Witton Becerra Mayorga, Adriana López, Alex L. Rojas

Python para las humanidades digitales

Catalogación en la publicación - Biblioteca Nacional de Colombia

Becerra Mayorga, Witton, autor
Python para las humanidades digitales / Witton Becerra Mayorga, Adriana López, Alex L. Rojas. -- Primera edición. -- Bogotá : Ecoe Ediciones, 2023.
279 páginas. -- (Computación y tecnología de la información. Programación y desarrollo de software)

Incluye datos curriculares de los autores -- Incluye índices de bloques de código y alfabético -- Incluye referencias bibliográficas.

ISBN 978-958-503-822-6 -- 978-958-503-823-3 (e-book)

1. Python (Lenguaje de programación de computadores) 2. Humanidades - Investigaciones - Procesamiento de datos 3. Programación (Computadores electrónicos) 4. Desarrollo de software I. López, Adriana, autora II. Rojas, Alex L., autor

CDD: 005.133 ed. 23 CO-BoBN– a1131254

Área: *Computación y tecnología de la información*
Subárea: *Programación y desarrollo de software*

ECOE
EDICIONES

info@ecoeediciones.com
www.ecoeediciones.com
Carrera 19 # 63 C 32
Teléfono: (+57) 321 226 46 09
Bogotá, Colombia

Primera edición: Bogotá, noviembre del 2023

ISBN: 978-958-503-822-6
e-ISBN: 978-958-503-823-3

Directora editorial: Ana María Rueda G.
Coordinadora editorial: Paula Bermúdez B.
Editora de adquisiciones: Alejandra Cely R.
Corrección de estilo: Liceth Bohórquez León
Diagramación: Álex L Rojas
Carátula: Sindy Nicol Pulido C.
Impresión: Carvajal Soluciones de Comunicación S.A.S.
Carrera 69 #15-24

Tabla de contenidos

Índice de figuras **v**

Índice de tablas **ix**

Prefacio **xi**

1 Introducción **1**
1.1 Instalación 1
1.2 Conceptos básicos 6
1.3 Acceso a datos 8
1.4 Errores comunes 8

I Obtención de datos **13**

2 Tipos de datos **15**
2.1 Secuencias de caracteres 16
2.2 Falsos y verdaderos 26
2.3 Números 31
2.4 Listas 34
2.5 Tuplas 45
2.6 Diccionarios 45
2.7 Ejercicios 49

3 Trabajando con datos externos **51**
3.1 Archivos de texto sin formato 52
3.2 Archivos con formato `JSON` 58
3.3 Archivos con formato `pdf` 61
3.4 Archivos de datos en formato `csv` y la librería `pandas` . 89
3.5 Ejercicios 99

4 Trabajando con datos en la *Web* **103**
4.1 `HTML` 104
4.2 *Scraping* con `BeautifulSoup` 113

4.3 Extracción de artículos en un currículo `CvLAC` 129
4.4 Ejercicios 138

II Manipulación de datos 141

5 Manipulación de texto 143
5.1 Expresiones regulares 144
5.2 Conjuntos de datos y más de la librería `pandas` 155
5.3 Ejercicios 197

6 Procesamiento natural del lenguaje con `spaCy` 199
6.1 Instalación de `spaCy` y descarga de corpus 200
6.2 `Token`, `Span` y `Doc` 203
6.3 Reconocimiento de entidades nombradas 218
6.4 Ejercicios 227

7 Redes 229
7.1 Conceptos básicos 229
7.2 Centralidad, densidad y diámetro 240
7.3 Comunidades 246
7.4 Ejercicios 251

Referencias 253

Índice de bloques de código 255

Índice alfabético 263

Índice de figuras

1.1 Interfaz de `Jupyter Lab` 3
1.2 Interfaz de `Google Colab` 4
1.3 Navegador `Anaconda` 5
1.4 Interfaz de `VS Code 2` 6
1.5 Cuatro instancias de la clase `cubo` 7
1.6 Organización de la carpeta de trabajo 9
1.7 Palabras reservadas, las cuales fueron obtenidas con el comando `help("keywords")` 10

2.1 Indexación secuencia de caracteres 21
2.2 Visualización de la clasificación Publindex de dos revistas. El verde indica que la condición es verdadera, de lo contrario es lila 29
2.3 Ejemplos del uso de la función `range()` 34
2.4 Visualización de la comprensión de listas en el Ejemplo 2.11 38
2.5 Traducción a código Morse de una secuencia de caracteres 48
2.6 Creación de un iterable con la función `zip()` 48

