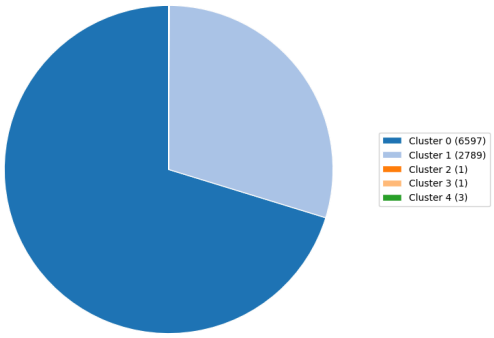


Figura 5.59: Distribución de documentos por *cluster* tras cálculo de Davies-Bouldin, Calinski-Harabasz y *Elbow Method*

Fuente: Elaboración propia.

Las Figuras 5.58 y 5.59 constituyen elementos de análisis relevantes, a partir de los cuales se desprenden las siguientes observaciones:

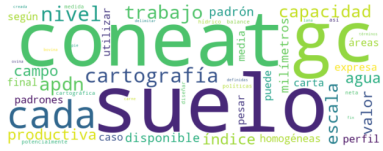
- ***Clusters* diferenciados:** Al aplanar las coordenadas, cada *cluster* se distingue claramente de los demás. Esto indica que existe un valor de k para el cual las distancias entre los grupos son todas diferentes.
- **Coincidencia con los centroides:** Los elementos de cada *cluster* coinciden con su centroide, lo que refleja una estructura interna simple y una alta cohesión de los grupos.
- **Máxima separación:** Todos los *clusters* se encuentran completamente separados. Esto sugiere que, más allá de cinco *clusters*, el modelo no lograría una división más significativa.
- ***Clusters* con coordenadas similares:** Los *clusters* 0 y 1 comparten la misma coordenada en el eje y, lo que probablemente indica una varianza similar entre ellos.

- ***Clusters* con pocos documentos:** Los *clusters* 2 y 3 contienen un solo documento cada uno, mientras que el *cluster* 4 incluye tres. Estos casos podrían corresponder a instancias aisladas y resultan convenientes de analizar en trabajos futuros (véase Capítulo 8).

En la Figura 5.60 se observa la nube de palabras tras realizar el *clustering*.

Cluster 0 – Ortoimágenes y cobertura nacional

Cluster 1 – Cobertura urbana con ortoimágenes



Cluster 2 – Suelo y cartografía agrícola

Cluster 3 – Mapas y educación agronómica

Cluster 4 – Georreferenciación y hojas aerofotogramétricas

Figura 5.60: *Clusters* obtenidos según Davies-Bouldin, Calinski-Harabasz y *Elbow Method* para la IDE de Uruguay

Fuente: Elaboración propia.

Las nubes de palabras ilustradas en la Figura 5.60 permiten describir a cada *cluster*, ya que nos brindan información descriptiva. A continuación, se realiza un análisis breve de cada uno de los *clusters* que conforman la estructura:

- ***Cluster* 0 – Ortoimágenes y cobertura nacional:** Se centra en ortoimágenes a nivel nacional, levantamientos aerofotogramétricos y productos derivados de estos estudios.
- ***Cluster* 1 – Cobertura urbana con ortoimágenes:** Incluye términos relacionados con zonas urbanas, uso de ortoimágenes y productos específicos para análisis de áreas urbanas.
- ***Cluster* 2 – Suelo y cartografía agrícola:** Contiene términos como CONEAT, escala, productiva, padrón, hídrico, campo y APDN, reflejando la cartografía y análisis de suelos

y productividad agrícola.

- ***Cluster* 3 – Mapas y educación agronómica:** Abarca términos como mapa, clases, facultad, agronomía, universidad y mapas tomados, indicando un enfoque en mapas académicos y educativos.
- ***Cluster* 4 – Georreferenciación y hojas aerofotogramétricas:** Incluye referencia, georreferenciación y hojas aerofotogramétricas, destacando la importancia de la ubicación precisa y la integración espacial.

El análisis exploratorio confirma, al igual que en el apartado anterior (ver 5.7.2.1), que los *clusters* 0 y 1 presentan una varianza similar. Este hallazgo resulta especialmente interesante, ya que ambos agrupan la mayor cantidad de información, lo que sugiere que la estructura de los datos es notablemente simple y homogénea.

En términos generales, los resultados obtenidos indican un *clustering* excepcionalmente consistente y poco habitual, en el que coinciden las métricas de *Elbow*, *Davies-Bouldin* y *Calinski-Harabasz*, reforzando la solidez del modelo y la coherencia entre los distintos métodos de evaluación aplicados.

5.7.3. Análisis comparativo

El análisis realizado de manera individual para la IDE de España (ver 5.7.1) y la IDE de Uruguay (ver 5.7.2) permite desarrollar un análisis comparativo entre ambas IDEs.

El objetivo de esta subsección es comparar elementos conceptualmente comparables con el fin de extraer hallazgos relevantes con el fin de encontrar o bien nuevos patrones o avalando resultados previos. Por este motivo, se procede a analizar el resultado de la estructura resultante tras realizar el análisis de *clustering* con el$k$ óptimo tras el cálculo de la métrica *Silhouette Score* (ver 5.7.3.1), Davies-Bouldin (ver 5.7.3.2), Calinski-Harabasz (ver 5.7.3.3) y *Elbow Method* (ver 5.7.3.4). Por último, se exponen consideraciones finales (ver 5.7.3.5).

5.7.3.1. *Silhouette Score*

En el caso de la IDE de España, tal como se observó en secciones previas, el número de *clusters* (k) es elevado y, aun así, se evidencia una superposición temática . Esto sugiere que, aunque la estructura general de los datos es razonable, existen temas que se solapan, lo cual refleja la complejidad del conjunto analizado. El análisis exploratorio del *clustering*, complementado con las nubes de palabras, permite profundizar en los detalles específicos del país. En este sentido, se identifican grupos de datos que reflejan temáticas muy particulares, lo cual aporta una comprensión más granular del dominio. Desde un punto de vista metodológico, se logra una extensión del aprendizaje de esta métrica a través del análisis de *clustering* para la IDE de España . Resulta especialmente útil cuando se requiere capturar matices o detalles finos en los datos, mostrando un buen desempeño en contextos con alta variabilidad temática como lo es el dominio de los metadatos de la IDE de España. En cuanto a la distribución de los documentos, el tamaño mínimo de los grupos es de 22 documentos (aproximadamente 1 % del total, 22/2200), lo que indica que no existen temáticas completamente aisladas. Esto refuerza la idea de una estructura relativamente equilibrada, donde incluso los grupos más pequeños mantienen cierta representatividad dentro del conjunto.

En el caso de la IDE de Uruguay, se observa una estructura de datos sólida, lo cual resulta coherente dado que se trata de un conjunto de información más sencillo y homogéneo. En la proyección PCA, se aprecia una coincidencia marcada entre los centroides, lo que evidencia una organización interna simple y altamente cohesiva, incluso más que la observada en España. A diferencia del caso de la IDE de España, en este escenario no es posible identificar con claridad los detalles específicos de cada *cluster*, ya que los temas abordados corresponden a niveles más generales. Esto se relaciona con la naturaleza de los datos, que presentan menor diversidad temática. En cuanto a la distribución de documentos, se observan dos *clusters* con información particularmente específica: el *cluster* 2, con 2 documentos, y el *cluster* 3, con 3 documentos, mostrando que la estructura resultante es menos equilibrada que en el caso de la IDE de España. Aun así, el total de 4 *clusters* resulta suficiente para representar la estructura general del conjunto. La alta homogeneidad de los datos hace que, al intentar capturar mayor detalle interno, el resultado tienda a generalidades, reflejando que las temáticas son, en

esencia, globales y poco diferenciadas. Por lo tanto, a pesar de que la métrica, en su esencia, permite capturar detalles, en este caso, nos encontramos con datos de índole genérica y con característica similares entre si.

5.7.3.2. Davies-Bouldin

En el caso de la IDE de España, la métrica de Davies-Bouldin refleja la existencia de una estructura temática compleja, con una elevada superposición entre los *clusters*. Esta métrica, que prioriza la simplicidad y la separación clara entre grupos, no logra representar adecuadamente la organización subyacente de los datos, ya que la IDE de España carece de una estructura global simple. La alta diversidad temática dificulta la formación de grupos compactos y bien definidos, lo que se traduce en valores de la métrica menos favorables. En las nubes de palabras, este comportamiento se manifiesta en la presencia de términos muy generales y numerosos subtemas que dificultan la identificación de categorías claras y deriva en que las etiquetas de cada *cluster* no sean robustas. En definitiva, el comportamiento de la métrica en este caso permite profundizar en su interpretación teórica: cuando el conjunto de datos presenta una gran variedad de temáticas, la métrica Davies-Bouldin tiende a penalizar la complejidad y a reflejar la ausencia de cohesión global.

En el caso de la IDE de Uruguay, los *clusters* se encuentran claramente diferenciados, evidenciando una estructura global simple y altamente cohesiva. Dado que la métrica de Davies-Bouldin tiende a favorecer configuraciones con grupos bien separados y de baja complejidad, los resultados obtenidos confirman que los datos presentan una organización interna homogénea y poco solapada. Este comportamiento sugiere que la estructura de los datos es intrínsecamente simple, incluso antes de la aplicación de la métrica. Por esta razón, el valor obtenido no aporta un aprendizaje teórico tan enriquecedor como en el caso de España, ya que la homogeneidad del conjunto impide observar contrastes significativos o comportamientos límite de la métrica. En síntesis, el análisis con Davies-Bouldin en Uruguay reafirma la solidez y simplicidad del agrupamiento, en consonancia con lo observado en las proyecciones PCA y en la distribución de los documentos entre *clusters*.

5.7.3.3. Calinski-Harabasz

En el caso de la IDE de España, la métrica de Calinski-Harabasz refleja una máxima separación entre los grupos con un valor de $k = 2$, lo que resulta particularmente bajo. Este comportamiento indica que, para lograr la mejor relación entre varianza intergrupal e intragrupal, el modelo tiende a generar un número reducido de *clusters*. Si bien esta configuración maximiza la separación, también implica una baja cohesión interna dentro de los grupos, lo que reduce la capacidad del modelo para capturar matices temáticos más finos. En consecuencia, el análisis resultante se torna excesivamente general, ya que agrupa información heterogénea bajo categorías amplias. En términos interpretativos, la métrica permite reforzar la comprensión de su propio funcionamiento teórico, mostrando cómo prioriza la separación global por sobre la homogeneidad interna. Sin embargo, en este caso particular, su aporte analítico es limitado debido a la complejidad temática del conjunto de datos de España, que requiere un enfoque más detallado para ser representado adecuadamente.

En el caso de la IDE de Uruguay, la métrica de Calinski-Harabasz alcanza su valor óptimo con $k = 5$, lo cual resulta razonable en función del tamaño y naturaleza del conjunto de datos. Este resultado sugiere una estructura de datos clara y bien separada, aunque de carácter general. La distribución observada en los grupos muestra que los documentos se agrupan de manera coherente, reflejando una estructura homogénea sin una gran cantidad de subtemas diferenciados. En consecuencia, el modelo captura adecuadamente la organización global de los datos, pero no profundiza en distinciones temáticas específicas, dado que el propio contenido se caracteriza por ser uniforme. En términos interpretativos, la métrica confirma la simplicidad estructural del conjunto de los metadatos de la IDE de Uruguay, donde la separación entre grupos es nítida y consistente. Esto refuerza la idea de que los datos presentan un patrón global bien definido, sin requerir una granularidad temática mayor.

5.7.3.4. Elbow Method

En el caso de la IDE de España, los resultados de la métrica *Elbow Method* muestran una clara superposición entre los grupos, lo cual evidencia una estructura de datos compleja y

parcialmente solapada. Este comportamiento permite comprender el equilibrio entre separación y cohesión interna que caracteriza a esta métrica: a diferencia de Calinski-Harabasz y Davies-Bouldin, la inercia ofrece una lectura más equilibrada entre ambos aspectos, permitiendo capturar la estructura global sin perder completamente el detalle de los subgrupos. En este sentido, la métrica aporta más información interpretativa que Calinski-Harabasz y Davies-Bouldin, aunque menos que *Silhouette Score*, ya que el nivel de solapamiento limita parcialmente la claridad de las separaciones. No obstante, su resultado confirma la existencia de una estructura densa y entrelazada, coherente con la diversidad temática observada en los metadatos de la IDE de España.

En el caso de la IDE de Uruguay, el equilibrio entre complejidad y rendimiento, que caracteriza a la métrica de inercia, no se evidencia plenamente. Sin embargo, la presencia de un *clustering* tan consistente confirma la fortaleza de la estructura interna de los datos, aunque limita la extensión del aprendizaje de la métrica.

5.7.3.5. Consideraciones finales

Los resultados de las métricas aplicadas al *clustering* de la IDE de España evidencian una extensión del aprendizaje de cada métrica, es decir, cada método produce una estructura de agrupamiento distinta. Por el contrario, en el caso de la IDE de Uruguay se observa una estructura general más homogénea, lo que provoca que las distintas métricas arrojen resultados prácticamente equivalentes. Esta falta de variación impide apreciar qué aspectos prioriza cada métrica, lo cual puede atribuirse a la naturaleza del conjunto de datos: un espacio temáticamente más uniforme, con alta cohesión interna, escasa dispersión y buena separación entre los grupos. En cambio, la IDE de España presenta mayor diversidad temática y presencia de subconjuntos diferenciados, lo que genera que cada métrica produzca configuraciones de *clustering* notablemente distintas y aporte perspectivas complementarias sobre la estructura subyacente. Por otro lado, se aprecian algunos documentados aislados en el caso de la IDE de Uruguay, probablemente casos puntuales, aunque casi la totalidad de la información refiere a Ortoimagenes (ver 5.60). Sin embargo, el análisis arrojado tras el estudio de la métrica *Silhouette Score* en el caso de la IDE de España (ver 5.7.3.1) obtenemos subtematicas

especificas y con información orientada a la gestión y planificación.

En definitiva, los metadatos de la IDE de España revelan la priorización de cada métrica en términos de cohesión, separación y dispersión, mientras que en el caso de la IDE de Uruguay evidencia una estructura estable y predecible. Así, ambas configuraciones, aunque opuestas, se complementan: una muestra la potencia diferenciadora de las métricas; la otra, los límites de su alcance ante un conjunto homogéneo.

5.8. IDE *Comparator*

IDE Comparator es el nombre elegido para la prueba conceptual, en línea con los objetivos específicos 4 y 5 (ver 4.1). En términos generales, *IDE Comparator* acompaña al investigador durante el proceso de investigación, facilitando y acelerando tareas relacionadas con la obtención, procesamiento y análisis de datos. El objetivo de esta sección es presentar de manera estructurada el propósito del prototipo (ver 5.8.1), su alcance (ver 5.8.2) y detallar cómo el prototipo guía a la investigación (ver 5.8.3).

5.8.1. Objetivo

IDE Comparator tiene como propósito centralizar la información que habitualmente se obtiene de forma independiente en distintas etapas de la investigación, unificando y consolidando los insumos necesarios para el análisis desde las fases iniciales de obtención, exploración y evaluación de los datos. De este modo, el prototipo sirve como base para realizar un estudio comparativo sobre el relacionamiento y la estructura de los datos, facilitando la evaluación de cómo se interrelaciona la información. En síntesis, su objetivo es sentar las bases para el desarrollo futuro de una herramienta más robusta, capaz de ofrecer un análisis más completo y automatizado (ver Capítulo 8).

5.8.2. Alcance

Dado que el prototipo se encuentra en una fase de experimentación, no se desarrolla ingeniería de requerimientos formal. Sin embargo, esta tarea debe ser llevada a cabo en líneas

futuras (ver Capítulo 8). El propósito principal de la prueba conceptual es demostrar la viabilidad del enfoque y analizar su factibilidad (ver Capítulo 8). Para permitir que el investigador pueda realizar un estudio comparativo sobre la relación de los datos, el prototipo centraliza todos los insumos necesarios, desde la recolección inicial mediante el *scraping* de la información, la visualización de los datos, hasta las fases de evaluación, que incluyen el análisis de métricas y el análisis de *clustering* (ver 4.2). Si bien actualmente se automatizan algunos flujos específicos, la prueba conceptual establece los lineamientos y caminos que guiarán el desarrollo de un sistema sólido y plenamente automatizado en el futuro (ver Capítulo 8). En esencia, la prueba conceptual busca centralizar información que normalmente se recopila de forma manual, ofreciendo al investigador un soporte integral para el análisis comparativo de los datos.

5.8.3. Cómo IDE Comparator guía la investigación

El desarrollo del prototipo se encuentra enmarcado en la metodología definida (ver 4.2), lo que asegura que las actividades realizadas sigan un enfoque estructurado y sistemático. Seguir una metodología de investigación es siempre recomendable, ya que permite organizar el trabajo, garantizar consistencia en los análisis y facilitar la interpretación de los resultados, al tiempo que guía al investigador en cada etapa del proceso. No obstante, estas afirmaciones requieren validación en trabajos futuros, por lo que su eficacia y aplicabilidad quedarán sujetas a análisis posteriores (ver Capítulo 8).

En definitiva, el prototipo actúa como un aliado del investigador, brindando soporte en todas las fases del análisis de datos: desde la fase de *Data Understanding* (ver 5.8.3.1), *Data Preparation* (ver 5.8.3.2), que incluye el *scraping* y preprocesamiento de texto y etapas de *Modeling* y *Evaluation* (ver 5.8.3.3).

5.8.3.1. Sobre la fase de *Data Understanding*

En la fase de *Data Understanding*, el prototipo proporciona la visualización y descarga de los metadatos de las IDEs (ver 5.8.3.1) y análisis descriptivo (ver 5.8.3.1), facilitando al investigador

la comprensión de la estructura y características de los datos antes de avanzar a las siguientes etapas del análisis.

Metadatos

El módulo de metadatos permite visualizar (ver 5.8.3.1) y descargar (ver 5.8.3.1) los registros almacenados en la base de datos. Esto permite que los datos puedan ser obtenidos sin necesidad de conectarse directamente a la base de datos, facilitando al investigador distintos tipos de análisis, incluso si únicamente desea utilizar el *scraper* o trabajar con los datos descargados.

Visualización de metadatos

En la Figura 5.61 se muestra la visualización de los metadatos en formato de tabla.

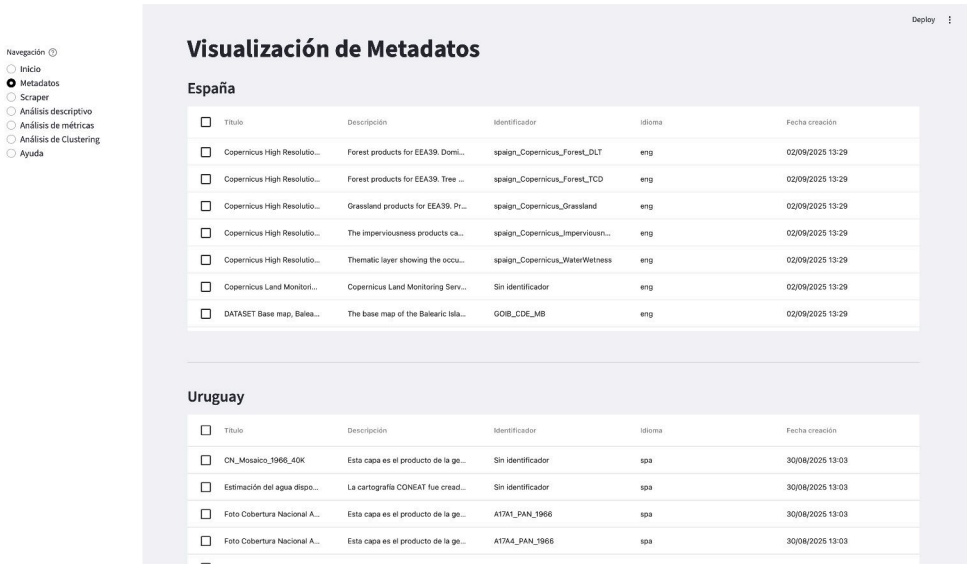
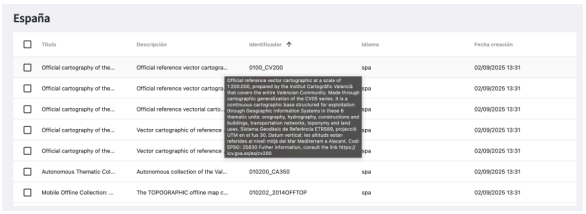


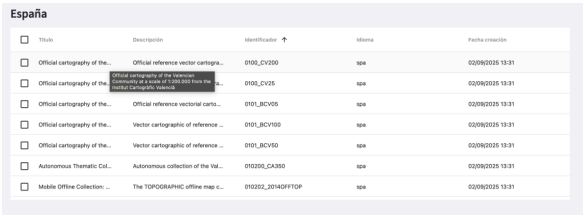
Figura 5.61: Pantalla de visualización de metadatos - IDE Comparator

Fuente: Elaboración propia.

