

# ÍNDICE

|  |             |
|--|-------------|
| <b>PRÓLOGO</b> .....                         | <b>XIII</b> |
| <b>CAPÍTULO 1. LECTURA DE FICHEROS</b> ..... | <b>1</b>    |
| Introducción.....                            | 1           |
| CSV .....                                    | 2           |
| TSV .....                                    | 7           |
| Excel .....                                  | 8           |
| JSON .....                                   | 15          |
| XML.....                                     | 19          |
| Conclusiones .....                           | 24          |
| Referencias .....                            | 24          |
| <b>CAPÍTULO 2. WEB SCRAPING</b> .....        | <b>25</b>   |
| Introducción.....                            | 25          |
| Ficheros incluidos en la página web.....     | 27          |

|   |           |
|---|-----------|
| URIs, URLs y URNs .....                           | 27        |
| Ejemplo: datos de contaminación en Madrid .....   | 28        |
| Datos que forman parte de la página.....          | 32        |
| Lo que oculta una página web .....                | 32        |
| Un poco de HTML.....                              | 34        |
| Navegación absoluta .....                         | 38        |
| Navegación relativa .....                         | 40        |
| Ejemplo: día y hora oficiales.....                | 41        |
| Datos que requieren interacción.....              | 43        |
| Selenium: instalación y carga de páginas .....    | 44        |
| Clic en un enlace .....                           | 46        |
| Cómo escribir texto .....                         | 47        |
| Pulsando botones.....                             | 48        |
| Localizar elementos.....                          | 50        |
| XPath .....                                       | 51        |
| Navegadores <i>headless</i> .....                 | 58        |
| Conclusiones .....                                | 58        |
| Referencias.....                                  | 59        |
| <b>CAPÍTULO 3. RECOLECCIÓN MEDIANTE APIS.....</b> | <b>61</b> |
| Introducción.....                                 | 61        |
| API Twitter .....                                 | 62        |
| Acceso a Twitter como desarrollador.....          | 62        |

|   |           |
|---|-----------|
| Estructura de un tweet .....                        | 65        |
| Descargando tweets.....                             | 69        |
| API-REST .....                                      | 72        |
| Ejemplo: API de Google Maps .....                   | 72        |
| Ejemplo: API de OMDb.....                           | 73        |
| Referencias .....                                   | 75        |
| <b>CAPÍTULO 4. MONGODB.....</b>                     | <b>77</b> |
| Introducción.....                                   | 77        |
| ¿De verdad necesito una base de datos? ¿Cuál? ..... | 78        |
| Consultas complejas.....                            | 79        |
| Esquema de datos complejo o cambiante .....         | 80        |
| Gran volumen de datos.....                          | 81        |
| Arquitectura cliente-servidor de MongoDB .....      | 81        |
| Acceso al servidor .....                            | 81        |
| Puesta en marcha del servidor.....                  | 82        |
| Bases de datos, colecciones y documentos .....      | 84        |
| Carga de datos .....                                | 85        |
| Instrucción insert.....                             | 85        |
| Importación desde ficheros CSV o JSON .....         | 87        |
| Ejemplo: inserción de tweets aleatorios .....       | 88        |
| Consultas simples.....                              | 89        |
| find, skip, limit y sort .....                      | 89        |

|  |            |
|--|------------|
| Estructura general de find .....                                 | 93         |
| Proyección en find .....   | 93         |
| Selección en find .....  | 94         |
| find en Python .....   | 99         |
| Agregaciones .....   | 100        |
| El pipeline .....  | 101        |
| \$group .....  | 101        |
| \$match .....  | 103        |
| \$project .....  | 104        |
| Otras etapas: \$unwind, \$sample, \$out, .....                   | 104        |
| \$lookup .....   | 106        |
| Ejemplo: usuario más mencionado .....                            | 107        |
| Vistas .....   | 108        |
| Update y remove .....  | 109        |
| Update total .....   | 109        |
| Update parcial .....   | 110        |
| Upsert .....   | 112        |
| Remove .....   | 113        |
| Referencias .....  | 114        |
| <b>CAPÍTULO 5. APRENDIZAJE AUTOMÁTICO CON SCIKIT-LEARN .....</b> | <b>115</b> |
| Introducción .....   | 115        |
| NumPy .....  | 115        |