3.1 Representación gráfica de la selección de caracteres de la secuencia de caracteres `NDP` 54
3.2 Creación de un objeto iterable con la función `zip()` . . . 69
3.3 Representación gráfica de las listas anidadas en el Ejemplo 3.6 74
3.4 Visualización de selección de columnas y filtración de filas en un conjunto de datos 92
3.5 Visualización de una partición de un conjunto de datos utilizando `groupby()` 97
3.6 Diagrama de dispersión del número de visitas vs. el número de descargas 98
3.7 Diagrama de caja para el promedio de descargas diarias . 98

4.1 Ilustración de una estructura básica en HTML 105

4.2 Bosquejo de la estructura de la plantilla CvLAC 114
4.3 Número de artículos por año por revista 135
4.4 Número de artículos por cantidad de autores 137

5.1 Visualización de los atributos `iloc` y `loc` en conjunto de datos 167
5.2 Distribución de revistas homologadas por categorías por año 176
5.3 Proporción de palabras en una frase para cada emoción y sentimiento con base en léxico Emolex 196

6.1 Representación gráfica de un objeto de tipo `Doc` 202
6.2 Relaciones de dependencia sintáctica para el objeto `span1` 207
6.3 Visualización de las dependencias sintácticas para cuatro frases similares 209
6.4 Representación gráfica de *tokens* como vectores 210
6.5 Proyección en dos dimensiones de los vectores de palabras más cercanos y más alejados a la palabra `'Trump'` 214
6.6 Entidades nombradas detectadas por `spaCy` 218
6.7 Entidades nombradas detectadas en un segmento del texto 'Las Campañas del Sur' 221
6.8 Pantallazo del menú para especificar el archivo de configuración del modelo de etiquetado 225
6.9 Pantallazo del archivo de configuración modificado 226
6.10 Entidades nombradas detectadas en un segmento del texto 'Las Campañas del Sur' con el modelo entrenado 227

7.1 Redes de coautoría 230
7.2 Redes de coautoría para dos artículos. El grosor de las aristas representa el número de artículos en los que los autores han colaborado 233
7.3 Red de colaboradores del profesor García Ubaque con base en su currículo CvLAC 236
7.4 Red de colaboradores del profesor García Ubaque con base en su currículo CvLAC en un posicionamiento circular . . 238
7.5 Red de colaboradores del profesor García Ubaque con base en su currículo CvLAC utilizando `pyvis` 240
7.6 Red de colaboradores del profesor García Ubaque con base en su currículo CvLAC eliminando el nodo más central . . 241

7.7 Red de coautoría de la Escuela de Matemáticas y Estadística 244
7.8 Parte de la red de coautoría de la Escuela de Matemáticas y Estadística de la UPTC con base en los currículos CvLAC 246
7.9 Comunidades de coautoría de la Escuela de Matemáticas y Estadística de la UPTC con base en los currículos CvLAC 248
7.10 Comunidades de coautoría más densas y conectadas entre ellas, de la Escuela de Matemáticas y Estadística de la UPTC con base en los currículos CvLAC 250

Índice de tablas

1.1 Librerías que utilizamos en este libro y están disponibles en la distribución `Anaconda` 4

3.1 Número de visitas y descargas de artículos de los once artículos en el volumen 45 de la revista "La Palabra" . . 91
3.2 Instrucciones básicas para filtrar filas, seleccionar columnas y calcular estadísticas en un conjunto de datos `c`, con columnas `col1`, `col2` y `col3`. La columna `col1` contiene datos numéricos y `col2` secuencias de caracteres. 93
3.3 Fecha de publicación de los artículos en la sección "Música y Literatura" 94
3.4 Estadísticas descriptivas de descargas para los artículos publicados hace menos de 5 meses 95
3.5 Sección y descargas diarias para artículos frecuentemente descargados 96
3.6 Promedio de descargas diarias por sección 96

4.1 Número de artículos publicados en la Revista de Salud Pública y Tecnura por año. 134
4.2 Número de artículos publicados por cantidad de autores en el currículo del profesor García Ubaque 136

5.1 Conjuntos de caracteres predefinidos para las expresiones regulares 145
5.2 Cuantificacores para expresiones regulares 146
5.3 Funciones del módulo `re` comúnmente usadas 151
5.4 Subconjunto de revistas con un segundo ISSN para los años 2016 y 2017 166
5.5 Instrucciones básicas para filtrar filas y seleccionar columnas con los atributos `iloc` y `loc` en un conjunto de datos `cd` 168