Dado que los títulos y descripciones suelen ser extensos, la tabla incorpora *tooltips* que permiten visualizar el contenido completo, como se ilustra en la Figura 5.62.



(a) *Tooltip* de la descripción en visualización de metadatos - IDE Comparator



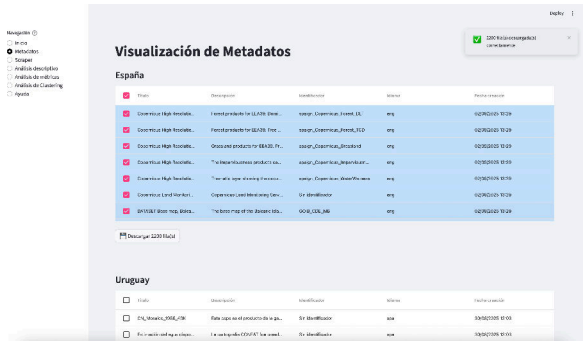
(b) *Tooltip* del título en visualización de metadatos - IDE Comparator

Figura 5.62: Visualización de *tooltips* en la tabla de metadatos - IDE Comparator.

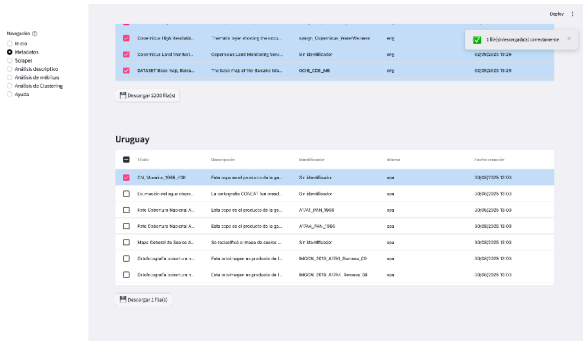
Fuente: Elaboración propia.

Descarga de metadatos

La aplicación permite descargar los metadatos, ofreciendo dos opciones o bien descargar registros seleccionados o descargar todos los registros. Una vez seleccionados los registros, la aplicación proporciona *feedback* indicando que la descarga se ha completado correctamente y mostrando la cantidad de filas descargadas (ver Figura 5.63).



(a) Descargar todos los metadatos

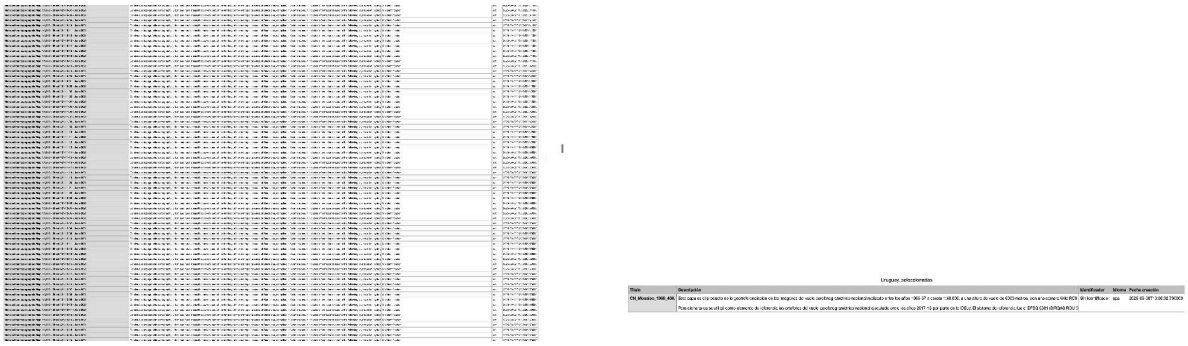


(b) Descargar un metadato

Figura 5.63: Acciones de descarga de metadatos - IDE Comparator

Fuente: Elaboración propia.

Los datos se descargan en formato CSV, permitiendo su análisis posterior de manera flexible. Ejemplos de los datos descargados para España y Uruguay se muestran en la Figura 5.64.



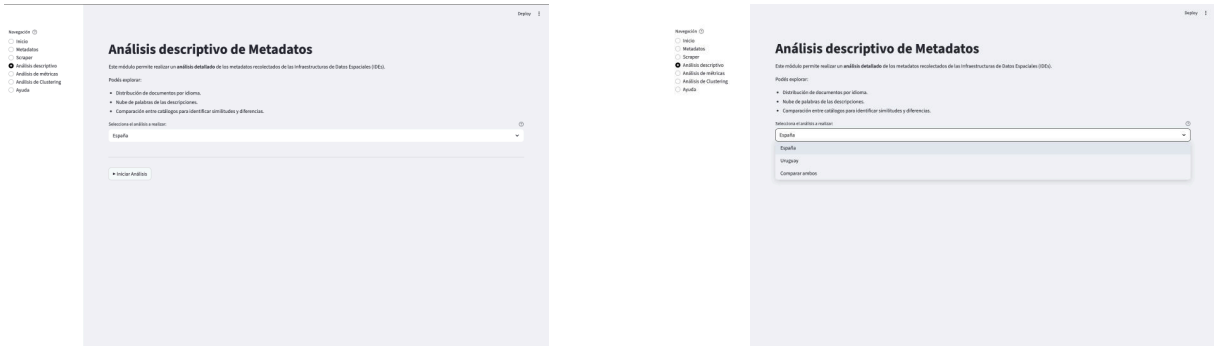
(a) Datos descargados de España (b) Datos descargados de Uruguay

Figura 5.64: Datos descargados-IDE Comparator

Fuente: Elaboración propia.

Análisis descriptivo

El objetivo del módulo de análisis descriptivo (ver Figura 5.65) es guiar al investigador, en la fase de *Data Understanding* (ver 4.2.2), en la exploración y comparación de los países definidos para el caso de estudio. Este módulo permite analizar los datos y obtener resultados que habiliten al investigador a tener un primer acercamiento de los datos.



(a) Pantalla módulo de análisis - IDE Comparator (b) Pantalla de selección de país - IDE Comparator

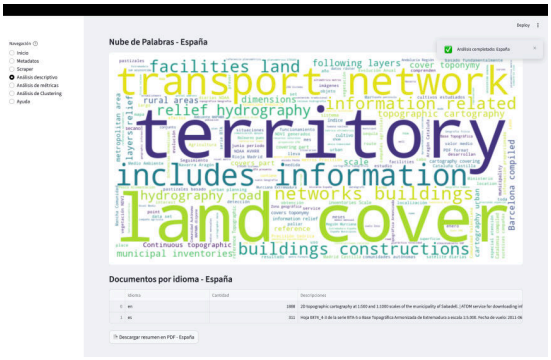
Figura 5.65: Pantallas del módulo de análisis descriptivo - IDE Comparator.

Fuente: Elaboración propia.

Se puede seleccionar un país para realizar un análisis individual o comparar los datos de ambas IDEs.

Los resultados del análisis individual (ilustrados en la Figura 5.66) constan de una nube de

palabras de los datos de la IDE y la cantidad de documentos por idioma con el fin de obtener un primer conocimiento de la información.



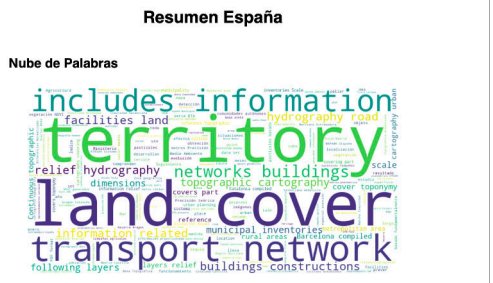
(a) Resultados país España

(b) Resultados país Uruguay

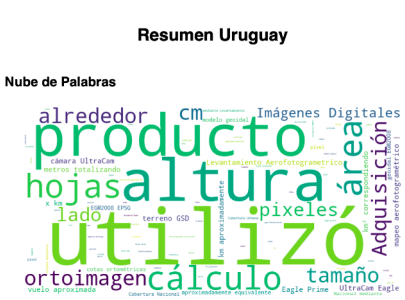
Figura 5.66: Resultados de análisis descriptivo individual - IDE Comparator

Fuente: Elaboración propia.

El PDF final contiene los mismos elementos: nube de palabras y documentos por idioma, proporcionando un marco de análisis exploratorio completo (ver Figura 5.67).



(a) PDF resultados España

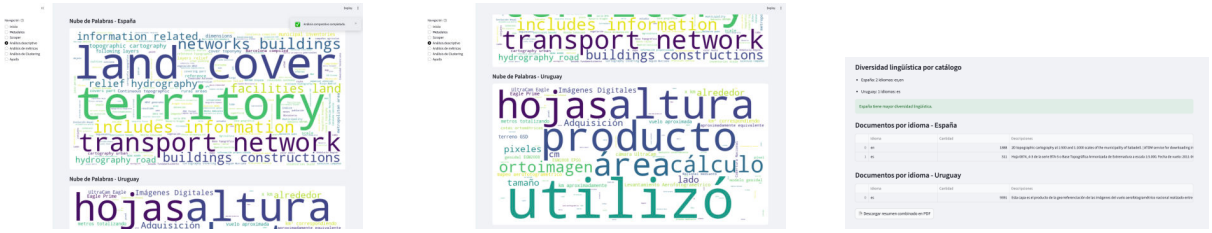


(b) PDF resultados Uruguay

Figura 5.67: Resultados de análisis descriptivo individual en PDF - IDE Comparator

Fuente: Elaboración propia.

El resultado del análisis comparativo (ver Figura 5.68) incluye, cantidad de documentos por idioma por país y una comparación en términos de diversidad lingüística.

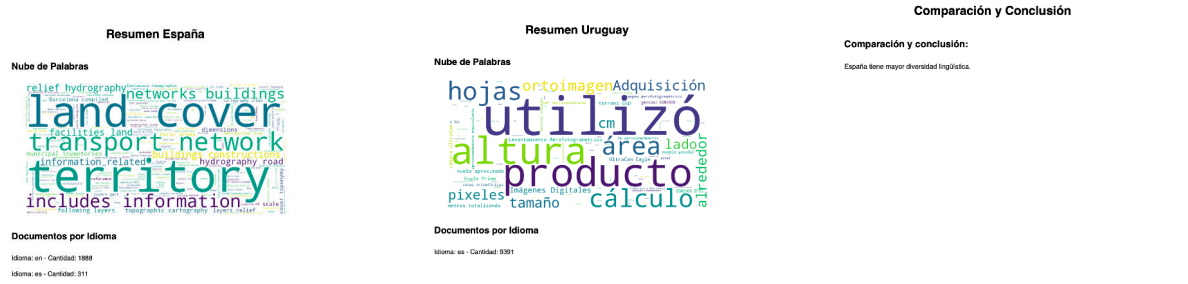


(a) IDE España (b) IDE Uruguay (c) Resultados de comparación

Figura 5.68: Resultados del análisis comparativo descriptivo - IDE Comparator.

Fuente: Elaboración propia.

El PDF comparativo refleja los mismos elementos: documentos por idioma, nube de palabras y muestra cuál de las IDEs tiene mayor diversidad en términos lingüísticos (ver Figura 5.69).



(a) PDF resultados España (b) PDF resultados Uruguay (c) Resultados de comparación en PDF

Figura 5.69: Resumen en PDF del análisis comparativo - IDE Comparator

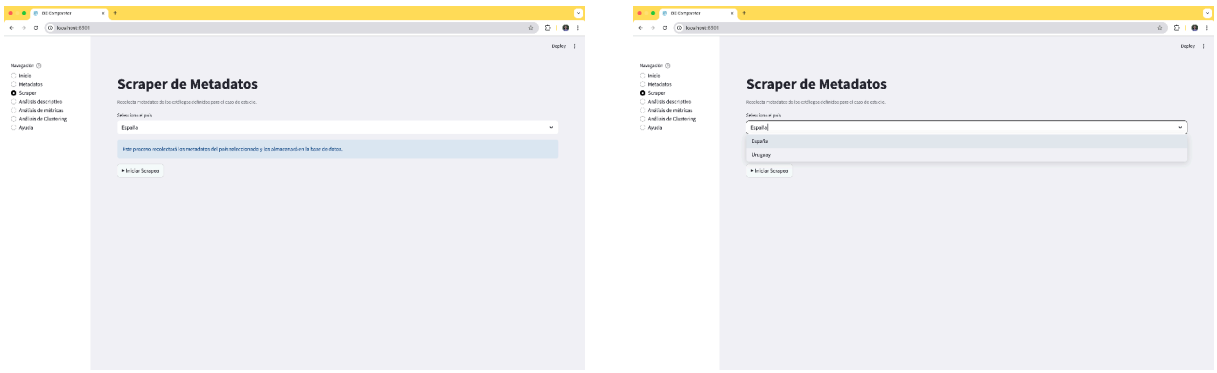
Fuente: Elaboración propia.

5.8.3.2. Sobre la fase de *Data Preparation*

El módulo Scraper consolida los resultados obtenidos del proceso de *Web Scraping* (ver 5.1), a la vez que actua de las veces de la fase de *Data Preparation* (ver 4.2.3) dentro de la metodología (ver 4.2) que se emplea.

El objetivo del módulo de Scraper es recolectar metadatos automáticamente de los países definidos en el caso de estudio, cubriendo todo el flujo desde la extracción hasta la estructuración y el preprocesamiento de los datos. Esto asegura que la información esté actualizada y libera al investigador de tareas repetitivas y propensas a errores.

El principal valor del módulo radica en la automatización del proceso, garantizando que los resultados estén siempre alineados con los datos públicos más recientes. Al ingresar al módulo, se despliega la pantalla de Scraper, como se muestra en la Figura 5.70, donde el investigador puede seleccionar el país cuyos datos desea *scrapear*.



(a) Pantalla Scraper (b) Selección de país en pantalla Scraper

Figura 5.70: Pantalla del módulo Scraper - IDE Comparator.

Fuente: Elaboración propia.

Una vez finalizado el *scrapeo*, los datos se almacenan automáticamente en la base de datos, permitiendo su posterior análisis y visualización dentro de la aplicación (ver 5.8.3.1).

5.8.3.3. Sobre la fase de *Modeling+Evaluation*

La fase de *Modeling* (ver 5.8.3.3) incluye la definición del modelo de representación vectorial a emplear en *Text Mining*. En este caso, el modelo que se utiliza está predefinido y corresponde a *Sentence Transformers* (ver 4.2.4). En trabajos futuros, podría resultar conveniente permitir la selección de distintos modelos para mayor flexibilidad (ver Capítulo 8). Sin embargo, en esta prueba conceptual se utiliza un modelo fijo.

Por otro lado, en la fase de *Evaluation* (ver 4.2.5) se emplean análisis de métricas (ver 5.8.3.3) y análisis de *clustering* (ver 5.8.3.3)

Análisis de métricas

El módulo Análisis de Métricas ilustrado en la Figura 5.71 permite explorar los datos de

cada país de forma individual o realizar una comparación entre ambos. El objetivo es brindarle al investigador un módulo orientado a la evaluación de los datos y por tanto brindando entendimiento de la estructura de los metadatos.

Figura 5.71: Pantalla principal del módulo de análisis de métricas

Fuente: Elaboración propia.

Parámetros de configuración

Antes de iniciar el análisis, se deben definir los siguientes parámetros (ilustrados en la Figura 5.71):

- **País:** País sobre el cual se realizará el análisis.
- **Límite de documentos:** Se puede trabajar con un subconjunto de documentos o con todos.
- ***k* a evaluar:** Número de *clusters* a analizar, indicando *k* mínimo y *k* máximo.
- ***Batch size*:** Tamaño del *batch* para procesar los documentos.
- **Métricas:** Métricas a evaluar (*Elbow Method*, *Silhouette Score*, Davies-Bouldin, Calinski-Harabasz).

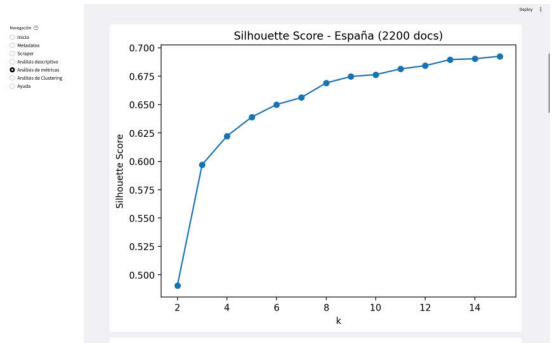
El resultado del análisis de métricas individual (ver Figura 5.72) incluye:

- **Gráficas:** se muestran gráficas de las métricas seleccionadas. Esto permite que el investigador pueda realizar un estudio de la evolución según el tamaño de la muestra

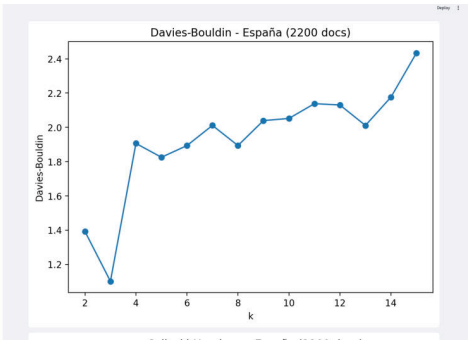
configurado y pueda analizar cómo se comportan esos datos según la métrica seleccionada.

- **Tabla resumen:** Incluye el país, cantidad de documentos, *k* óptimo, valor de la métrica y, para *Silhouette Score*, su interpretación.

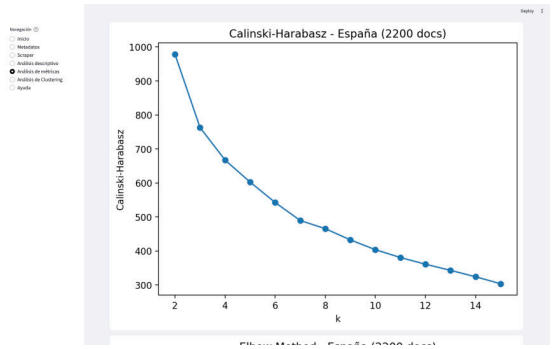
(a) Resultado análisis individual.



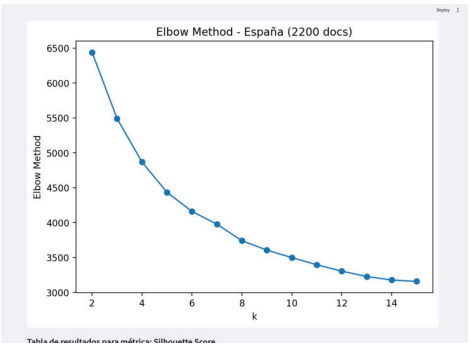
(b) Resultado análisis individual.