|   |            |
|---|------------|
| pandas ( <i>Python Data Analysis Library</i> ).....         | 117        |
| El conjunto de datos sobre los pasajeros del Titanic.....   | 118        |
| Cargar un DataFrame desde fichero .....                     | 119        |
| Visualizar y extraer información .....                      | 120        |
| Transformar DataFrames .....                                | 124        |
| Salvar a ficheros .....                                     | 125        |
| Aprendizaje automático .....                                | 126        |
| Nomenclatura .....  | 127        |
| Tipos de aprendizaje .....                                  | 128        |
| Proceso de aprendizaje y evaluación de modelos.....         | 129        |
| Etapa de preprocesado .....                                 | 133        |
| Biblioteca scikit-learn .....                               | 136        |
| Uso de scikit-learn.....                                    | 136        |
| Preprocesado .....  | 137        |
| Clasificación .....   | 140        |
| Regresión .....   | 142        |
| Análisis de grupos .....                                    | 144        |
| Otros aspectos de scikit-learn.....                         | 146        |
| Conclusiones .....  | 151        |
| Referencias .....   | 152        |
| <b>CAPÍTULO 6. PROCESAMIENTO DISTRIBUIDO CON SPARK.....</b> | <b>153</b> |
| Introducción.....   | 153        |

|   |            |
|---|------------|
| Conjuntos de datos distribuidos resilientes ..... | 157        |
| Creación de RDDs.....                             | 160        |
| Acciones .....                                    | 162        |
| collect, take y count .....                       | 163        |
| reduce y aggregate.....                           | 164        |
| Salvar RDDs en ficheros.....                      | 167        |
| Transformaciones.....                             | 169        |
| map y flatMap .....                               | 169        |
| filter.....                                       | 171        |
| RDDs de parejas .....                             | 172        |
| Transformaciones combinando dos RDDs.....         | 175        |
| Ejemplo de procesamiento de RDD .....             | 177        |
| Conclusiones .....                                | 180        |
| Referencias.....                                  | 180        |
| <b>CAPÍTULO 7. SPARKSQL Y SPARKML .....</b>       | <b>181</b> |
| SparkSQL .....                                    | 181        |
| Creación de DataFrames .....                      | 182        |
| Almacenamiento de DataFrames .....                | 188        |
| DataFrames y MongoDB .....                        | 190        |
| Operaciones sobre DataFrames .....                | 193        |
| Spark ML .....                                    | 212        |
| Clasificación con SVM.....                        | 214        |

|   |            |
|---|------------|
| Regresión lineal .....                              | 218        |
| Análisis de grupos con k-means .....                | 219        |
| Persistencia de modelos .....                       | 220        |
| Referencias .....                                   | 221        |
| <b>CAPÍTULO 8. VISUALIZACIÓN DE RESULTADOS.....</b> | <b>223</b> |
| Introducción.....                                   | 223        |
| La biblioteca matplotlib .....                      | 223        |
| Gráficas .....                                      | 230        |
| Gráfica circular .....                              | 230        |
| Gráfica de caja.....                                | 233        |
| Gráfica de barras.....                              | 236        |
| Histograma.....                                     | 241        |
| Conclusiones .....                                  | 244        |
| Referencias .....                                   | 245        |
| <b>APÉNDICE. INSTALACIÓN DEL SOFTWARE.....</b>      | <b>247</b> |
| Introducción.....                                   | 247        |
| Python y sus bibliotecas.....                       | 247        |
| Windows 10 .....                                    | 248        |
| Linux.....  | 250        |
| Mac OS.....   | 251        |
| MongoDB .....                                       | 253        |
| Windows 10 .....                                    | 253        |

|                              |            |
|------------------------------|------------|
| Linux.....                   | 256        |
| Mac OS.....                  | 257        |
| Apache Spark y PySpark.....  | 258        |
| Windows 10.....              | 258        |
| Linux.....                   | 259        |
| Mac OS.....                  | 260        |
| <b>ÍNDICE ANALÍTICO.....</b> | <b>261</b> |