5.6 Filas con la homologación por año de la revista "SEL Studies in English Literature" 169
5.7 Subconjunto de revistas con ISSN igual a `1467-9442` o `0347-0520` 171
5.8 Número de artículos publicados por autor en el primer número de la Revista Tecnura 186
5.9 Número de artículos por año en la Revista Tecnura por Luis Fernando Pedraza Martínez 189
5.10 Nombre del autor con más artículos publicados, por año, en la Revista Tecnura 190
5.11 Nombre de los autores con más artículos publicados, por año, en la Revista Tecnura 191
5.12 *Top* 10 de los autores que más publican en la Revista Tecnura 192
5.13 Asignación de puntajes en emociones y sentimientos para la palabra guerra 194
5.14 Asignación de puntajes en emociones y sentimientos para la palabra loco 194

6.1 Atributos del objeto `Token` 204
6.2 Atributos para los tokens en la variable `span1` 205
6.3 Categoría gramatical y lema para los tokens en `span1`. . 206
6.4 Tipos de entidades que involucran nombres 217
6.5 Tipos de entidades que involucran números 217

Prefacio

Las Humanidades Digitales son un campo interdisciplinar, que combinan el estudio tradicional de las Ciencias Humanas con la programación, para abordar los fenómenos y objetos acostumbrados de estas disciplinas de una forma distinta, gracias a la programación, la cual permite abordar las investigaciones desde una perspectiva diferente.

Si bien, bajo el espectro de este concepto subyacen muchas posibilidades las que aquí se quieren expresar consisten, básicamente, en la acumulación, manipulación, análisis y estudio de datos de la literatura, la lingüística, la historia, el derecho, el periodismo etc. La base principal de estas disciplinas es, por ejemplo, los objetos físicos como libros, al ser estos digitalizados permiten su abordaje de manera distinta gracias a la recolección de datos mediante programación que permite diseñar nuevos protocolos de investigación para estudiar las humanidades de manera más amplia, con resultados más completos, y que generan novedades respecto a los tradicionales objetos de estudio.

Como antecedentes, existen varios proyectos que generaron esta nueva forma de ver el estudio de las humanidades, por ejemplo: "Michael Hart creó el Proyecto Gutenberg en 1971, que se convirtió en uno de los primeros proyectos de digitalización de libros. El objetivo del proyecto era hacer que los libros de dominio público estuvieran disponibles en línea y gratuitamente. En 2019, el Proyecto Gutenberg había digitalizado más de 60,000 libros." (Warwick, 2012, p. 13). En este mismo sentido se crea en la década siguiente el Text Encoding Initiative (TEI) que: "es un consorcio internacional que desarrolla y mantiene un estándar para la codificación de textos electrónicos. La TEI se creó en 1987 para proporcionar una metodología para la creación de textos electrónicos, y ha sido fundamental en la creación de proyectos de digitalización de textos. La TEI es ampliamente utilizada en proyectos de humanidades digitales que implican la digitalización y análisis de textos." (Warwick, 2012, p. 20). Hasta aquí de lo que trata el asunto es de la digitalización de textos.

Como se observa, la iniciativa de estos proyectos se fundamenta en la digitalización de contenido físico, especialmente libros. En este sentido, en principio, se propuso la digitalización de varios textos patrimonio de la humanidad o como móvil se tuvo en cuenta la acumulación y el rescate de ese material mediante la digitalización de estos. Así, el Proyecto Perseus: "es una iniciativa que comenzó en 1987 con el objetivo de crear una biblioteca digital de textos y recursos en áreas como la literatura, la historia y la filosofía. El proyecto también incluye herramientas de análisis y visualización de datos. La biblioteca digital del Proyecto Perseus se ha convertido en una de las más utilizadas por los investigadores de humanidades digitales." (Schreibman et al., 2008, p. 31). De la misma manera, los autores señalan los estudios de la Biblioteca de Alejandría como un pilar para esta mirada de esta nueva manera de acumular los textos: "En 1991, la UNESCO lanzó una iniciativa para recrear la Biblioteca de Alejandría en formato digital. El proyecto involucra la digitalización y análisis de textos antiguos y la creación de una biblioteca digital que se pueda acceder desde cualquier parte del mundo. Los estudios de la Biblioteca de Alejandría han sido fundamentales en la creación de una red global de bibliotecas digitales." (Schreibman et al., 2008, p. 33).

Como se observa, estos grandes referentes proponen la creación de fondos bibliográficos digitales. La pregunta es, entonces, qué pasó después de confirmar que se podían crear numerosas cantidades de textos digitales. Acá nace la perspectiva actual de las Humanidades Digitales que consiste en estudiar mediante programas por computador estos textos.