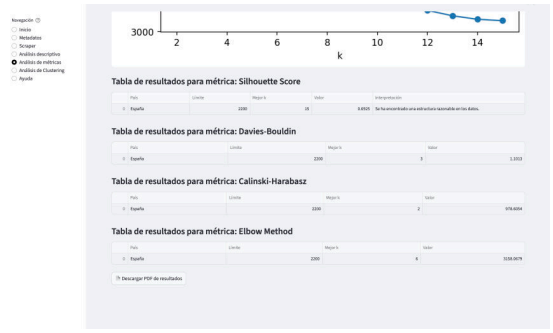
(c) Resultado análisis individual.



(d) Resultado análisis individual.



(e) Resultado análisis individual.

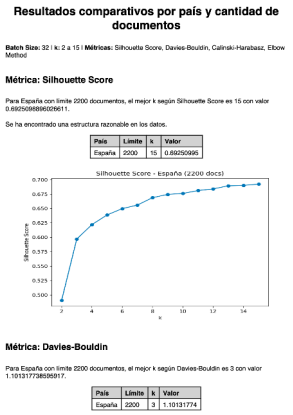


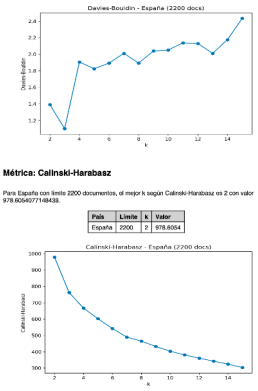
(f) Resultado análisis individual

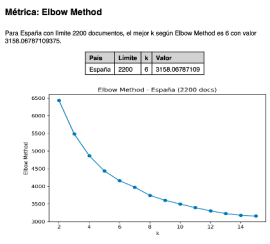
Figura 5.72: Resultados del análisis individual de análisis de métricas-IDE Comparator

Fuente: Elaboración propia.

La Figura 5.73 ilustra el resultado del resumen en pdf del análisis a nivel individual.







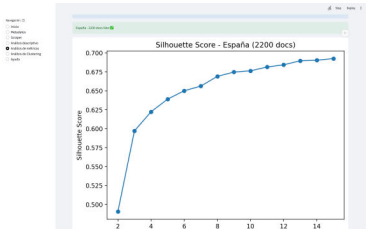
(a) Resumen PDF individual. (b) Resumen PDF individual. (c) Resumen PDF individual.

Figura 5.73: Resumen en PDF del análisis de métricas-IDE Comparator

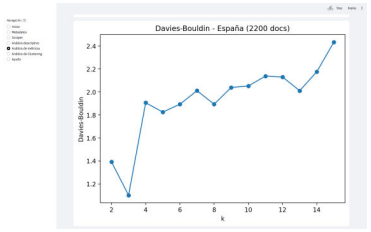
Fuente: Elaboración propia.

Por otro lado, el resultado del análisis comparativo (ver Figura 5.74) incluye:

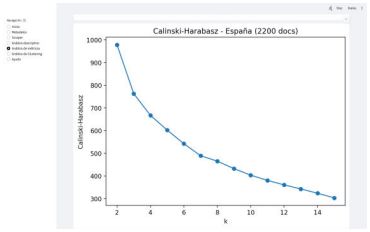
- **Gráficas:** se muestran gráficas de las métricas seleccionadas para ambos países. Esto permite que el investigador pueda realizar un estudio de la evolución según el tamaño de la muestra configurado y pueda analizar cómo se comportan esos datos según la métrica seleccionada, permitiendo profundizar en un análisis en un contexto de índole comparativo.
- **Tabla resumen:** Incluye el país, cantidad de documentos, *k* óptimo, valor de la métrica y, para *Silhouette Score*, su interpretación.



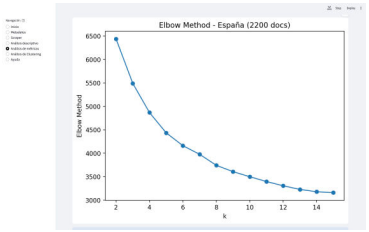
(a) Resultados del análisis comparativo de métricas.



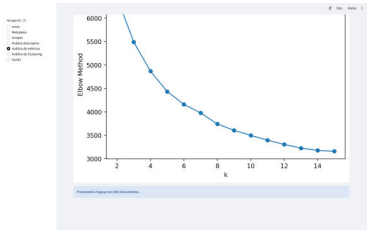
(b) Resultados del análisis comparativo de métricas.



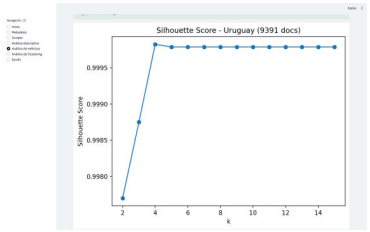
(c) Resultados del análisis comparativo de métricas.



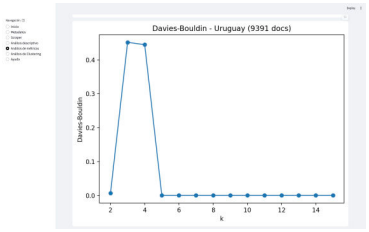
(d) Resultados del análisis comparativo de métricas.



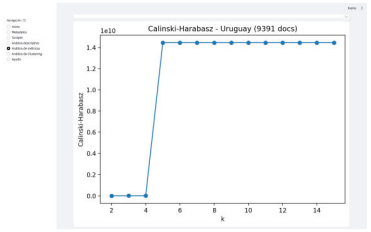
(e) Resultados del análisis comparativo de métricas.



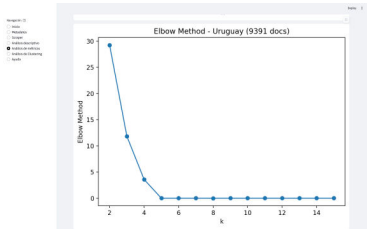
(f) Resultados del análisis comparativo de métricas.



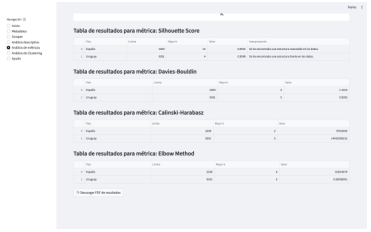
(g) Resultados del análisis comparativo de métricas.



(h) Resultados del análisis comparativo de métricas.



(i) Resultados del análisis comparativo de métricas.

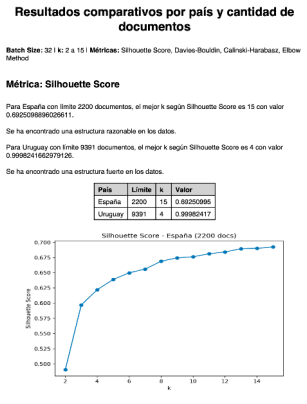


(j) Resultados del análisis comparativo de métricas.

Figura 5.74: Resultados del análisis comparativo de métricas-IDE Comparator

Fuente: Elaboración propia.

Esta información permite ser guardada en PDF con el objetivo de que el investigador pueda trabajar luego con los resultados (ver Figura 5.75).



Resultados comparativos por país y cantidad de documentos

Batch Size: 32 | **k:** 2 a 15 | **Métricas:** Silhouette Score, Davies-Bouldin, Calinski-Harabasz, Elbow Method

Métrica: Silhouette Score

Para España con límite 2200 documentos, el mejor k según Silhouette Score es 15 con valor 0.692509889602611.

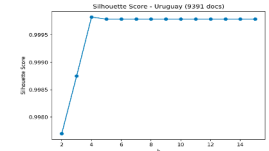
Se ha encontrado una estructura razonable en los datos.

Para Uruguay con límite 9391 documentos, el mejor k según Silhouette Score es 4 con valor 0.999824166297912 6.

Se ha encontrado una estructura fuerte en los datos.

País	Límite	k	Valor
España	2200	15	0.69250995
Uruguay	9391	4	0.99982417

(a) Resumen en PDF del análisis comparativo de métricas.



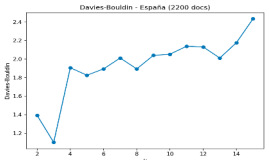
Métrica: Davies-Bouldin

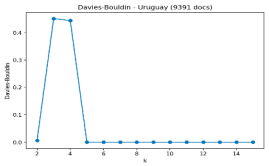
Para España con límite 2200 documentos, el mejor k según Davies-Bouldin es 3 con valor 1.1013177383698917.

Para Uruguay con límite 9391 documentos, el mejor k según Davies-Bouldin es 5 con valor 0.0001548227992795082.

País	Límite	k	Valor
España	2200	3	1.10131774
Uruguay	9391	5	0.00015482

(b) Resumen en PDF del análisis comparativo de métricas.





Métrica: Calinski-Harabasz

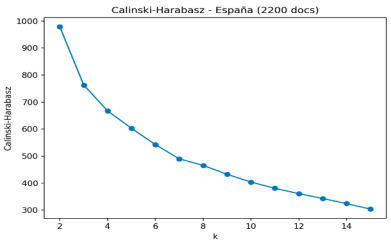
Para España con límite 2200 documentos, el mejor k según Calinski-Harabasz es 2 con valor 978.6054077148438.

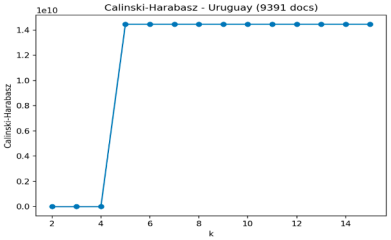
Para Uruguay con límite 9391 documentos, el mejor k según Calinski-Harabasz es 5 con valor 14452959232.0.

País	Límite	k	Valor
España	2200	2	978.6054

(c) Resumen en PDF del análisis comparativo de métricas.

Uruguay	9391	5	14452959000.0





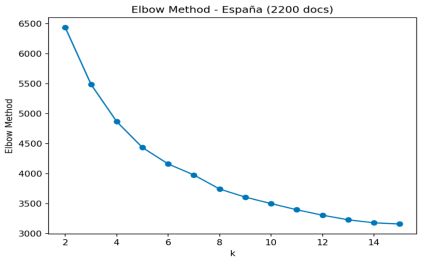
Métrica: Elbow Method

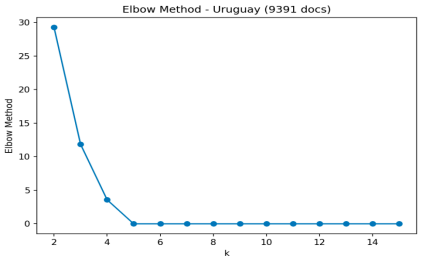
Para España con límite 2200 documentos, el mejor k según Elbow Method es 6 con valor 3158.06787109375.

Para Uruguay con límite 9391 documentos, el mejor k según Elbow Method es 5 con valor 1.4413641835631097e-08.

(d) Resumen en PDF del análisis comparativo de métricas.

País	Límite	k	Valor
España	2200	6	3158.06787109
Uruguay	9391	5	1e-08





(e) Resumen en PDF del análisis comparativo de métricas.

Figura 5.75: Resumen en PDF del análisis comparativo de métricas-IDE Comparator.

Fuente: Elaboración propia.

Análisis de *clustering*

El objetivo del módulo de análisis de *clustering* es brindarle al investigador la posibilidad de profundizar en la distribución de la información, es decir, cómo se agrupan los metadatos. En la Figura 5.76 se muestra la pantalla principal del módulo de análisis de *clustering*.

Parámetros del análisis:

- **País:** País a analizar.
- **Cantidad de documentos:** se puede elegir un subconjunto o todos los documentos disponibles.
- **Métrica y parámetros del modelo:** Métrica para calcular el k óptimo y *batch size* del modelo.

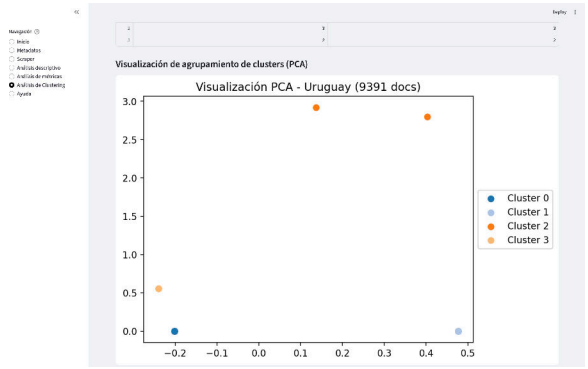
Figura 5.76: Pantalla principal del módulo de análisis de *clustering*

Fuente: Elaboración propia.

Luego de configurar los parámetros, se presiona Generar Clusters, y la aplicación muestra (ver 5.77):

- ***k* óptimo:** Se informa con qué *k* se realizó la *clusterización*.
- **Documentos por *cluster*:** Se informa la distribución de cantidad de documentos por *cluster* con el fin de generar un entendimiento de la distribución de los *clusters*.
- **PCA:** Se permite visualizar los *clusters* en un gráfica que reduce las dimensiones de los datos (PCA) para un mayor entendimiento de la agrupación de la información.

(a) Resultados de análisis de *clustering*.



(b) Resultados de análisis de *clustering*.

Figura 5.77: Resultados del análisis de *clustering*, mostrando cantidad de documentos por *cluster* y visualización en PCA.

Fuente: Elaboración propia.

5.9. Ayuda y documentación

IDE Comparator incluye una sección de ayuda y documentación, como se muestra en la Figura 5.78, disponible para los usuarios que la requieran dentro de la aplicación.

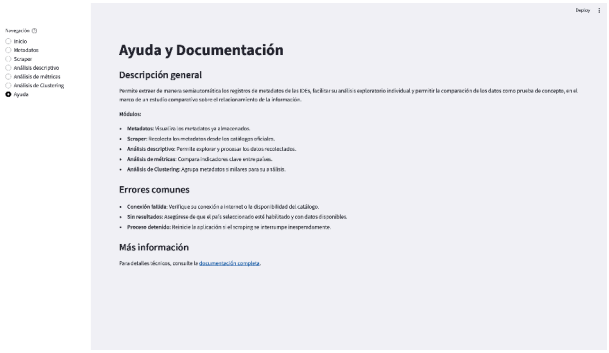


Figura 5.78: Pantalla ayuda y documentación-IDE Comparator

Fuente: Elaboración propia.

Para información más detallada sobre la operación del prototipo, así como manuales técnicos y de usuario, se remite a los anexos correspondientes (ver A.3).

6 Conclusiones

El presente proyecto ha permitido desarrollar un marco conceptual para comparar portales de Infraestructuras de Datos Espaciales en base a técnicas de *Text Mining*. Se realiza una recapitulación de los objetivos (ver 6.1), se expone el resumen de los hallazgos (ver 6.2) y las contribuciones del estudio (ver 6.3).

6.1. Recapitulación de los objetivos

A continuación, se detallan los objetivos planteados en el capítulo 4, explicando cómo cada uno fue alcanzado a lo largo del desarrollo del proyecto.

6.1.1. Objetivo General

El objetivo principal del proyecto es desarrollar un marco conceptual para comparar portales de Infraestructuras de Datos Espaciales en base a técnicas de *Text Mining*.

El objetivo se logró identificando y comparando categorías semánticamente equivalentes de cada IDE. Una vez definidas estas categorías, se procede a extraer los datos a través de *Web Scraping*, se exploran los datos a través de un análisis descriptivo. Posteriormente, se estudia la naturaleza de los datos, en particular las relaciones que existen entre ellos, es decir, determinando qué tan similares o diferentes son. Este análisis se realiza mediante métricas de validación interna y técnicas de *clustering*, lo que permite evaluar de manera objetiva la estructura y cohesión de los grupos. Finalmente, todo este flujo se integra en la herramienta IDE Comparator, que centraliza el proceso completo y facilita la visualización y comparación de los resultados obtenidos.

6.1.2. Objetivos Específicos

Cada uno de los objetivos específicos también se cumplió, como se describe a continuación:

- **OE1: Diseñar un marco conceptual, especificar los módulos que lo integran y sus**

principales características.

Se propuso llevar a cabo la identificación y relevamiento de categorías que sean semánticamente comparables de cada IDE. Luego, los datos son extraídos a través del proceso consolidado de *Web Scraping* y se exploran los datos a través de un análisis de índole exploratorio (descriptivo) con el fin de obtener un primer acercamiento a la información que se va a comparar. Una vez realizado el estudio inicial, se emplean métricas de validación interna y técnicas de *clustering*, permitiendo una evaluación objetiva de la estructura y cohesión de los *clusters* resultante. Finalmente, todo este flujo se integra en la herramienta IDE *Comparator*, que centraliza el proceso completo y facilita la visualización y comparación de los resultados obtenidos.

- **OE2: Proponer una metodología semiautomática de relevamiento de datos textuales de portales de Infraestructuras de Datos Espaciales.**

  Se desarrolló una metodología semiautomática que consiste en identificar y utilizar las APIs desde las cuales las IDEs obtienen sus datos, para luego realizar las solicitudes correspondientes mediante *requests*. Se considera semiautomática porque requiere la intervención manual para localizar los *endpoints* adecuados, pero automatiza el proceso de extracción y procesamiento posterior. La metodología releva datos textuales, ya que implementa un modelo de preprocesamiento de información que permite estructurar y normalizar el contenido obtenido.

- **OE3: Proponer una metodología para la comparación de portales de Infraestructuras en base a técnicas de *Text Mining*.**

  Se emplean técnicas de *clustering* y métricas de validación interna, buscando analizar la naturaleza de la información a través del estudio comparativo del relacionamientos de los datos. El análisis se realiza de forma individual para cada IDE y, posteriormente, se comparan los resultados obtenidos a fin de evaluar el relacionamiento de los datos de cada IDE.

- **OE4: Desarrollar una prueba de concepto y construir un prototipo para validar la propuesta.**

Para cumplir el OE4, se desarrolló un prototipo integral que centraliza todas las etapas de la metodología. Este prototipo permite ejecutar el proceso completo de *Web Scraping*, recopilando los datos, seguido de su preprocesamiento y utilizando técnicas de *Text Mining* para el estudio del relacionamiento de los datos. Posteriormente, el prototipo integra y aplica las herramientas de comparación y análisis definidas en la propuesta, facilitando la evaluación de similitudes semánticas entre descripciones y la validación de la metodología planteada. Por tanto, el prototipo no solo asegura la consistencia y reproducibilidad del flujo de trabajo, sino que también permite al investigador observar y analizar los resultados de manera directa, cumpliendo así el objetivo de construir una prueba de concepto funcional.

- **OE5: Validar el prototipo con un caso de estudio.**

  Se validó el prototipo utilizando un caso de estudio concreto, que consiste en las APIs de las IDEs de España y Uruguay. Dentro del prototipo, se cargaron específicamente las APIs mencionadas, permitiendo ejecutar el flujo completo de extracción, preprocesamiento y análisis. Posteriormente, se aplicaron las herramientas de comparación definidas en la metodología, evaluando la similitud y las relaciones semánticas entre los metadatos de ambos portales. Esta validación permitió comprobar la funcionalidad del prototipo en un contexto real y confirmar que la propuesta metodológica puede aplicarse efectivamente a distintas IDEs.

6.2. Resumen de los hallazgos

Los hallazgos más relevantes obtenidos a lo largo del desarrollo del proyecto incluyen:

- **Comparar los metadatos de las IDEs** es una herramienta valiosa en un contexto comparativo, ya que estos representan información clave sobre cada país. En particular, el análisis de las categorías de los metadatos permite comprender mejor qué tipo de información se está disponibilizando y cuáles son las áreas temáticas con mayor desarrollo o cobertura.

- **Comprender y detectar el *endpoint* de las solicitudes marcó un avance importante**, ya que permitió optimizar la eficiencia del proceso y encaminar la solución hacia una ejecución más automática y ágil.

- **El uso de MongoDB en lugar de archivos CSV representó una mejora significativa**, ya que permitió disponer de una estructura de datos ordenada, escalable y fácilmente integrable. Además, la incorporación de Docker contribuyó a garantizar la portabilidad y reproducibilidad del prototipo.

- **La elección de la *feature* de comparación resultó un hallazgo en sí misma**, ya que implicó identificar un atributo común que condensara la esencia informativa de los metadatos de cada categoría. Esta decisión permitió obtener entendimiento sólido del conjunto de datos.

- **La reducción de sesgos mediante el uso de un modelo multilingüe permitió desarrollar una metodología extensible a distintos idiomas**, lo cual representa un aporte significativo considerando la diversidad lingüística presente en el estado del arte de las IDEs. Además, el empleo de *Sentence Transformers* proporcionó resultados superiores frente a los modelos estadísticos clásicos, consolidando su efectividad para futuras aplicaciones en otras IDEs

- **La detección y eliminación de *noise* en las descripciones de la *feature* seleccionada resultó fundamental**, ya que permitió mejorar la precisión de los resultados en un entorno comparativo.

- **La posibilidad de cuantificar los datos textuales permitió entender numéricamente la organización de los grupos**, ofreciendo así una herramienta valiosa para el análisis comparativo y para extraer conclusiones sobre la distribución de la información.

- **Utilizar modelos matemáticos para comprender datos textuales permitió establecer comparaciones más precisas y objetivas**, evitando sesgos humanos. La evaluación de la curva de inercia es un ejemplo de cómo esta técnica puede cuantificar el relacionamiento de los datos, comprendiendo su complejidad y estructura.

- **El análisis de *clustering* permitió comprender la naturaleza de la información**, al revelar cómo se organizan los datos en grupos. Esto facilita un conocimiento más profundo de los metadatos y, por extensión, de las características de cada país.
- **Desarrollar un prototipo que permite asistir al investigador a lo largo de cada fase de la metodología**, garantizando un flujo de trabajo ordenado y consistente, y apoyando la ejecución de las tareas de manera eficiente.

6.3. Contribuciones del estudio

El proyecto ha generado los siguientes *outputs*:

1. **Presentación en las Jornadas del IPGH 2025 [162]:**

   Tuve la oportunidad de participar en las Jornadas del IPGH (ver Figura 6.1) el 17 de septiembre de 2025, donde compartí mi investigación a nivel nacional. Esta participación me permitió exponer mis resultados y contribuir al conocimiento local.

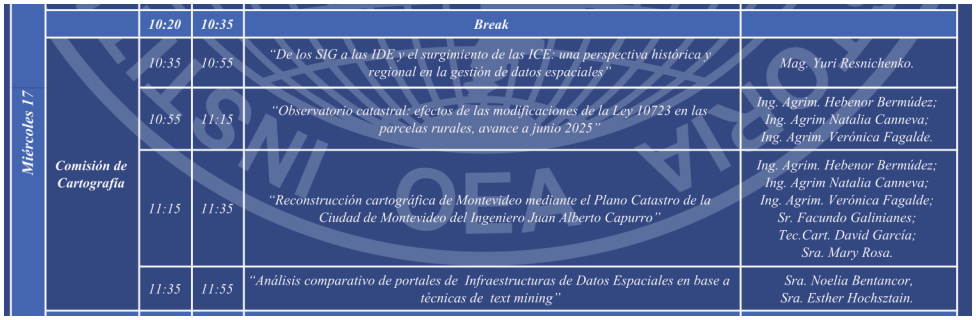


Figura 6.1: Presentación Jornadas IPGH.

Fuente: IPGH [162].

2. **Comparación de datos semánticamente comparables:**

   El marco conceptual desarrollado permite comparar datos de categorías semánticamente equivalentes, facilitando el análisis de cómo se interrelacionan y proporcionando una comprensión más profunda de la estructura de la información. Por extensión, se puede aplicar a distintos campos, por ejemplo:

- **Educación:** Comparar planes de estudio de distintas facultades. Por ejemplo, al analizar la misma carrera, se puede evaluar la estructuración de las materias para identificar similitudes, coherencia y patrones de organización.
- **Estudio de clientes por productos:** Analizar y comparar el perfil de los clientes de distintos productos, evaluando similitudes y relaciones entre consumidores y productos, para identificar patrones de comportamiento y segmentaciones relevantes.

3. **Análisis individual:**

   El marco conceptual también permite extenderse a escenarios individuales, donde se requiere comprender el relacionamiento de la información entre elementos como clientes, estudiantes o pacientes.

4. **Prototipo integrado:**

   El prototipo centraliza toda la metodología que el investigador realiza de forma separada, actuando como una guía unificada. Permite extraer y visualizar los datos en un solo lugar, eliminando procesos dispersos y redundantes, lo que simplifica el análisis y mejora la eficiencia.

7 Lecciones aprendidas

En este punto, me gustaría reflexionar sobre distintas aristas que, a mi entender, resumen las lecciones aprendidas.

Quiero detenerme en el acto de investigar, que involucró múltiples aprendizajes (ver 7.1), en la forma de trabajo, que fue clave en el proceso de investigación (ver 7.2) y en las herramientas, que me permitieron alcanzar resultados significativos (ver 7.3).

7.1. Sobre investigar

Al comenzar el proyecto, no tenía del todo claro qué involucraba investigar. Había tenido un primer acercamiento durante la materia Pasantía Académica, pero se trataba de un proyecto de corta duración. Posteriormente, cursé Metodología de la Investigación, donde aprendí sobre distintas metodologías y enfoques, adquiriendo herramientas fundamentales para abordar el proceso de investigación de manera más estructurada.

Luego de haber culminado el proyecto, miro atrás y en retrospectiva puedo decir que gran parte del tiempo la dediqué a pensar y reflexionar sobre cómo generar resultados. Sin embargo, más que los resultados en sí, lo verdaderamente significativo fueron las ideas que surgieron a lo largo del proceso. Por ejemplo, en el caso de la métrica Calinski-Harabasz, aunque los resultados no fueron del todo concluyentes, aprendí que para comparar datos de manera efectiva es preferible trabajar con métricas normalizadas o absolutas, ya que esto permite establecer comparaciones más precisas y consistentes. En este caso, el *output* más que un resultado es un aprendizaje que me permite luego tomar mejores decisiones.

Por último, la investigación implicó un estudio constante, explorando *papers* en busca de inspiración y nuevas ideas. A medida que el proyecto avanzaba, cada temática explorada revelaba nuevas aristas, generando preguntas que abrían caminos inesperados y permitían continuar profundizando, incluso cuando parecía que el tema estaba concluido, desembocando

en muchas preguntas y trabajos futuros.

7.2. Sobre la forma de trabajo

Al realizar la tesis de manera individual, encontrar una forma de trabajo fue clave. La constancia se convirtió en un pilar fundamental, ya que aprendí que, en un proyecto de investigación, las ideas se construyen de manera gradual y requieren tiempo para desarrollarse y madurar. La gestión del tiempo también resultó esencial: organizar las tareas de manera eficiente me permitió avanzar. Asimismo, la comunicación clara y el *feedback* constante jugaron un papel importante. Las reuniones semanales con mi tutora, Esther Hochsztain, aseguraban que cada semana tuviera un avance relevante, y sus comentarios me ayudaban a mejorar la presentación de los resultados y la estructura del documento. Comenzar la documentación con tiempo también fue clave, ya que me permitió aprender cómo redactar los resultados, incorporando sus sugerencias de forma temprana.

7.3. Sobre las herramientas

La elección de las herramientas resultó clave para avanzar de manera eficiente. Durante este trabajo, aprendí que cada herramienta tiene un propósito y que su utilidad depende del contexto en el que se aplique; es fundamental seleccionar lo mejor según la necesidad específica del proyecto.

Por ejemplo:

- **Python:** Profundicé en su uso y comprobé que es un lenguaje muy potente para análisis de datos, gracias a su versatilidad y a la gran cantidad de librerías disponibles. Fue muy acertado utilizar Python en un contexto de *Text Mining*.
- **Streamlit:** Aunque no es ideal para un prototipo final, resulta práctico para probar ideas conceptuales de manera rápida, permitiendo visualizar resultados de forma clara y dinámica.
- **MongoDB:** Extender y aplicar mis conocimientos en bases de datos me permitió

consolidar aprendizajes previos y llevarlos a la práctica.

- **Inteligencia Artificial:** Sirve para acelerar tareas específicas y repetitivas, como identificar *papers* relevantes o explorar literatura, aunque no reemplaza el juicio crítico ni la interpretación de los resultados del investigador.
- **LaTeX:** Fue muy útil para comunicar ideas, presentar resultados y generar diagramas claros.

En conjunto, estas herramientas no solo facilitaron el desarrollo del proyecto, sino que también ampliaron mis competencias técnicas y permitieron consolidar un enfoque más estructurado y eficiente en la investigación.

8 Futuros trabajos

En este capítulo se presentan las líneas futuras de investigación y desarrollo derivadas de los resultados obtenidos en el proyecto. Dado que se utiliza una metodología basada en CRISP-DM (ver 4.2), las propuestas de trabajo futuro se organizan siguiendo las fases de este enfoque. Al tratarse de una metodología iterativa, las mejoras y extensiones sugeridas no solo buscan perfeccionar cada fase de manera individual, sino también generar *feedback* para optimizar el ciclo completo de análisis y procesamiento de datos.

Se discuten líneas futuras en el marco de las fase de *Business Understanding* (ver 8.1), *Data Understanding* (ver 8.2), *Modeling* (ver 8.4), *Evaluation* (ver 8.5 y *Deployment* (ver 8.6).

8.1. Sobre la fase de *Business Understanding*

- **Revisión continua del estado del arte:** El modelo es iterativo y las herramientas evolucionan rápidamente, por lo que es fundamental mantener una revisión constante de la literatura y los avances en el campo.
- **Ingeniería de requerimientos:** Aplicar técnicas de recolección de requerimientos para comprender las expectativas de los usuarios y traducirlas en funcionalidades del prototipo.
- **Estudio en profundidad de *Elbow Method*:** El estudio de *elbow* para ambas IDEs arroja resultados interesantes, las derivadas de la curva de la inercia son un punto interesante de comparación. Por lo tanto, la revisión de literatura se hace necesaria para poder avanzar en investigaciones futuras.
- **Investigar métricas:** Se debe recopilar información de otras métricas de validación interna con el fin de continuar con el estudio comparativo con otras métricas.

8.2. Sobre la fase de Data *Understanding*

- **Exploración continua de los datos:** Sería valioso continuar profundizando en el análisis exploratorio de los datos, por ejemplo, identificando términos más frecuentes o patrones semánticos relevantes. Dado el carácter abierto de este tipo de estudios, la exploración constante puede revelar nueva información o relaciones que aporten una comprensión más completa de los metadatos analizados.

- **Profundizar en la comprensión de las IDEs:** Sería interesante continuar explorando qué otros aspectos pueden ser objeto de comparación entre las IDEs. Los datos disponibles ofrecen múltiples perspectivas sobre cada país, por lo que podrían incorporarse nuevas categorías o dimensiones de análisis que amplíen la mirada comparativa. Por ejemplo, podría examinarse la cantidad de datos publicados, su nivel de actualización o el grado de accesibilidad que cada IDE proporciona. Estas líneas propuestas serían muy interesantes de realizar.

8.3. Sobre la fase de *Data Preparation*

- **Eliminación de registros irrelevantes:** En algunos *clusters* de la IDE de Uruguay se observan agrupaciones con muy pocos documentos, incluso con un único registro o dos. Sería recomendable analizar estos casos, ya que la existencia de *clusters* tan reducidos puede no aportar información relevante. Asimismo, convendría evaluar la eliminación de registros que no contribuyan significativamente al análisis.

- **Iterar en *feature selection*:** En este caso, se utiliza *abstract* como elemento textual a procesar. Sin embargo, es conveniente continuar probando con otros atributos. Por ejemplo, se pueden emplear combinaciones de atributos relevantes. Se necesita continuar perfeccionando esta selección.

8.4. Sobre la fase de *Modeling*

- **Evaluación comparativa de modelos de representación textual:** Sería conveniente continuar probando distintos modelos que representen texto en espacios vectoriales, con el fin de analizar si los resultados se mantienen consistentes entre ellos. Este tipo de análisis permitiría determinar si los hallazgos dependen del modelo utilizado o si son generalizables, evitando posibles sesgos y favoreciendo una iteración continua sobre la selección del modelo más adecuado.

- **Optimización de rendimiento de modelos de representación textual:** Revisar la velocidad de procesamiento de los *embeddings* generados por *Sentence Transformers* y evaluar si el modelo actual puede mejorar en términos de eficiencia. Esto incluye comparar distintas versiones del modelo, ajustar parámetros de *batch* o utilizar técnicas de reducción de dimensionalidad para acelerar la generación de *embeddings* sin comprometer la calidad de la representación. Además, analizar el impacto del hardware en el rendimiento y documentar recomendaciones para su uso en flujos de trabajo a gran escala. Este trabajo futuro surge de la experiencia previa con el modelo, ya que en la computadora anterior que tenía, la generación de *embeddings* era muy lenta (aproximadamente media hora), mientras que tras actualizar la misma, se logró una mejora significativa en los tiempos de respuesta. Esto debe ser estudiado, ya que puede constituir una problemática a futuro.

8.5. Sobre la fase de *Evaluation*

- **Análisis comparativo en base a distintos algoritmos de *clustering*:** Sería interesante probar otros algoritmos de *clustering* para evaluar si los resultados obtenidos se mantienen y determinar si es posible obtener soluciones más robustas o eficientes. Esta comparación permitiría identificar el algoritmo más adecuado para los datos analizados.

- **Precisión de las métricas:** Conviene continuar estudiando la precisión de las métricas utilizadas, ya que pequeñas variaciones, como los decimales en los datos de Uruguay, podrían afectar los resultados. Es importante evaluar el *trade-off* entre precisión y

facilidad de visualización, ajustando los parámetros según sea necesario para optimizar ambos aspectos.

- **Refinamiento de criterios de evaluación:** Conviene continuar trabajando en la definición y ajuste de los criterios de evaluación utilizados, con el fin de garantizar que las comparaciones entre modelos o configuraciones sean consistentes y significativas. Por ejemplo, según la experiencia obtenida, resulta más relevante comparar métricas absolutas en lugar de relativas, ya que estas últimas dificultan la generación de conceptos comparables entre distintos modelos o configuraciones.
- **Extensión del uso de técnicas de *Text Mining*:** Se podrían emplear técnicas de clasificación de textos para categorizar documentos de manera supervisada, lo que permitiría contrastar los resultados obtenidos con métodos no supervisados y enriquecer el análisis de los metadatos.

8.6. Sobre la fase de *Deployment*

Esta sección explora dos aristas de los trabajos futuros: la documentación (ver 8.6.1) y el prototipo (ver 8.6.2).

8.6.1. Documentación

- **Divulgación de resultados:** Resumir los hallazgos obtenidos y presentarlos en un artículo científico, contribuyendo a la difusión del conocimiento generado por el proyecto.
- **Difusión académica:** Continuar participando en congresos, compartiendo los resultados y recibiendo *feedback* pertinente.

8.6.2. Prototipo

- **Integración en Docker Compose:** Incorporar la aplicación desarrollada en Streamlit dentro de un *Docker Compose*, de manera que todo el entorno quede completamente dockerizado y sea independiente de configuraciones externas.

- **Automatización de la obtención de registros:** Ajustar el prototipo para que la cantidad de registros de España y Uruguay se obtenga de manera automática, evitando valores *hardcodeados* en el código y facilitando la recolección dinámica de datos según las necesidades del usuario.
- **Configuración de la base de datos por el usuario:** Sería recomendable permitir que el usuario pueda seleccionar y configurar la base de datos directamente desde el prototipo, aumentando la flexibilidad y adaptabilidad de la aplicación.
- **Extensión del análisis descriptivo:** Sería posible ampliar la funcionalidad del análisis de metadatos utilizando herramientas de inteligencia artificial, como OpenAI, para comparar descripciones y extraer conclusiones sobre la orientación y características de los datos.
- **Mejora de la usabilidad para usuarios no técnicos:** Sería conveniente incluir explicaciones contextuales, como tooltips, en campos clave del prototipo (por ejemplo, el parámetro k en el comparador de estimaciones), de manera que todos los conceptos relevantes estén claramente definidos para usuarios sin experiencia técnica.
- **Implementación de registro de eventos (*logging*):** Sería recomendable incorporar *logs* que registren tanto los errores como los registros que se van cargando, lo que permitiría un seguimiento más detallado del funcionamiento del prototipo y facilitaría la detección de problemas.
- **Ampliación de compatibilidad con múltiples IDEs:** Sería recomendable extender el prototipo para que el usuario pueda registrar y analizar datos de cualquier IDE, superando la limitación actual que lo restringe al caso de estudio específico.
- **Incorporación del análisis de tasa de cambio:** Sería interesante incluir en el prototipo un módulo que calcule automáticamente la tasa de cambio y genere resultados basados en la evolución de los datos a medida que el volumen de datos aumenta. Esto permitiría evaluar cómo evolucionan los datos de cada IDE a medida que el volumen de datos aumenta con el fin de comprender la naturaleza del relacionamiento de los datos.
- **Manejo de cambios en las fuentes de datos:** Sería recomendable implementar un mecanismo que detecte cambios en las páginas o APIs de las IDEs, evitando que las

solicitudes (*requests*) estén rígidamente *hardcodeadas*. Esto reduciría el acoplamiento del prototipo y permitiría mantener su funcionalidad ante posibles modificaciones en las fuentes de datos.

- **Ingeniería de requerimientos y validación con usuarios:** Sería recomendable realizar un proceso de ingeniería de requerimientos y obtener retroalimentación de usuarios, incluyendo investigadores, para validar el prototipo. Este enfoque permitiría iterar sobre el diseño y las funcionalidades, asegurando que la aplicación cumpla con las necesidades del usuario en un contexto experimental.

- ***Labeling* automático:** Como extensión del prototipo desarrollado, se propone implementar un sistema de *labeling* automático de documentos. Esta funcionalidad permitiría automatizar procesos manuales y por tanto poder entender las principales temáticas subyacentes de cada IDE, evitando el trabajo manual del investigador.

- **Código:** El código debe ser refactorizado para poder aumentar la legibilidad y mantenibilidad.

Bibliografía

[1] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp-dm 1.0: Step-by-step data mining guide," CRISP-DM Consortium, Tech. Rep., 2000, technical report. [Online]. Available: https://www.the-modeling-agency.com/crisp-dm.pdf

[2] R. Mitchell, *Web Scraping with Python: Collecting More Data from the Modern Web*. O'Reilly Media, Inc, 2018.

[3] K. Parikh, D. Singh, D. Yadav, and M. Rathod, "Detection of web scraping using machine learning," *Open access international journal of Science and Engineering*, vol. 3, no. 1, pp. 114–118, 2018.

[4] Q. Niu, I. A. Kandhro, A. Kumar, S. Shah, M. Hasan, H. M. Ahmed, and F. Liang, "Web scraping tool for newspapers and images data using jsonify," *Journal of Applied Science and Engineering*, vol. 26, pp. 465–474, junio 2022. [Online]. Available: https://doi.org/10.6180/jase.202304_26(4).0002

[5] R. Lawson, *Web scraping with Python*. Packt Publishing Ltd, 2015.

[6] R. Banerjee, "Website scraping," *Happiest Minds Technologies*, 2014.

[7] K. V. Rajkumar, K. Sri Nithya, C. T. Sai Narasimha, V. Shariff, V. J. Manasa, and N. S. Koti Mani Kumar Tirumanadham, "Scalable web data extraction for xtree analysis: Algorithms and performance evaluation," in *2024 Second International Conference on Inventive Computing and Informatics (ICICI)*, 2024, pp. 447–455.

[8] A. V. Saurkar, K. G. Pathare, and S. A. Gode, "An overview on web scraping techniques and tools," *International Journal on Future Revolution in Computer Science & Communication Engineering*, vol. 4, no. 4, pp. 363–367, 2018.

[9] E. J. Farley and L. Pierotte, "An emerging data collection method for criminal justice researchers," *Justice Research and statistics association*, pp. 1–9, 2017.

[10] E. Persson, “Evaluating tools and techniques for web scraping,” 2019.