Al tener datos digitalizados en abundancia, surgen posteriormente, las metodologías para estudiar los textos. Por eso, hoy en día, la perspectiva de las Humanidades Digitales consiste en estudiar estas grandes cantidades de texto mediante programas de computador a partir de modelos de estudio más amplios y completos en palabras de Jockers:

> Digital Humanities is not just about the application of digital technologies, but also about the development of new theories and methods for humanities research, such as social network analysis, data mining, visualization, and machine learning. These approaches enable researchers to analyze large and complex datasets, uncover patterns and relationships that are not immediately apparent, and generate new insights into human culture and society." (2013, p. 34)

En conclusión, en lo que se profundiza hoy en las Humanidades Digitales

es metodologías como análisis de redes sociales, minería de datos, la visualización de datos mediante grafos que muestren los textos y sus relaciones y el ya famoso *machine learning* que ha producido una revolución como es `ChatGPT`.

Este libro está escrito para cualquier humanista que esté interesado en aprender a programar en `Python`. En nuestra experiencia, este interés usualmente viene acompañado de un problema real que el individuo quiere resolver y ha escuchado que puede ser resuelto con `Python`. Por esta razón, recomendamos que a medida que avancemos con el material, no solo se practique con los ejemplos del libro, sino con material relacionado con el problema que atañe al humanista.

En la primera parte del libro, nos enfocamos en aprender las estructuras básicas para almacenar información en `Python` y su manipulación básica. También, extraemos datos de archivos de tipo texto, archivos `pdf` y páginas en internet con la ayuda de librerías como `BeatifulSoup` y `pypdf`. Los archivos de texto y `pdf` pueden estar disponibles en nuestro disco duro o en internet, por ende, es necesario tener una conección a internet para trabajar en el material propuesto. Para lograr extraer y manipular la información en estos archivos, es fundamental conocer y manejar las estructuras básicas de `Python`, al igual que otras estructuras disponibles en librerías escritas específicamente para manipular cierta clase de información. Por esta razón, es importante tener acceso a una distribución de `Python` que tenga instaladas estas librerías por defecto. Como explicamos en el primer capítulo, emplearemos la distribución `Anaconda`, así, no nos preocupamos por instalar ninguna librería extra, ni por incompatibilidades.

A diferencia de otros libros introductorios de `Python`, donde el uso de conceptos matemáticos es el eje central para presentar todo el material, en este libro nos enfocamos en trabajar con texto. Existen varios formatos para almacenar texto, pero no los consideramos todos en este libro. Nos enfocamos en texto sin formato, y los formatos `csv`, `HTML` y `JSON`. No obstante, si es necesario interactuar con algún otro formato, creemos que con el material que aprendamos en este libro, poseemos las herramientas para leerlo y manipularlo.

Como humanistas digitales, sabemos que el texto es una fuente extremadamente rica de información. Por ejemplo, en el tiempo desde que iniciamos a leer este libro, se han creado cientos de mensajes de texto, correos, comentarios en diferentes plataformas, etc. Además, de todo el

texto que se encuentra disponible en internet y en bases de datos. Una vez que tenemos acceso a texto, debemos organizarlo de una manera estructurada y asegurarnos que no tenga errores, en la medida de lo posible.

En la segunda parte del libro, consideramos herramientas más especializadas para el manejo de texto y de conjuntos de datos. En el Capítulo 5, iniciamos profundizando en el uso de expresiones regulares para lograr generalizar patrones de caracteres. Luego, retomamos el estudio de los conjuntos de datos y aprendemos otras herramientas para su manipulación.

En los dos últimos capítulos del libro, consideramos dos herramientas comúnmente utilizadas por los humanistas digitales: grafos y procesamiento natural del lenguaje. En el Capítulo 6, hacemos una introducción al procesamiento natural del lenguaje, utilizando la librería `spaCy`. Mientras que, en el Capítulo 7, aprendemos a graficar redes y obtener estadísticas básicas de estas. Nuestro tratamiento de estos dos temas es de carácter introductorio, debido a que un tratamiento profundo requiere de conocimiento de material más avanzado.

Todos los ejemplos de este libro pueden ser replicados, ya que todas las fuentes de información son de libre acceso. Así que, invitamos a nuestros lectores a repetir y modificar nuestros ejemplos. Para terminar, este libro fue digitado utilizando `Quarto`[1], y todas las figuras fueron creadas por los autores utilizando LaTeX, `Tikz`, `mermaid` y `Python`.

[1] https://quarto.org/docs/books