[11] S. D. S. Sirisuriya, “Importance of web scraping as a data source for machine learning algorithms-review,” in *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2023, pp. 134–139.

[12] S. A. Khan and S. Bhide, “Web scrapping tools used in healthcare sector,” *International Journal For Multidisciplinary Research*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:273465421

[13] S. M, S. S. B, and M. R, “A novel approach for news extraction using webscraping technique,” *NCICCNDA*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:186435667

[14] C. Lotfi, S. Srinivasan, M. Ertz, and I. Latrous, *Web Scraping Techniques and Applications: A Literature Review*. SCRS, 2021, pp. 381–394.

[15] I. Williamson, A. Rajabifard, and M. Feeney, *Developing Spatial Data Infrastructures: From Concept to Reality*. CRC Press, 2003. [Online]. Available: https://books.google.com.uy/books?id=OmuMV-tstRkC

[16] S. Schweers, K. E. Kinder-Kurlanda, S. Müller, and P. Siegers, “Conceptualizing a spatial data infrastructure for the social sciences: An example from germany,” *Journal of Map & Geography Libraries*, vol. 12, pp. 100 – 126, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:130339609

[17] D. J. Saab, “A conceptual investigation of the ontological commensurability of spatial data infrastructures among different cultures,” *Earth Science Informatics*, vol. 2, pp. 283–297, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:42672381

[18] L. Pashova and T. Bandrova, “A brief overview of current status of european spatial data infrastructures − relevant developments and perspectives for bulgaria,” *Geo-spatial Information Science*, vol. 20, pp. 108 – 97, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:81990474

[19] A. Rajabifard and I. P. Williamson, “Spatial data infrastructures: an initiative to facilitate spatial data sharing,” *Global environmental databases-present situation; future directions*, vol. 2, 2002.

[20] R. L. R. Borba, “Ecossistema para infraestrutura de dados espaciais híbrida, coproduzida, colaborativa, convergente e compartilhável,” Tese de Doutorado, Universidade Federal do Rio de Janeiro, COPPE, Rio de Janeiro, Mar. 2017, orientador: Jano Moreira de Souza.

[21] J. McLaughlin and S. Nichols, “Developing a national spatial data infrastructure,” *Journal of Surveying Engineering*, vol. 120, no. 2, pp. 62–76, 1994.

[22] S. K. S. Paixão, S. Nichols, and D. Coleman, “Towards a spatial data infrastructure: Brazilian initiatives,” *Revista Brasileira de Cartografia*, vol. 60, no. 2, 2009.

[23] M. Basaraner, “Revisiting cartography: towards identifying and developing a modern and comprehensive framework,” *Geocarto International*, vol. 31, pp. 71–91, 01 2016.

[24] I. P. Williamson, A. Rajabifard, and A. Binns, “Challenges and issues for sdi development,” *International Journal of Spatial Data Infrastructures Research*, vol. 1, pp. 24–35, 2006.

[25] S. Calzati and B. van Loenen, “A fourth way to the digital transformation: The data republic as a fair data ecosystem,” *Data & Policy*, vol. 5, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259132389

[26] J. Crompvoets, A. Bregt, A. Rajabifard, and I. Williamson, “Assessing the worldwide developments of national spatial data clearinghouses,” *International Journal of Geographical Information Science*, vol. 18, no. 7, pp. 665–689, 2004. [Online]. Available: https://doi.org/10.1080/13658810410001702030

[27] I. Masser, “All shapes and sizes: The first generation of national spatial data infrastructures,” *International Journal of Geographical Information Science*, vol. 13, pp. 67–84, 01 1999.

[28] S. Coetzee and B. Wolff-Piggott, “A review of sdi literature: Searching for signs of inverse infrastructures,” in *Cartography-Maps Connecting the World*, ser. Lecture Notes in Geoinformation and Cartography. Springer, 2015, pp. 113–127.

[29] A. Rajabifard, M.-E. Feeney, and I. Williamson, “Future directions for sdi development,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 2002, pp. 11–22, 08 2002.

[30] D. J. Coleman and J. D. McLaughlin, “Defining global geospatial data infrastructure (ggdi): components, stakeholders and interfaces,” *Geomatica*, vol. 52, no. 2, pp. 129–143, 1998.

[31] M. Craglia and A. Annoni, “Inspire: An innovative approach to the development of spatial data infrastructures in europe,” in *Research and Theory in Advancing Spatial Data Infrastructure Concepts*, H. Onsrud, Ed. Redlands, California: ESRI Press, 2007, pp. 7–32.

[32] I. Masser, “What is a spatial data infrastructure?” in *GSDI 4 Capetown*, 01 2000.

[33] C. Cömert, “Web services and national spatial data infrastructure (nsdi),” in *Proceedings of Geo-Imagery Bridging Continents, XXth ISPRS Congress*, vol. XXXV, Istanbul, Turkey, July 2004, pp. 12–23.

[34] L. Grus, J. Crompvoets, and A. Bregt, “Defining national spatial data infrastructures as complex adaptive systems,” in *GSDI-9 Conference Proceedings*, Santiago, Chile, 2006.

[35] I. Masser, “Emerging frameworks in the information age: The spatial data infrastructure (sdi) phenomenon,” in *The SAGE Handbook of GIS and Society*. London: Sage Publications, 2011, pp. 271–286.

[36] D. Vandenbroucke and K. Janssen, “Spatial data infrastructures in europe: State of play spring 2005,” European Commission (EUROSTAT & DGENV), Summary report of a study commissioned by the EC, 2005.

[37] K. Layne and J. Lee, “Developing fully functional e-government: A four stage model,” *Government Information Quarterly*, vol. 18, no. 2, pp. 122–136, 2001.

[38] M. Warnest, A. Rajabifard, and I. P. A. Williamson, “Collaborative approach to building national sdi in federated state systems: case study of australia,” in *GSDI-8 Conference*, Cairo, Egypt, 2005.

[39] N. R. Budhathoki, B. Bruce, and Z. Nedovic-Budic, “Reconceptualizing the role of the user of spatial data infrastructure,” *GeoJournal*, vol. 72, no. 3–4, pp. 149–160, 2008.

[40] I. Williamson, A. Rajabifard, J. Wallace, and R. Bennett, “Spatially enabled society,” in *Proceedings of the FIGURE Congress 2010, Facing the Challenges - Building the Capacity*, Sydney, Australia, 2011.

[41] D. J. Coleman, Y. Georgiadou, and J. Labonte, “Volunteered geographic information: the nature and motivation of produsers,” *International Journal of Spatial Data Infrastructures Research*, vol. 4, no. 1, pp. 332–358, 2009.

[42] S. Hennig and M. Belgui, “User-centric sdi: Addressing users requirements in third-generation sdi, the example of nature-sdiplus,” *Geoforum Perspektiv*, vol. 10, no. 20, 2012.

[43] S. Enemark and A. Rajabifard, “Spatially enabled society,” *Geoforum Perspektiv*, vol. 10, no. 20, 2012.

[44] S. Hennig, I. Gryl, and R. Vogler, “Spatial data infrastructures, spatially enabled society and the need for society’s education to leverage spatial data,” *International Journal of Spatial Data Infrastructures Research*, vol. 8, pp. 98–127, 2013.

[45] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2006.

[46] T. D. Abdusalomovna *et al.*, “Text mining,” *European Journal of Interdisciplinary Research and Development*, vol. 13, pp. 284–289, 2023.

[47] N. Altman and M. Krzywinski, “Points of significance: Clustering,” *Nature Methods*, vol. 14, pp. 545–546, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:7997450

[48] S. Chen, L. Wu, and J. Zhuo, “The application of unsupervised learning tf-idf algorithm in word segmentation of ideological and political education,” *Wireless Communications and Mobile Computing*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251476378

[49] W. Medhat, A. H. Yousef, and H. K. Mohamed, “Corpora preparation and stopword list generation for arabic data in social network,” *ArXiv*, vol. abs/1410.1135, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:15266684

[50] A. R. Rivas, E. L. Iglesias, M. Lourdes, and B. Diz, “Study of query expansion techniques and their application in the biomedical information retrieval,” *The Scientific World Journal*, vol. 2014, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:15648850

[51] C. Baroukh, S. L. Jenkins, R. Dannenfelser, and A. Ma’ayan, “Genes2wordcloud: a quick way to identify biological themes from gene lists and free text,” *Source Code for Biology and Medicine*, vol. 6, pp. 15 – 15, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:108823

[52] G. Schoier, G. Borruso, and P. Tossut, “A text mining analysis on big data extracted from social media,” *Computational Science and Its Applications – ICCSA 2020*, vol. 12252, pp. 351 – 364, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:222005867

[53] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: A review,” *ACM Computing Surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[54] T. W. Liao, “Clustering of time series data—a survey,” *Pattern Recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[55] I. Bose and X. Chen, “Detecting the migration of mobile service customers using fuzzy clustering,” *Information & Management*, vol. 52, no. 2, pp. 227–238, 2015.

[56] D. Grant and B. Yeo, “A global perspective on tech investment, financing, and ict on manufacturing and service industry performance,” *International Journal of Information Management*, vol. 43, pp. 130–145, 2018.

[57] S. Samoilenko and K.-M. Osei-Bryson, “Representation matters: an exploration of the socio-economic impacts of ict-enabled public value in the context of sub-saharan economies,” *International Journal of Information Management*, vol. 49, pp. 69–85, 2019.

[58] W.-B. Xie, Y.-L. Lee, C. Wang, D.-B. Chen, and T. Zhou, “Hierarchical clustering supported by reciprocal nearest neighbors,” *Information Sciences*, vol. 539, pp. 297–307, 2020.

[59] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 2005.

[60] R. Xu and D. C. Wunsch, *Clustering*. John Wiley & Sons, 2009.

[61] G. J. Oyewole and G. A. Thopil, “Data clustering: application and trends,” *Artificial intelligence review*, vol. 56, no. 7, pp. 6439–6475, 2023.

[62] F. Schwenker and E. Trentin, “Pattern classification and clustering: a review of partially supervised learning approaches,” *Pattern Recognition Letters*, vol. 37, pp. 4–14, 2014.

[63] W. Härdle, L. Simar, and M. Fengler, *Applied Multivariate Statistical Analysis*. Springer International Publishing, 2024. [Online]. Available: https://books.google.com.uy/books?id=6SMlEQAAQBAJ

[64] E. Braun, B. R. H. Geurten, and M. Egelhaaf, “Identifying prototypical components in behaviour using clustering algorithms,” *PLoS ONE*, vol. 5, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:10790752

[65] S. P. Lloyd, “Least squares quantization in pcm,” *IEEE Trans. Inf. Theory*, vol. 28, pp. 129–136, 1982. [Online]. Available: https://api.semanticscholar.org/CorpusID:10833328

[66] J. Mcqueen, “Some methods for classi cation and analysis of multivariate observations,” in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967. [Online]. Available: https://api.semanticscholar.org/CorpusID:235144755

[67] M. Shutaywi and N. N. Kachouie, “Silhouette analysis for performance evaluation in machine learning with applications to clustering,” *Entropy*, vol. 23, no. 6, p. 759, 2021.

[68] D. L. Nkweteyim, “Clustering by partitioning around medoids using distance-based similarity measures on interval-scaled variables,” *Nigerian Journal of Technological Development*, vol. 15, pp. 1–6, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:139931019

[69] D. Arthur and S. Vassilvitskii, “k-means++: The advantages of careful seeding,” in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA ’07)*, ACM-SIAM. New Orleans, LA, USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[70] ——, “K-means++: The advantages of careful seeding,” in *SODA ’07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, vol. 8, 01 2007, pp. 1027–1035.

[71] D. G. Cortés, E. Onieva, I. P. López, L. Trinchera, and J. Wu, “Autoencoder-enhanced clustering: A dimensionality reduction approach to financial time series,” *IEEE Access*, vol. 12, pp. 16 999–17 009, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267348405

[72] U. G. Inyang, U. A. Umoh, I. Nnaemeka, and S. A. Robinson, “Unsupervised characterization and visualization of students’ academic performance features,” *Comput. Inf. Sci.*, vol. 12, pp. 103–116, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:155424508

[73] A. Ullah, K. Haydarov, I. ul Haq, K. Muhammad, S. Rho, M. Y. Lee, and S. W. Baik, “Deep learning assisted buildings energy consumption profiling using smart meter data,” *Sensors (Basel, Switzerland)*, vol. 20, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211078877

[74] C. Hennig, “What are the true clusters?” *Pattern Recognition Letters*, vol. 64, pp. 53–62, 2015.

[75] G. Milligan and M. Cooper, “An examination of procedures for determining the number of clusters in a data set,” *Psychometrika*, vol. 50, pp. 159–179, 1985.

[76] A. Bagirov, R. Aliguliyev, and N. Sultanova, “Finding compact and well-separated clusters: Clustering using silhouette coefficients,” *Pattern Recognition*, vol. 135, 2023.

[77] D. M. Saputra, D. Saputra, and L. D. Oswari, “Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method,” in *Sriwijaya international conference on information technology and its applications (SICONIAN 2019)*. Atlantis Press, 2020, pp. 341–346.

[78] S. M. Zobaed, R. N. Gottumukkala, and M. A. Salehi, “Privacy□preserving clustering of unstructured big data for cloud□based enterprise search solutions,” *Concurrency and Computation: Practice and Experience*, vol. 34, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:218869768

[79] A. D. Topal and A. K. Geçer, “Examination of student satisfaction with e-courses by clustering analysis,” *Innoeduca. International Journal of Technology and Educational Innovation*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266254111

[80] A. T. Muharram, R. E. Nalawati, B. Warsuta, I. M. Malik Matin, A. Pradiptyas, and M. Natanael, “Comparison of elbow, silhouette and dbi methods for clustering nutritional status of toddlers using k-means clustering,” in *2024 12th International Conference on Cyber and IT Service Management (CITSM)*, 2024, pp. 1–6. [Online]. Available: https://doi-org.proxy.timbo.org.uy/10.1109/CITSM64103.2024.10775935

[81] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, “A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, no. 1, p. 31, Feb 2021. [Online]. Available: https://doi.org/10.1186/s13638-021-01910-w

[82] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[83] E. Villatoro-Tello, S. Parida, P. Motlícek, and O. Bojar, “Inferring highly-dense representations for clustering broadcast media content,” *Prague Bull. Math. Linguistics*, vol. 115, pp. 31–50, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:229010328

[84] L. E. E. Awong and T. Zielinska, “Comparative analysis of the clustering quality in self-organizing maps for human posture classification,” *Sensors (Basel, Switzerland)*, vol. 23, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:262072579

[85] P. Rousseeuw, M. Hubert, and A. Struyf, “Clustering in an object-oriented environment,” *Journal of Statistical Software*, vol. 01, 02 1997.

[86] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 2, p. 224–227, Feb. 1979. [Online]. Available: https://doi.org/10.1109/TPAMI.1979.4766909

[87] R. Punhani, V. Arora, A. Sai Sabitha, and V. K. Shukla, “Segmenting e-commerce customer through data mining techniques,” *Journal of Physics: Conference Series*, vol. 1714, no. 1, p. 012026, jan 2021. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1714/1/012026

[88] E. Ansari, M. H. Sadreddini, B. S. Bigham, and F. Alimardani, “A combinatorial cooperative-tabu search feature reduction approach,” *Operations Research eJournal*, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:120291885

[89] F. Yao, J. Coquery, and K.-A. L. Cao, “Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets,” *BMC Bioinformatics*, vol. 13, pp. 24 – 24, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:17336655

[90] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y. Wang, and J.-B. Huang, “A closer look at few-shot classification,” *ArXiv*, vol. abs/1904.04232, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:102351185

[91] N. Flann, M. Recker, B. Xu, X. Qi, and L. Ye, “Clustering educational digital library usage data: A comparison of latent class analysis and k-means algorithms,” *Journal of Educational Data Mining*, vol. 5, pp. 38–68, 2013.

[92] R. E. Febrita, W. F. Mahmudy, and A. P. Wibawa, “High dimensional data clustering using self-organized map,” *Knowl. Eng. Data Sci.*, vol. 2, pp. 31–40, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:198325877

[93] G. E. Raptis, C. P. Katsini, C. Alexakos, A. P. Kalogeras, and D. N. Serpanos, “Cavectir: Matching cyber threat intelligence reports on connected and autonomous vehicles using machine learning,” *Applied Sciences*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253659236

[94] M. Oliveira and A. R. S. Marçal, “Clustering lidar data with k-means and dbscan,” in *International Conference on Pattern Recognition Applications and Methods*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257358151

[95] J. C. R. Thomas, M. S. Peñas, and M. Mora, “New version of davies-bouldin index for clustering validation based on cylindrical distance,” *2013 32nd International Conference of the Chilean Computer Science Society (SCCC)*, pp. 49–53, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:13035201

[96] Y. Suh, “Discovering customer segments through interaction behaviors for home appliance business,” *J. Big Data*, vol. 12, p. 57, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:276877259

[97] G. K. Nave, S. Padhee, A. Alambo, T. Banerjee, N. R. Shah, and D. M. Abrams, “Clustering of pain dynamics in sickle cell disease from sparse, uneven samples,” *ArXiv*, vol. abs/2108.13963, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237365072

[98] D. S. Renjith, A. Sreekumar, and M. Jathavedan, “An empirical research and comparative analysis of clustering performance for processing categorical and numerical data extracts from social media,” *Acta Scientiarum. Technology*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247876493

[99] M. Aljeri, “Big data-driven approach to analyzing spatio-temporal mobility pattern,” *IEEE Access*, vol. 10, pp. 98 414–98 426, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252325757

[100] S. Suraya, M. Sholeh, and U. Lestari, “Evaluation of data clustering accuracy using k-means algorithm,” *International Journal of Multidisciplinary Approach Research and Science*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266488584

[101] I. Lorencin, D. Frank, N. Tanković, and T. Horvat, “K-means clustering of intracellular calcium signal transients,” *Acta Polytechnica Hungarica*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268966082

[102] S. Hatamikia, K. Maghooli, and A. M. Nasrabadi, “The emotion recognition system based on autoregressive model and sequential forward feature selection of electroencephalogram signals,” *Journal of Medical Signals and Sensors*, vol. 4, pp. 194 – 201, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:37932248

[103] G. Liu, “A new index for clustering evaluation based on density estimation,” *ArXiv*, vol. abs/2207.01294, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250264477

[104] H. You and G. Rumbe, “Comparative study of classification techniques on breast cancer fna biopsy data,” *Int. J. Interact. Multim. Artif. Intell.*, vol. 1, pp. 5–12, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:7763719

[105] J. Sublime, A. Troya-Galvis, and A. Puissant, “Multi-scale analysis of very high resolution satellite images using unsupervised techniques,” *Remote. Sens.*, vol. 9, p. 495, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:19124648

[106] H. Yin, A. Aryani, S. Petrie, A. Nambissan, A. Astudillo, and S. Cao, “A rapid review of clustering algorithms,” *ArXiv*, vol. abs/2401.07389, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:266999735

[107] G. jiang Duan and C. Zou, “A clustering effectiveness measurement model based on merging similar clusters,” *PeerJ Computer Science*, vol. 10, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268196691

[108] O. Chikumbo and V. Granville, “Optimal clustering and cluster identity in understanding high-dimensional data spaces with tightly distributed points,” *Mach. Learn. Knowl. Extr.*, vol. 1, pp. 715–744, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:198351463

[109] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, “Disentangling image distortions in deep feature space,” *Pattern Recognit. Lett.*, vol. 148, pp. 128–135, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211507167

[110] X. Han, C. Armenakis, and M. Jadidi, “Modeling vessel behaviours by clustering ais data using optimized dbscan,” *Sustainability*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237701939

[111] M. Zhai, S. Wang, Y. Wang, and D. Wang, “An interpretable prediction method for university student academic crisis warning,” *Complex & Intelligent Systems*, vol. 8, pp. 323 – 336, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235576377

[112] S. Kallel, M. Amayri, and N. Bouguila, “Clustering and interpretability of residential electricity demand profiles,” *Sensors (Basel, Switzerland)*, vol. 25, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:277293211

[113] A. Karim, S. Azam, B. Shanmugam, and K. Kannoorpatti, “Efficient clustering of emails into spam and ham: The foundational study of a comprehensive unsupervised framework,” *IEEE Access*, vol. 8, pp. 154 759–154 788, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221474601

[114] S. Zouinina, Y. Bennani, N. Rogovschi, and A. Lyhyaoui, “Data anonymization through collaborative multi-view microaggregation,” *Journal of Intelligent Systems*, vol. 30, pp. 327 – 345, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:222127209

[115] K. Pearson, “Liii. on lines and planes of closest fit to systems of points in space,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[116] C. Shen, S. Asante-Okyere, Y. Y. Ziggah, L. Wang, and X. Zhu, “Group method of data handling (gmdh) lithology identification based on wavelet analysis and dimensionality reduction as well log data pre-processing techniques,” *Energies*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:145966196

[117] A. V., S. I., R. V., and S. B., “Automatic fast video object detection and tracking on video surveillance system,” *ICTACT Journal on Image and Video Processing*, vol. 03, pp. 479–484, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:15743590

[118] Y. Peng, P. H. M. Albuquerque, I. F. do Nascimento, and J. V. F. Machado, “Between nonlinearities, complexity, and noises: An application on portfolio selection using kernel principal component analysis,” *Entropy*, vol. 21, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:132455926

[119] O. Uzga-Rebrovs and G. Kulesova, “Initial dataset dimension reduction using principal component analysis,” *Information Technology and Management Science*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:230558886

[120] D. Eslami, H. Izadbakhsh, O. Ahmadi, and M. Zarinbal, “Statistical modeling and monitoring of image data in the presence of temporal and spatial correlations,” *Scientia Iranica*, pp.–, 2022. [Online]. Available: https://scientiairanica.sharif.edu/article_22828.html

[121] R. Kambo and A. Yerpude, “Classification of basmati rice grain variety using image processing and principal component analysis,” *ArXiv*, vol. abs/1405.7626, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:13485027

[122] V. Kshirsagar, M. R. Baviskar, and M. E. Gaikwad, “Face recognition using eigenfaces,” *2011 3rd International Conference on Computer Research and Development*, vol. 2, pp. 302–306, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:8707245

[123] L. Xuan, L. Qian, J. Chen, X. □. Bai, and B. Wu, “State-of-charge prediction of battery management system based on principal component analysis and improved support vector machine for regression,” *IEEE Access*, vol. 8, pp. 164 693–164 704, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221847888

[124] T. Schneider, A. B. Bedrikow, and K. Stahl, “Enhanced prediction of thermomechanical systems using machine learning, pca, and finite element simulation,” *Adv. Model. Simul. Eng. Sci.*, vol. 11, p. 14, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270886304

[125] K. Y. Yeung and W. L. Ruzzo, “Principal component analysis for clustering gene expression data,” *Bioinformatics (Oxford, England)*, vol. 17, no. 9, pp. 763–774, 2001. [Online]. Available: https://doi.org/10.1093/bioinformatics/17.9.763

[126] S. Sadhukhan and V. K. Yadav, “Forecasting, capturing and activation of carbon-dioxide co2: Integration of time series analysis, machine learning, and material design,” *ArXiv*, vol. abs/2307.14374, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260203131

[127] M. Ghorbani and E. K. P. Chong, “Stock price prediction using principal components,” *PLoS ONE*, vol. 15, no. 3, p. e0230124, Mar 2020.

[128] X. Zhong and D. Enke, “Forecasting daily stock market return using dimensionality reduction,” *Expert Systems with Applications*, vol. 67, pp. 126–139, Jan 2017.

[129] H. Yu, R. Chen, and G. Zhang, “A svm stock selection model within pca,” in *Proceedings of Computational Science*, vol. 31, Jan 2014, pp. 406–412.

[130] G. Pasini, “Principal component analysis for stock portfolio management,” *International Journal of Pure and Applied Mathematics*, vol. 115, no. 1, pp. 153–167, Jun 2017.

[131] A. Nagpal, M. Sabharwal, and R. Tripathi, “A hybrid feature selection approach for urinary tract infection detection and prediction in iot-fog environment,” *Multidisciplinary Science Journal*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:267010692

[132] A. Omogbai, “Application of multiview techniques to nhanes dataset,” *ArXiv*, vol. abs/1608.04783, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:17445990

[133] S. Ezekiel, A. A. Alshehri, L. Pearlstein, X.-W. Wu, and A. Lutz, “Iot anomaly detection using multivariate,” *International Journal of Innovative Technology and Exploring Engineering*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:210164074

[134] F. Harvey, A. Iwaniak, S. Coetzee, and A. K. Cooper, “Sdi past, present and future: a review and status assessment,” in *Spatial Enabling Government, Industry and Citizens*. Enschede, The Netherlands: GSDI Association Press, 2012, pp. 23–46.

[135] A. E. Mulder, G. Wiersma, and B. Van Loenen, “Status of national open spatial data infrastructures: A comparison across continents,” *International Journal of Spatial Data Infrastructures Research*, vol. 15, pp. 56–87, 2020.

[136] A. Trystuła, M. Dudzińska, and R. Źróbek, “Evaluation of the completeness of spatial data infrastructure in the context of cadastral data sharing,” *Land*, vol. 9, no. 8, 2020. [Online]. Available: https://www.mdpi.com/2073-445X/9/8/272

[137] A. Negi, “A brief survey on text mining, its techniques, and applications,” *International Journal of Mobile Computing and Application*, vol. 8, no. 1, pp. 1–6, 2021.

[138] M. F. Samiyeva and M. A. Madyarova, “Text mining and its development stages,” *Science and Education*, vol. 4, no. 4, pp. 1346–1352, 2023.

[139] I. Kaczmarek, A. Iwaniak, A. Świetlicka, M. Piwowarczyk, and A. Nadolny, “A machine learning approach for integration of spatial development plans based on natural language processing,” *Sustainable Cities and Society*, vol. 76, p. 103479, 11 2021.

[140] L. M. Baptista and A. Figueiras, “Evaluating narrative in geoportals for territorial public policies,” *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 56, no. 4, pp. 303–319, 2021.

[141] M. A. Khder, “Web scraping or web crawling: State of art, techniques, approaches and application,” *International Journal of Advances in Soft Computing & Its Applications*, vol. 13, no. 3, 2021.

[142] S. Bickel, T. C. Spruegel, B. R. Schleich, and S. Wartzack, “How do digital engineering and included ai based assistance tools change the product development process and the involved engineers,” *Proceedings of the Design Society: International Conference on Engineering Design*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:201235043

[143] D. L. Cogburn, “Analyzing trends and topics in internet governance and cybersecurity debates found in twelve years of igf transcripts,” in *Hawaii International Conference on System Sciences*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:102352924

[144] IBM. (n.d.) Conceptos básicos de ayuda de crisp-dm. Accedido el 06 de octubre de 2025. [Online]. Available: https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

[145] IDEE, “Infraestructura de datos espaciales de españa,” https://www.idee.es/, accedido el 27 de septiembre de 2025.

[146] IDE Uruguay, “Visualizador de la infraestructura de datos espaciales de uruguay,” https://visualizador.ide.uy/, accedido el 27 de septiembre de 2025.

[147] ——, “Catálogo de metadatos de la ide de uruguay,” https://visualizador.ide.uy/geonetwork/srv/spa/catalog.search#/home, accedido el 27 de septiembre de 2025.

[148] IDEE, “Catálogo de metadatos de la ide de españa,” https://www.idee.es/csw-inspire-idee/srv/eng/catalog.search#/home, accedido el 27 de septiembre de 2025.

[149] H. Dang, B. Nguyen, N. Ziems, and M. Jiang, “Embedding mental health discourse for community recommendation,” *ArXiv*, vol. abs/2307.03892, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259376588

[150] H. Westermann, J. Savelka, V. R. Walker, K. D. Ashley, and K. Benyekhlef, “Sentence embeddings and high-speed similarity search for fast computer assisted annotation of legal documents,” *ArXiv*, vol. abs/2112.11494, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:229367900

[151] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410/

[152] H. Kroll, P. Sackhoff, B. M. Thang, M. Ksouri, and W.-T. Balke, “A library perspective on supervised text processing in digital libraries: An investigation in the biomedical domain,” *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:274149876

[153] A. Mahboub, M. E. Za’ter, B. Alfrou, Y. Estaitia, A. Jaljuli, and A. Hakouz, “Evaluation of semantic search and its role in retrieved-augmented-generation (rag) for arabic language,” *ArXiv*, vol. abs/2403.18350, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268723878

[154] J. Gaarder and P. Møller, *Sophie’s World: A Novel about the History of Philosophy*, ser. Berkley book. Berkley Books, 1996. [Online]. Available: https://books.google.com.uy/books?id=dZL_yErvfDkC

[155] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[156] D. Merkel, “Docker: lightweight linux containers for consistent development and deployment,” *Linux journal*, vol. 2014, no. 239, p. 2, 2014.

[157] G. van Rossum and the Python development team, *Python 3 Reference Manual*, Python Software Foundation, 2023. [Online]. Available: https://docs.python.org/3/reference/

[158] MongoDB Inc., “MongoDB: The application data platform,” https://www.mongodb.com, 2024, accessed: 2025-07-05.

[159] V. Bisht, R. Choyal, A. S. Negi, and E. K. Singh, “Utilizing python for web scraping and incremental data extraction,” pp. 1450–1455, Dec 2023. [Online]. Available: https://doi.org/10.1109/ICACRS58579.2023.10404702

[160] H. Poincaré, *Science and Method*. New York: Dover, 1902, “A fact in itself is nothing. It is valuable only for the idea attached to it, or for the proof which it furnishes.”.

[161] C. de Boor, *A Practical Guide to Splines*. Springer, 2001.

[162] Instituto Panamericano de Geografía e Historia, “Jornadas anuales 2025,” Instituto Geográfico Militar, Montevideo, Uruguay, Sep. 2025, accedido: 2025-10-12. [Online]. Available: https://ipgh.org.uy/pages/eventos/2025-1.html

[163] N. Bentancor, “Ide comparator,” GitHub repository, 2025, accedido: 13-Oct-2025. [Online]. Available: https://github.com/NoeliaBentancor/IDEComparator.git

A Anexos

A.1. Resultados análisis métricas

A.1.1. *Silhouette Score*

A.1.1.1. IDE Uruguay

País	Cantidad de datos	K óptimo	*Silhouette Score*
Uruguay	100	3	0.98362756
	1000	3	0.99836266
	2000	3(=)	0.9991813
	3000	3(=)	0.9994541
	4000	3(=)	0.9995906
	5000	3(=)	0.9996725
	6000	3(=)	0.9997271
	7000	4	0.9997641
	8000	4(=)	0.9997936
	9000	4(=)	0.99981654
	9391	4(=)	0.99982417

Tabla A.1: Evolución del valor Silhouette Score(sin redondeo) en función de la cantidad de datos en Uruguay.

: Elaboración propia.

A.1.1.2. IDE España

País	Cantidad de datos	k óptimo	*Silhouette Score*
España	100	11	0.875730574131012
	300	15	0.623317301273346
	500	11	0.5025307536125183
	700	15	0.6420274972915649
	900	15	0.7263484597206116
	1100	11	0.7735884189605713
	1300	11	0.8084207773208618
	1500	13	0.8037937879562378
	1700	14	0.7635549902915955
	1900	14	0.7124417424201965
	2100	15	0.7088155746459961
	2200	15	0.6925098896026611

Tabla A.2: Evolución de *Silhouette Score* y de k en función de la cantidad de datos en España

Fuente: Elaboración propia.

A.1.2. Davies Bouldin Index

A.1.2.1. IDE Uruguay

País	**Cantidad de datos**	**k óptimo**	**Valor Davies-Bouldin**
Uruguay	100	4	9.889649453454232e-07
	1000	4	9.999575887305112e-06
	2000	4	2.0640571815092386e-05
	3000	4	3.2770106944239685e-05
	4000	4	4.374076785929748e-05
	5000	4	5.228216181603007e-05
	6000	4	6.100691707440391e-05
	7000	5	0.00012771787256878533
	8000	5	0.0001398902961261304
	9000	5	0.00015126498295262706
	9391	5	0.00015482279922795082

Tabla A.3: Evolución del índice Davies Bouldin y de k en función de la cantidad de datos en Uruguay

Fuente: Elaboración propia.

A.1.2.2. IDE España

País	Cantidad de datos	k óptimo	Valor Davies-Bouldin
España	100	15	0.3494871777341576
	300	2	1.216123697436212
	500	2	1.537612610815216
	700	3	1.159401678084403
	900	3	1.15864266446037
	1100	3	1.158349179507332
	1300	3	1.158193544046657
	1500	3	1.1532338997103617
	1700	3	1.1038212026917515
	1900	2	1.2025513317987628
	2100	3	1.0779498697131882
	2200	3	1.101317738595917

Tabla A.4: Evolución del índice Davies Bouldin y de k en función de la cantidad de datos en España.

Fuente: Elaboración propia.

A.1.3. Calinski-Harabasz

A.1.3.1. IDE Uruguay

País	Cantidad de documentos	Número de clusters	Calinski-Harabasz
Uruguay	100	4	2143513000000.0
	1000	4	20434747000.0
	2000	4	4788594700.0
	3000	4	1898689900.0
	4000	4	1065470340.0
	5000	4	745703100.0
	6000	4	547626560.0
	7000	5	2475080700.0
	8000	5	7772312000.0
	9000	5	12695791000.0
	9391	5	14452959000.0

Tabla A.5: Resultados del índice de Calinski-Harabasz para Uruguay

A.1.3.2. IDE España

País	Cantidad de documentos	Número de clusters	Calinski-Harabasz
España	100	15	439.4090270996094
	300	2	119.99584197998047
	500	2	137.54347229003906
	700	2	236.60400390625
	900	2	353.6512451171875
	1100	2	475.0137023925781
	1300	2	598.07861328125
	1500	2	721.0642700195312
	1700	2	862.7117919921875
	1900	2	999.5468139648438
	2100	2	995.9691772460938
	2200	2	978.6054077148438

Tabla A.6: Resultados del índice de Calinski-Harabasz para España, con distintos límites de documentos analizados

Fuente: Elaboración propia.

A.1.4. *Elbow Method* (inercia)

A.1.4.1. IDE Uruguay

País	Cantidad de documentos	Número de clusters	Inercia
Uruguay	100	4	6.30171370814836e-13
	1000	4	3.152061131637268e-13
	2000	4	4.079461107028143e-14
	3000	4	3.8514425481327164e-14
	4000	4	2.436512126227034e-14
	5000	2	2.0221541455200942e-13
	6000	4	5.496479465148807e-14
	7000	5	9.99074489804741e-10
	8000	5	5.206964193149588e-09
	9000	5	5.581187956238409e-09
	9391	5	1.4416745131029529e-08

Tabla A.7: Valores de inercia para Uruguay en función de la cantidad de documentos,

Número de clusters k	Inercia $I_{\text{Uruguay}}(k)$
2	29.25926399230957
3	11.841351509094238
4	3.577669620513916
5	0.00000001444410191453471 3
6	0.00000001446082986689134 5
7	0.00000001443404951118054 7
8	0.00000001446437813967804 7
9	0.00000001447286468447828 2
10	0.00000001439895402910451 6
11	0.00000001445951625100860 8
12	0.00000001448034225859373 7
13	0.00000001448817865679075 1
14	0.00000001453537556983519 6
15	0.00000001452883147123884 6

Tabla A.8: Inercia para Uruguay

Fuente: Elaboración propia.

A.1.4.2. IDE España

País	Límite de datos	K óptimo	Inercia (WCSS)
España	100	10	2.484591007232666
	300	6	564.2022705078125
	500	7	1229.50146484375
	700	7	1224.37158203125
	900	7	1225.6334228515625
	1100	6	1219.321044921875
	1300	7	1219.32177734375
	1500	6	1470.033935546875
	1700	8	1910.950927734375
	1900	7	2518.796142578125
	2100	6	2774.57666015625
	2200	6	3158.06787109375

Tabla A.9: Evolución de la inercia (*Elbow Method*) en función de la cantidad de documentos en España.

Fuente: Elaboración propia.

Número de clusters k	Inercia $I_{\text{España}}(k)$
2	6438.748046875
3	5490.5751953125
4	4867.09619140625
5	4434.484375
6	4160.37890625
7	3976.66015625
8	3741.52294921875
9	3606.37060546875
10	3498.67333984375
11	3395.831787109375
12	3304.90185546875
13	3228.02099609375
14	3177.53076171875
15	3158.06787109375

Tabla A.10: Inercia completa para España considerando 2200 documentos.

Fuente: Elaboración propia.

A.1.5. Cálculo derivada de la inercia

A.1.5.1. IDE Uruguay

```
import numpy as np
from scipy.interpolate import CubicSpline
import matplotlib.pyplot as plt

k = np.arange(2, 16)
I = np.array([29.25926399230957, 11.841351509094238,
   3.577669620513916,
              1.4444101914534713e-08, 1.4460829866891345e-08,
              1.4434049511180547e-08, 1.4464378139678047e-08,
                 1.4472864684478282e-08,
              1.4398954029104516e-08, 1.4459516251008608e-08,
                 1.4480342258593737e-08,
              1.4488178656790751e-08, 1.4535375569835196e-08,
                 1.4528831471238846e-08])
```

```

spline = CubicSpline(k, I)

k_cont = np.linspace(2, 15, 300)
dI_cont = spline.derivative()(k_cont)
plt.axhline(0, color='gray', linestyle=':', lw=1)

plt.figure(figsize=(8,5))
plt.plot(k_cont, dI_cont, 'orange', lw=2, label="Derivada I'(k)")
plt.axhline(0, color='gray', linestyle=':', lw=1)
plt.xlabel('Número de clusters k')
plt.ylabel("Derivada de la inercia I'(k)")
plt.title("Derivada de la inercia para Uruguay (9391 documentos)")
plt.legend()
plt.grid(alpha=0.3)
plt.show()
```

A.1.5.2. IDE España

```
import numpy as np
from scipy.interpolate import CubicSpline
import matplotlib.pyplot as plt

# Datos de inercia para España
k = np.arange(2, 16)
I = np.array([6438.748046875, 5490.5751953125, 4867.09619140625,
   4434.484375,
              4160.37890625, 3976.66015625, 3741.52294921875,
                 3606.37060546875,
              3498.67333984375, 3395.831787109375, 3304.90185546875,
                 3228.02099609375,
              3177.53076171875, 3158.06787109375])
```

```
11
12 # Spline cúbica
13 spline = CubicSpline(k, I)
14
15 # Malla continua
16 k_cont = np.linspace(2, 15, 300)
17 I_cont = spline(k_cont)
18 dI_cont = spline.derivative()(k_cont)
19
20 # Graficar
21 plt.figure(figsize=(8,5))
22 plt.plot(k_cont, I_cont, 'b-', lw=2, label='Inercia I(k)')
23 plt.plot(k_cont, dI_cont, 'orange', lw=2, linestyle='--', label="
    Derivada I'(k)")
24 plt.axhline(0, color='gray', linestyle=':', lw=1)
25 plt.scatter(k_cont[np.argmin(np.abs(dI_cont))],
26             I_cont[np.argmin(np.abs(dI_cont))],
27             color='red', label='Aprox. punto elbow')
28 plt.xlabel('Número de clusters k')
29 plt.ylabel('Inercia / Derivada')
30 plt.title('Curva de inercia y derivada para España (2200 documentos)'
    )
31 plt.legend()
32 plt.grid(alpha=0.3)
33 plt.show()
```

A.2. Repositorio del proyecto

El desarrollo completo de la herramienta se encuentra disponible en el repositorio público de GitHub:

- **URL principal**: `https://github.com/NoeliaBentancor/IDEComparator`[163]

- **Licencia**: MIT
- **Estructura técnica**:

```
IDEComparator/
    app.py
    config.py
    requirements.txt
    .env.example
    README.md
    LICENSE
    docker-compose.yml
    entities/
        enums/
            __init__.py
            country.py
            metric.py
        __init__.py
        document.py
        pdf_analysis_report.py
        pdf_base_report.py
        pdf_clustering_report.py
        scraper.py
    sections/
        __init__.py
        analysis.py
        clustering.py
        comparison_metrics.py
        database.py
        scraper.py
    utils/
```

```
        __init__.py
        clustering_utils.py
        database.py
        embedding_utils.py
        pdf_generator.py
        silhouete_analysis.py
        string_utils.py
        text_utils.py
        worldcloud_utils.py
    manual/
        manual.pdf
```

A.3. Manuales

En esta sección se detalla el manual de instalación (ver A.3.1) y el manual orientado al usuario (ver A.3.2).

A.3.1. Manual de instalación

A.3.1.1. Prerequisitos

- Python 3.10 o superior
- pip (gestor de paquetes de Python)
- Git
- Docker (Se recomienda utilizar 20.10 o superior.)
- Docker Compose: Se requiere Docker Compose versión 1.29 o superior.

A.3.1.2. Clonación del repositorio

```
git clone IDEComparator [163]
```

A.3.1.3. Ejecutar docker

```
docker compose up
```

A.3.1.4. Agregar archivo .env

```
# Spain IDE configuration
BASE_URL_SPAIN=https://www.idee.es/csw-inspire-idee/srv/eng/q
DB_COLLECTION_SPAIN=land_cover_metadata_spain

# Uruguay IDE configuration
BASE_URL_URUGUAY=https://visualizador.ide.uy/geonetwork/srv/spa/q
DB_COLLECTION_URUGUAY=land_cover_metadata_uruguay

# Database configuration
DB_NAME=geospatial_data
DB_HOST=localhost
DB_PORT=27017
DB_USERNAME=admin
DB_PASSWORD=password
```

A.3.1.5. Instalación de dependencias

El archivo “requirements.txt” contiene las siguientes bibliotecas de Python necesarias como muestra la Figura A.1:

```
streamlit
python-dotenv
PyPDF2
pillow
pymongo
requests
langdetect
sentence-transformers
wordcloud
matplotlib
pandas
nltk
scikit-learn
kneed
reportlab
autopep8
seaborn
streamlit-aggrid
```

Figura A.1: Dependencias-IDE Comparator.

A.3.1.6. Ejecutar la aplicación

```
streamlit run app.py
```

Una vez ejecutado este comando, se deberá visualizar la pantalla de inicio (ver A.3.2.1).

A.3.2. Manual de usuario

El objetivo del manual de usuario es proporcionar instrucciones claras sobre el uso del prototipo. Se muestra el inicio (ver A.3.2.1), el módulo de metadatos (ver A.3.2.2), scraper (ver A.3.2.3), análisis descriptivo (ver A.3.2.4), análisis de métricas (ver A.3.2.5), análisis de *clustering* (ver A.3.2.6) y el módulo de ayuda y documentación (ver A.3.2.7).

A.3.2.1. Inicio

En la Figura A.2 se ilustra la pantalla principal de IDE Comparator.

Figura A.2: Pantalla principal-IDE Comparator

Fuente: Elaboración propia.

En esta pantalla, se detallan las funcionalidades principales de la aplicación.

A.3.2.2. Metadatos

El módulo de metadatos permite visualizar y descargar los registros almacenados en la base de datos. Esto permite que los datos puedan ser obtenidos sin necesidad de conectarse directamente a la base de datos, facilitando al usuario distintos tipos de análisis, incluso si únicamente desea utilizar el *scraper* o trabajar con los datos descargados.

Este módulo permite tanto la visualización (ver A.3.2.2) como la descarga (ver A.3.2.2) de metadatos.

Visualización de metadatos

En la Figura A.3 se muestra la visualización de los metadatos en formato de tabla.

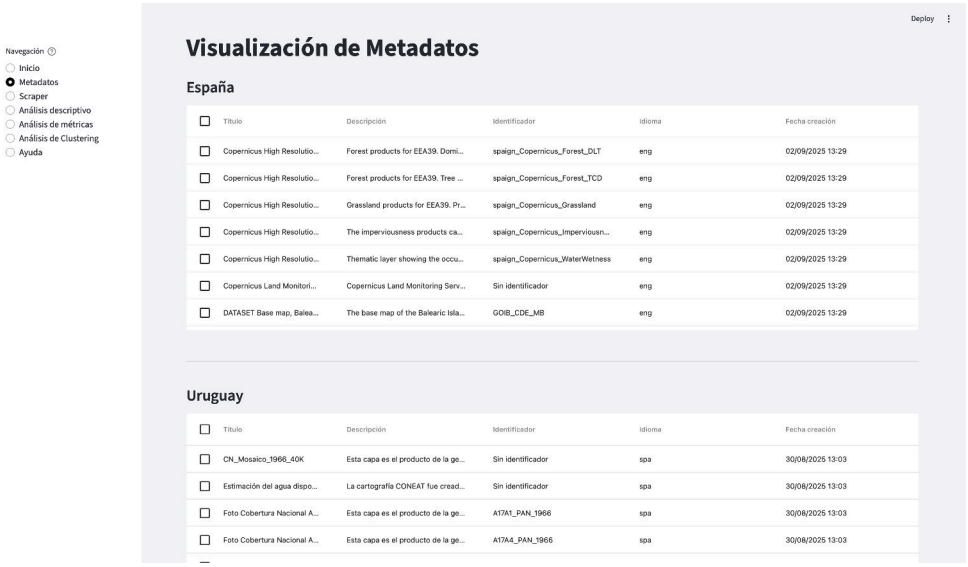
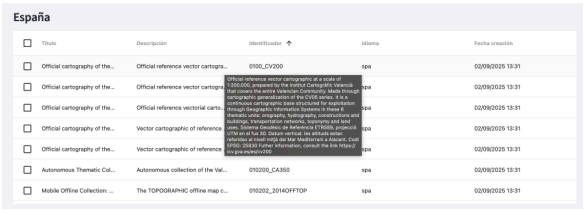


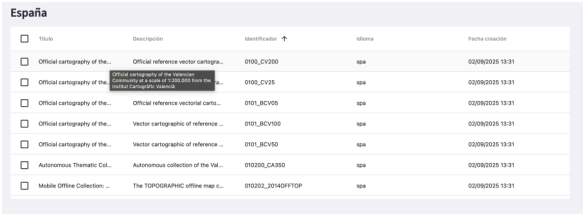
Figura A.3: Pantalla de visualización de metadatos - IDE Comparator

Fuente: Elaboración propia.

Dado que los títulos y descripciones suelen ser extensos, la tabla incorpora *tooltips* que permiten visualizar el contenido completo, como se ilustra en la Figura A.4.



(a) *Tooltip* de la descripción en visualización de metadatos - IDE Comparator



(b) *Tooltip* del título en visualización de metadatos - IDE Comparator

Figura A.4: Visualización de *tooltips* en la tabla de metadatos - IDE Comparator.

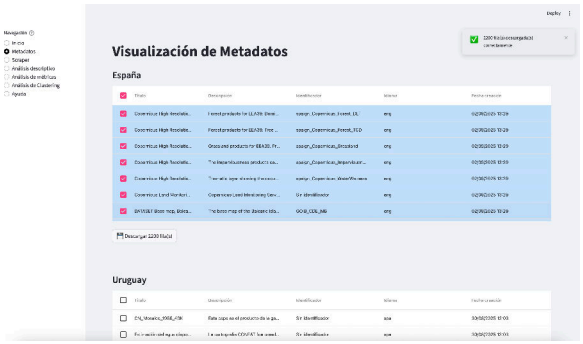
Fuente: Elaboración propia.

Descarga de metadatos

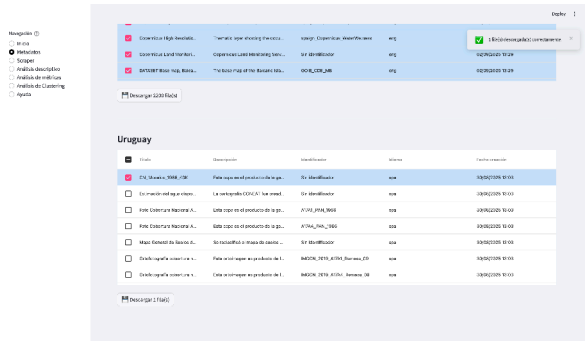
La aplicación permite descargar los metadatos, ofreciendo dos opciones:

1. **Descargar todos los registros**
2. **Descargar registros seleccionados**

Una vez seleccionados los registros, la aplicación proporciona *feedback* indicando que la descarga se ha completado correctamente y mostrando la cantidad de filas descargadas (ver Figura A.5).



(a) Descargar todos los metadatos

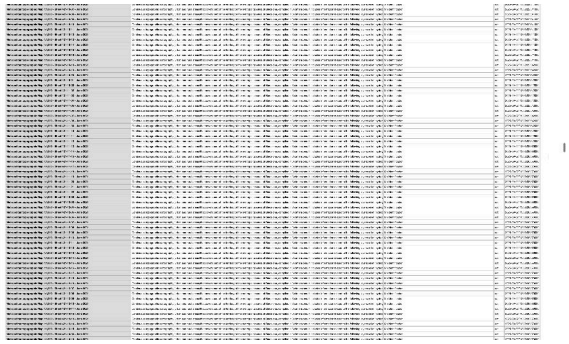


(b) Descargar un metadato

Figura A.5: Acciones de descarga de metadatos - IDE Comparator.

Fuente: Elaboración propia.

Los datos se descargan en formato CSV, permitiendo su análisis posterior de manera flexible. Ejemplos de los datos descargados para España y Uruguay se muestran en la Figura A.6.



(a) Datos descargados de España

(b) Datos descargados de Uruguay

Figura A.6: Datos descargados-IDE Comparator

Fuente: Elaboración propia.

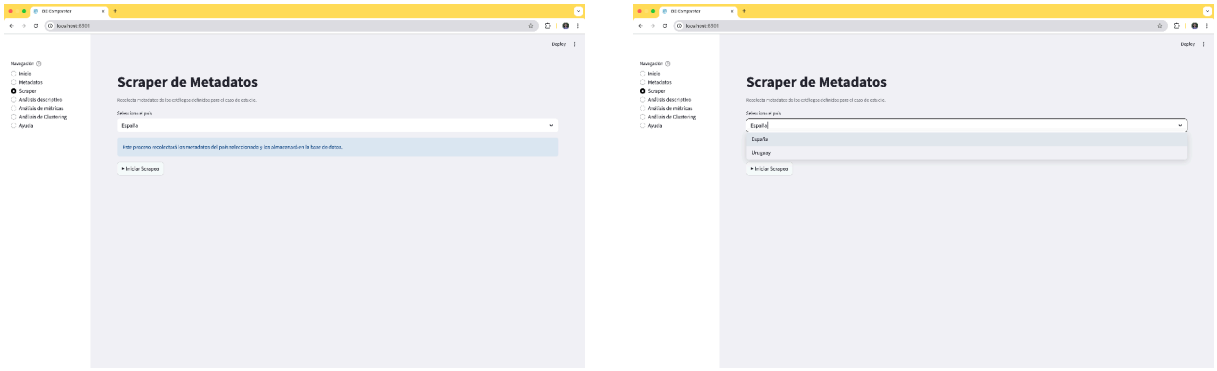
A.3.2.3. Scraper

El módulo Scraper permite recolectar automáticamente los metadatos, centralizando los resultados en la aplicación. Esto asegura que los datos se mantengan actualizados y facilita el

análisis. Su principal valor radica en la automatización del proceso, garantizando que los resultados estén siempre alineados con los datos públicos más recientes.

Pantalla principal del Scraper

Al ingresar al módulo, se despliega la pantalla de Scraper como se muestra en la Figura A.7a.



(a) Pantalla Scraper (b) Selección de país en pantalla Scraper

Figura A.7: Pantalla del módulo Scraper - IDE Comparator.

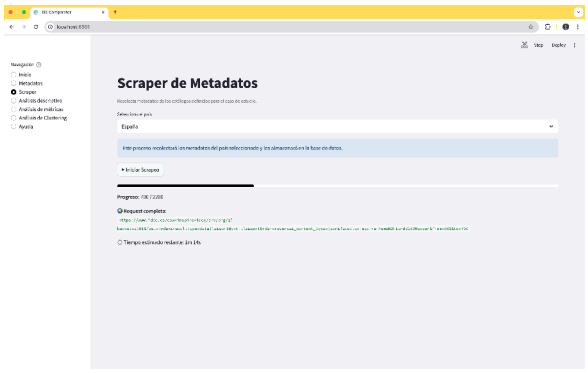
Fuente: Elaboración propia.

Para iniciar el proceso:

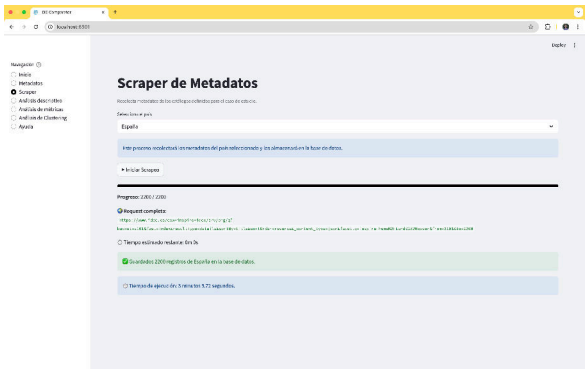
1. Seleccione uno de los países disponibles (ver Figura A.7b).
2. Presione Iniciar Scrapeo.

Feedback del proceso de Scraping

Durante la ejecución del Scraper, la aplicación proporciona *feedback* en tiempo real para informar sobre el progreso del proceso como se ilustra en la Figura A.8.



(a) *Feedback* de progreso intermedio



(b) *Feedback* de finalización de scraping

Figura A.8: Mensajes de *feedback* del módulo Scraper.

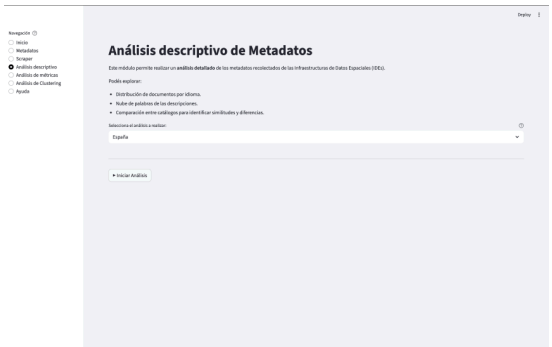
Fuente: Elaboración propia.

Almacenamiento de los datos

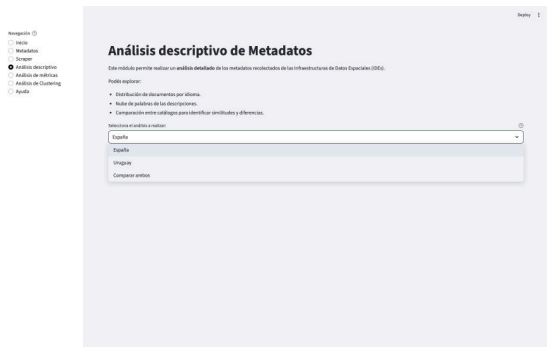
Una vez finalizado el *scraping*, los datos se almacenan automáticamente en la base de datos, permitiendo su posterior análisis y visualización dentro de la aplicación.

A.3.2.4. Análisis descriptivo

El módulo Análisis ilustrado en la Figura A.9a permite realizar un análisis individual (ver A.3.2.4) o comparar ambas IDEs (ver A.3.2.4).



(a) Pantalla módulo de análisis - IDE Comparator



(b) Pantalla de selección de país - IDE Comparator

Figura A.9: Pantallas del módulo de análisis descriptivo - IDE Comparator.

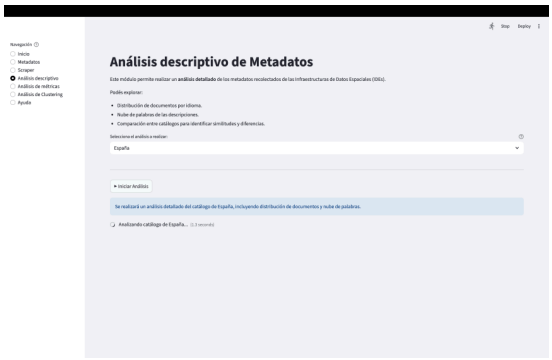
Fuente: Elaboración propia.

Análisis individual

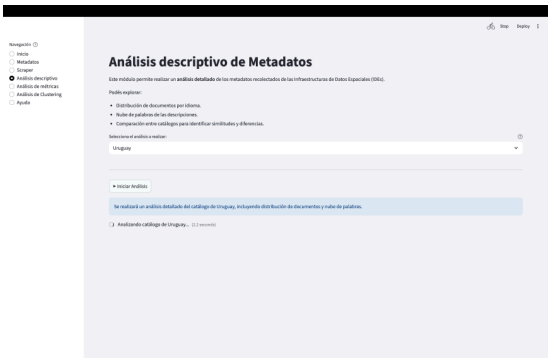
Para realizar un análisis individual:

1. Seleccionar el país deseado (ver Figura A.15a).
2. Presionar Iniciar Análisis.

Durante el proceso, la aplicación muestra un *feedback* indicando que el análisis está en curso (ver Figura A.10b).



(a) Selección país España - IDE Comparator



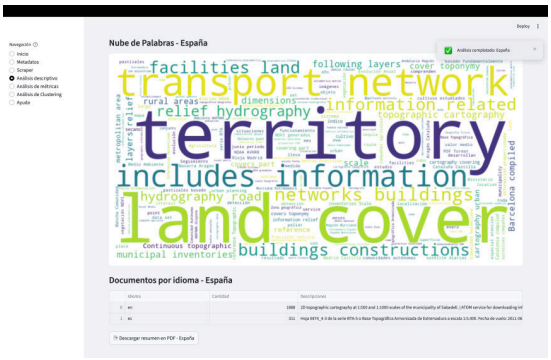
(b) Selección país Uruguay - IDE Comparator

Figura A.10: Feedback de análisis descriptivo individual - IDE Comparator.

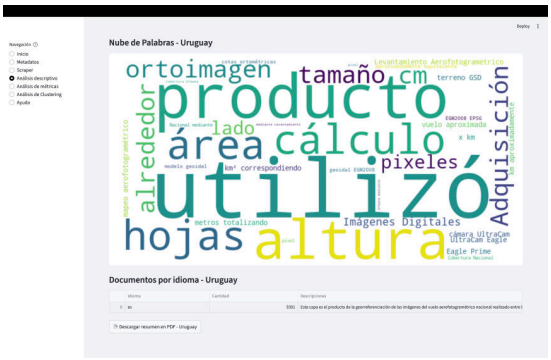
Fuente: Elaboración propia.

Al finalizar el análisis, se muestra la pantalla de resultados (ver Figura A.11) que permite:

- Descargar el informe en PDF.
- Visualizar la nube de palabras por país y documento.
- Consultar documentos por idioma.



(a) Resultados país España

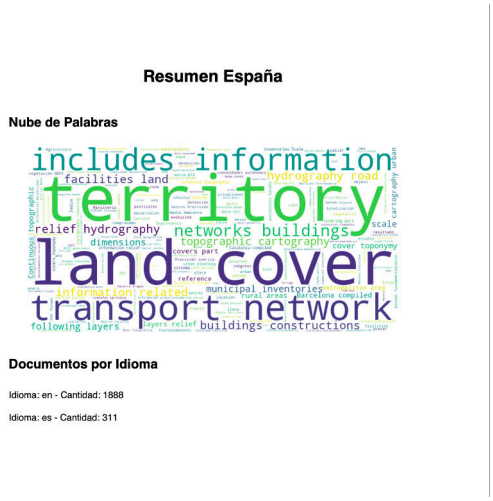


(b) Resultados país Uruguay

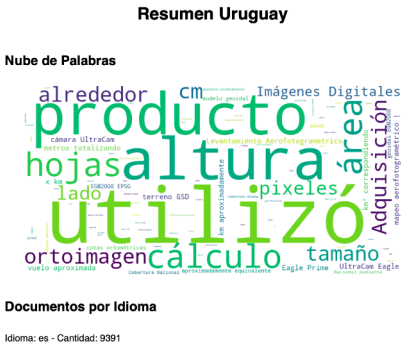
Figura A.11: Resultados de análisis descriptivo individual - IDE Comparator.

Fuente: Elaboración propia.

El PDF final contiene los mismos elementos: nube de palabras y documentos por idioma, proporcionando un marco de análisis exploratorio completo (ver Figura A.9).



(a) PDF resultados España



(b) PDF resultados Uruguay

Figura A.12: Resultados de análisis descriptivo individual en PDF - IDE Comparator.

Fuente: Elaboración propia.

Análisis comparativo

Para realizar un análisis comparativo:

- Seleccionar Comparar ambos países (ver Figura A.13).

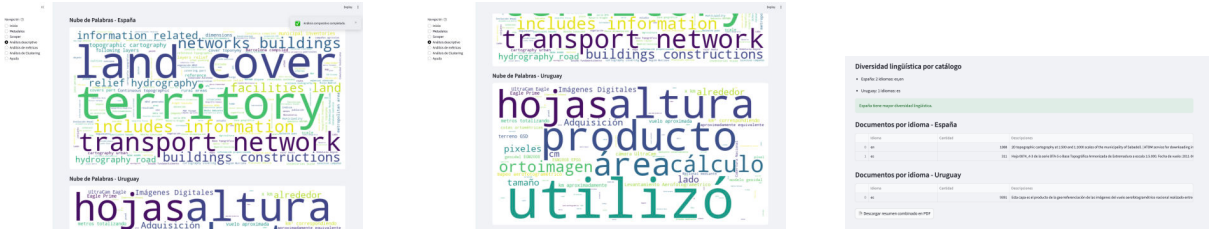
La aplicación mostrará *feedback* de proceso, indicando que la comparación está en curso (ver Figura A.13).

Figura A.13: Feedback durante análisis comparativo - IDE Comparator

Fuente: Elaboración propia.

El resultado del análisis comparativo ilustrado en la Figura A.14 incluye:

- Nube de palabras por país.
- Diversidad lingüística por país, indicando cuál de las dos IDEs posee mayor diversidad en términos lingüísticos.



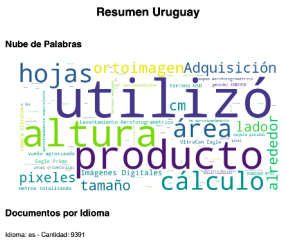
(a) IDE Comparator España (b) IDE Comparator Uruguay (c) Resultados de comparación

Figura A.14: Resultados del análisis comparativo - IDE Comparator.

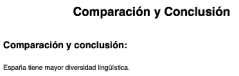
Fuente: Elaboración propia.

El PDF comparativo refleja los mismos elementos: documentos por idioma, nube de palabras, diversidad lingüística y cuál de las IDEs tiene mayor diversidad en términos lingüísticos (ver Figura A.15).

(a) PDF resultados España



(b) PDF resultados Uruguay



(c) Resultados de comparación en PDF

Figura A.15: Resumen en PDF del análisis comparativo - IDE Comparator.

Fuente: Elaboración propia.

A.3.2.5. Análisis de métricas

El módulo Análisis de Métricas ilustrado en la Figura A.16 permite explorar los datos de cada país de forma individual o realizar una comparación entre ambos.

Figura A.16: Pantalla principal del módulo de análisis de métricas.

Parámetros de configuración

Antes de iniciar el análisis, se deben definir los siguientes parámetros (ilustrados en la Figura A.16):

- **País:** País sobre el cual se realizará el análisis.
- **Límite de documentos:** Se puede trabajar con un subconjunto de documentos o con todos.

- ***k* a evaluar:** Número de *clusters* a analizar, indicando *k* mínimo y *k* máximo.
- ***Batch size*:** Tamaño del *batch* para procesar los documentos.
- **Métricas:** Métricas a evaluar (*Elbow Method*, *Silhouette Score*, Davies-Bouldin, Calinski-Harabasz).

Flujo de análisis individual:

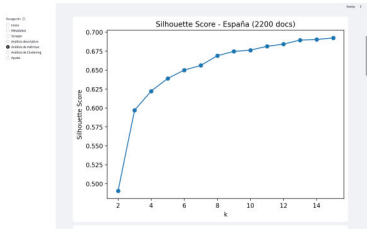
1. Configurar los parámetros.
2. Presionar **Iniciar Análisis**.
3. Visualizar el *feedback* de procesamiento (ver Figura A.17).

Figura A.17: Feedback de procesamiento durante análisis individual.

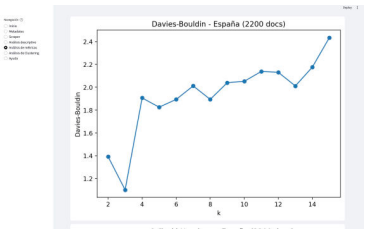
El resultado del análisis de métricas individual (ver Figura A.18) incluye:

- **Gráficas:** se muestran gráficas de las métricas seleccionadas.
- **Tabla resumen:** Incluye el país, cantidad de documentos, *k* óptimo, valor de la métrica y, para *Silhouette Score*, su interpretación.

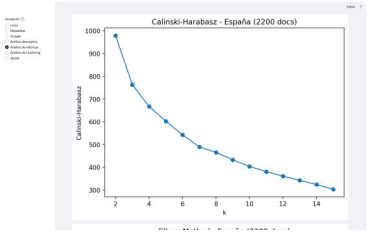
(a) Resultado análisis individual.



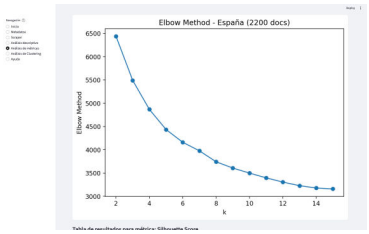
(b) Resultado análisis individual.



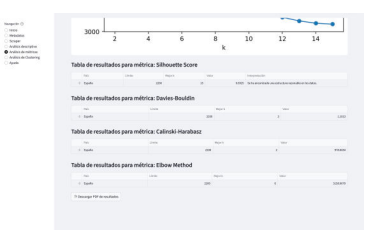
(c) Resultado análisis individual.



(d) Resultado análisis individual.



(e) Resultado análisis individual.

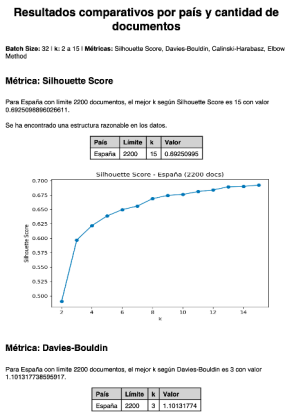


(f) Resultado análisis individual

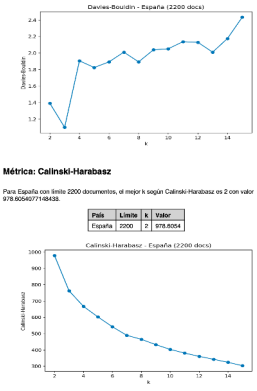
Figura A.18: Resultados del análisis individual de análisis de métricas-IDE Comparator.

Fuente: Elaboración propia.

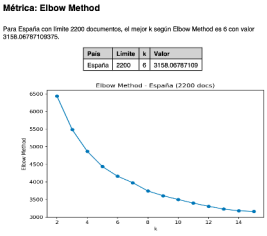
La Figura A.19 ilustra el resultado del resumen en PDF del análisis a nivel individual.



(a) Resumen PDF individual.



(b) Resumen PDF individual.



(c) Resumen PDF individual.

Figura A.19: Resumen en PDF del análisis de métricas-IDE Comparator

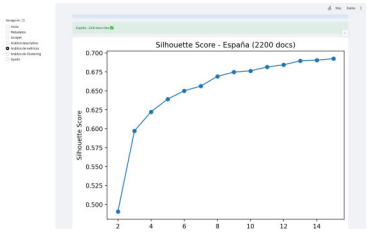
Fuente: Elaboración propia.

Flujo de análisis comparativo:

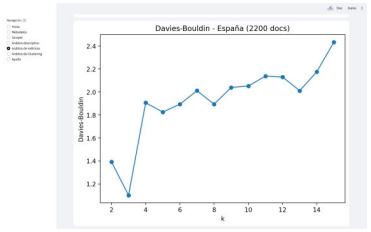
1. Configurar los parámetros para ambos países.
2. Presionar **Iniciar Análisis Comparativo**.
3. La aplicación muestra *feedback* mientras procesa los documentos.

Por otro lado, el resultado del análisis comparativo (ver Figura A.20) incluye:

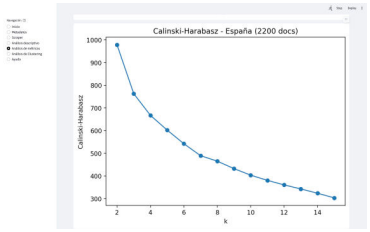
- **Gráficas:** se muestran gráficas de las métricas seleccionadas para ambos países.
- **Tabla resumen:** Incluye el país, cantidad de documentos, *k* óptimo, valor de la métrica y, para *Silhouette Score*, su interpretación.



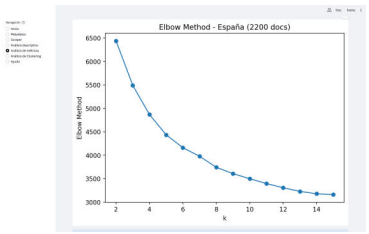
(a) Resultados del análisis comparativo de métricas.



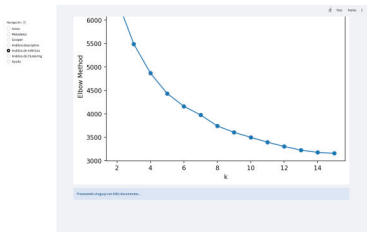
(b) Resultados del análisis comparativo de métricas.



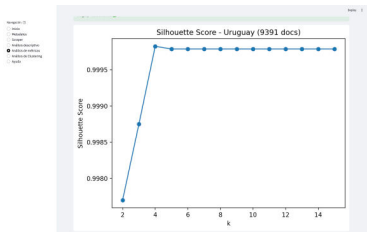
(c) Resultados del análisis comparativo de métricas.



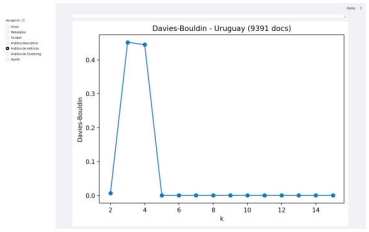
(d) Resultados del análisis comparativo de métricas.



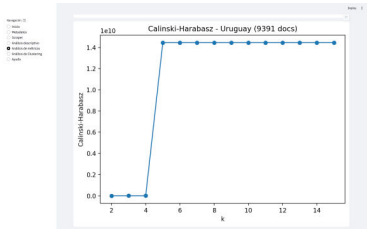
(e) Resultados del análisis comparativo de métricas.



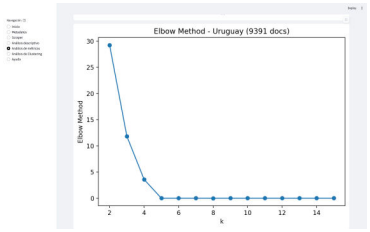
(f) Resultados del análisis comparativo de métricas.



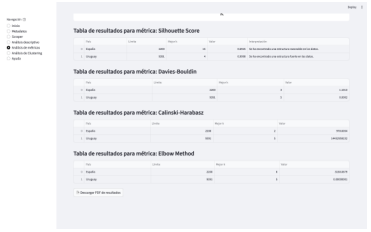
(g) Resultados del análisis comparativo de métricas.



(h) Resultados del análisis comparativo de métricas.



(i) Resultados del análisis comparativo de métricas.

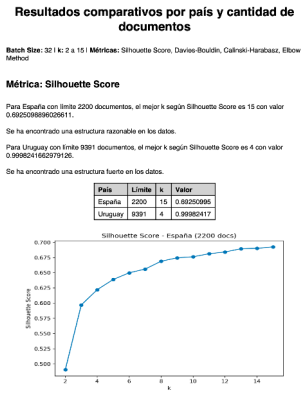


(j) Resultados del análisis comparativo de métricas.

Figura A.20: Resultados del análisis comparativo de métricas-IDE Comparator.

Fuente: Elaboración propia.

Esta información permite ser guardada en PDF (ver Figura A.21).



Resultados comparativos por país y cantidad de documentos

Batch Size: 32 | **k:** 2 a 15 | **Métricas:** Silhouette Score, Davies-Bouldin, Calinski-Harabasz, Elbow Method

Métrica: Silhouette Score

Para España con límite 2200 documentos, el mejor k según Silhouette Score es 15 con valor 0.692509889602611.

Se ha encontrado una estructura razonable en los datos.

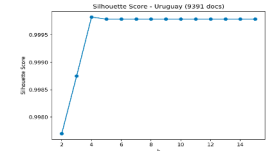
Para Uruguay con límite 9391 documentos, el mejor k según Silhouette Score es 4 con valor 0.9998241662979126.

Se ha encontrado una estructura fuerte en los datos.

País	Límite	k	Valor
España	2200	15	0.69250995
Uruguay	9391	4	0.99982417

Silhouette Score - España (2200 docs)

(a) Resumen en PDF del análisis comparativo de métricas.



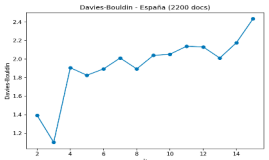
Métrica: Davies-Bouldin

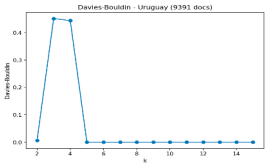
Para España con límite 2200 documentos, el mejor k según Davies-Bouldin es 3 con valor 1.1013177383698917.

Para Uruguay con límite 9391 documentos, el mejor k según Davies-Bouldin es 5 con valor 0.0001548227992795082.

País	Límite	k	Valor
España	2200	3	1.10131774
Uruguay	9391	5	0.00015482

(b) Resumen en PDF del análisis comparativo de métricas.





Métrica: Calinski-Harabasz

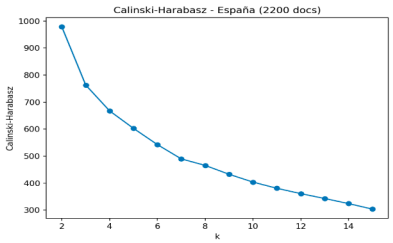
Para España con límite 2200 documentos, el mejor k según Calinski-Harabasz es 2 con valor 978.6054077148438.

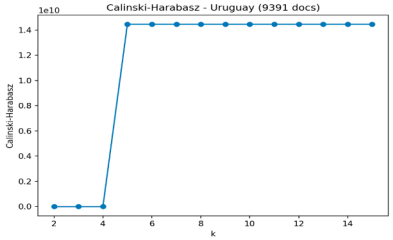
Para Uruguay con límite 9391 documentos, el mejor k según Calinski-Harabasz es 5 con valor 14452959232.0.

País	Límite	k	Valor
España	2200	2	978.6054

(c) Resumen en PDF del análisis comparativo de métricas.

Uruguay	9391	5	14452959000.0





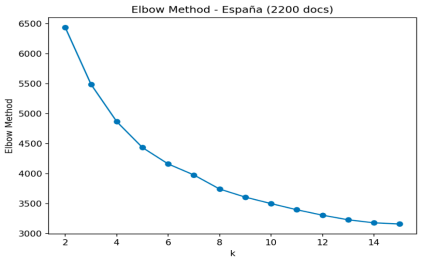
Métrica: Elbow Method

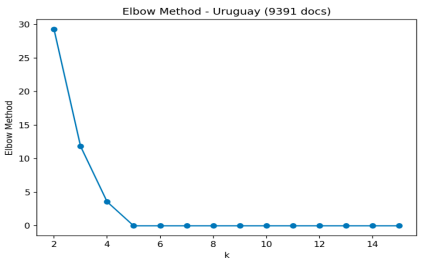
Para España con límite 2200 documentos, el mejor k según Elbow Method es 6 con valor 3158.06787109375.

Para Uruguay con límite 9391 documentos, el mejor k según Elbow Method es 5 con valor 1.4413641835631097e-08.

(d) Resumen en PDF del análisis comparativo de métricas.

País	Límite	k	Valor
España	2200	6	3158.06787109
Uruguay	9391	5	1e-08





(e) Resumen en PDF del análisis comparativo de métricas.

Figura A.21: Resumen en PDF del análisis comparativo de métricas-IDE Comparator.

Fuente: Elaboración propia.

A.3.2.6. Análisis de *clustering*

En la Figura A.22 se muestra la pantalla principal del módulo de análisis de *clustering*.

Parámetros del análisis:

- **País:** País a analizar.
- **Cantidad de documentos:** se puede elegir un subconjunto o todos los documentos disponibles.
- **Métrica y parámetros del modelo:** Métrica para calcular el k óptimo y *batch size* del modelo.

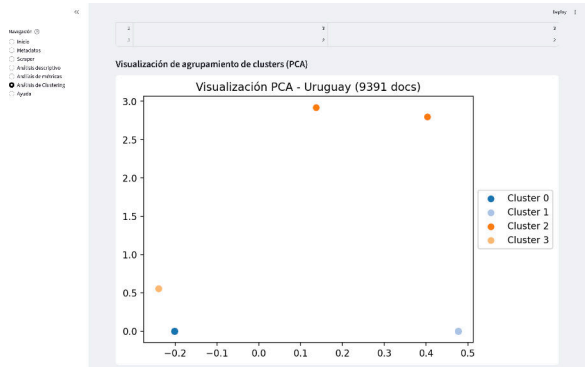
Figura A.22: Pantalla principal del módulo de análisis de clustering.

Fuente: Elaboración propia.

Luego de configurar los parámetros, se presiona **Generar Clusters**, y la aplicación muestra(ver A.23):

- ***k* óptimo:** Se informa con qué *k* se realizó la *clusterización*.
- **Documentos por *cluster*:** Se informa la distribución de cantidad de documentos por *cluster* con el fin de generar un entendimiento de la distribución de los *clusters*.
- **PCA:** Se permite visualizar los *clusters* en un gráfica que reduce las dimensiones de los datos (PCA) para un mayor entendimiento de la agrupación de la información.

(a) Resultados de análisis de *clustering*.



(b) Resultados de análisis de *clustering*.

Figura A.23: Resultados del análisis de *clustering*, mostrando cantidad de documentos por *cluster* y visualización en PCA.

Fuente: Elaboración propia.

A.3.2.7. Ayuda y documentación

IDE Comparator incluye una sección de ayuda y documentación, como se muestra en la Figura A.24, disponible para los usuarios que la requieran dentro de la aplicación.

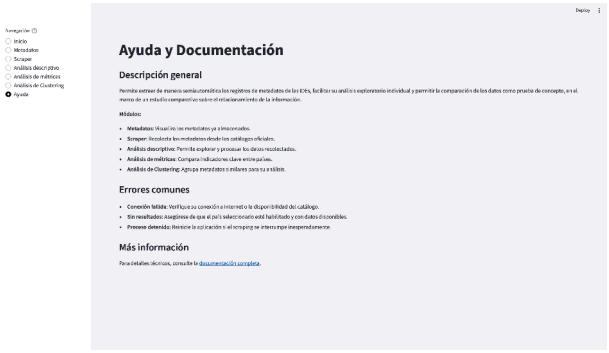


Figura A.24: Pantalla ayuda y documentación-IDE Comparator

Fuente: Elaboración propia.