MÁSTER EN
CIENCIA DE DATOS

Tipología y ciclo de vida de los datos

(M2.851)

PRÁCTICA 1

Juan Miguel Iglesias Labraca

Elías Páez de la Rosa

8 de noviembre de 2021

EJERCICIO 1

1. CONTEXTO.

La búsqueda de la información se ha centrado en encontrar una web que cumpla los siguientes requisitos:

- Contenga una tabla con un mínimo aceptable de registros.
- Que la información esté estructurada dentro del código HTML para facilitar la conversión a un dataframe.
- Que la información sea interesante o relevante.
- Que para llegar a la información se deba interactuar primero con la web con el objetivo de poner en práctica técnicas más difíciles de webscraping.

Tras varias pruebas en diferentes webs se ha optado finalmente por escoger GoodReads ya que, además de su popularidad en el mundo de la lectura, cumple los requisitos anteriormente mencionados.
GoodReads dispone de numerosos listados de libros en los que se incluye el autor y varias métricas como, por ejemplo, la cantidad de usuarios que lo han valorado y las puntuaciones medias.

2. TÍTULO.

El título del dataset va a ser: “GoodReads_Best_Books_2021”.

Se entiende que la palabra “Best” es muy subjetiva, en este caso GoodReads considera como “mejores” a una combinación entre la cantidad de usuarios que han votado y que han comentado un libro en concreto ponderado por las puntuaciones otorgadas.

3. DESCRIPCIÓN DEL DATASET.

El dataset extraído se contiene los siguientes campos:

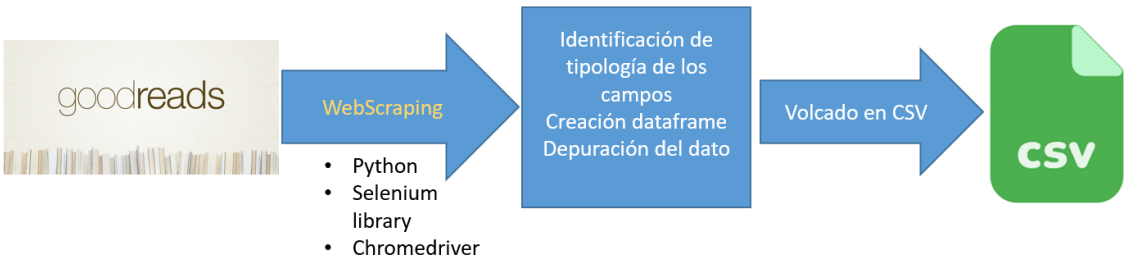
- **ID**: identificador único del registro, también sirve como indicador del puesto de los mejores libros.
- **Título**: Título del libro
- **Autor**: Autor/es del libro
- **Avg_rating**: puntuación media que han dado los usuarios de GoodReads al libro.
- **Ratings**: Cantidad de usuarios de GoodReads que han puntuado el libro

- **Score**: puntuación acumulada específica para este listado de libros de GoodReads y que marca el puesto dentro de este listado. Un mismo libro puede tener diferentes puntuaciones dependiendo del listado en el que se muestra, por ejemplo, el mejor libro de ciencia ficción del 2021 podría ser el décimo puesto en los mejores libros de ciencia ficción del siglo XXI.
- **People_voted**: cantidad de usuarios de GoodReads que han votado a ese libro en este listado en concreto.

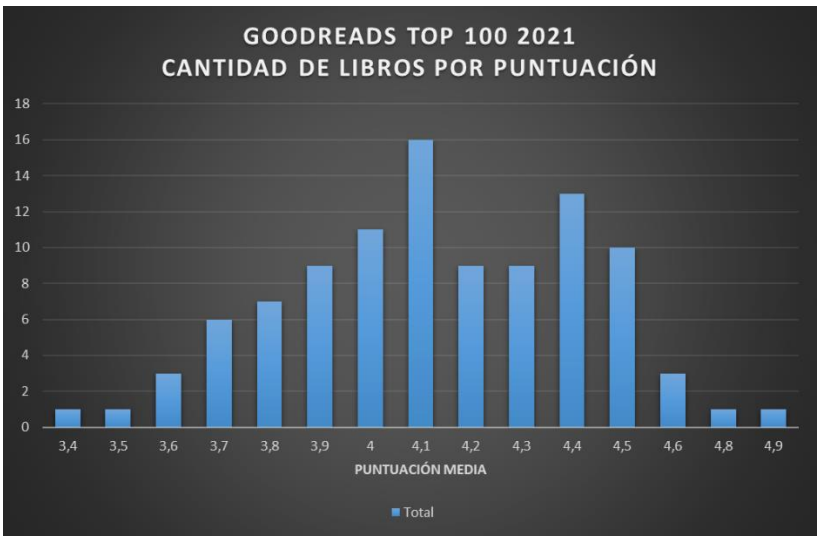
	A	B	C	D	E	F	G
1	**ID**	**Titulo**	**Autor**	**Avg_rating**	**Ratings**	**Score**	**People_voted**
2	1	Project Hail Mary	Andy Weir	4,54	132751	11,32	114
3	2	The Four Winds	Kristin Hannah	4,32	277604	8,881	90
4	3	Malibu Rising	Taylor Jenkins Reid	4,12	212764	6,734	69
5	4	A Court of Silver Flames (A Court of Thorns and Roses, #4)	Sarah J. Maas	4,39	18792	6,7	68
6	5	People We Meet on Vacation	Emily Henry	4,11	21254	5,183	53
7	6	Karma: A Yogi s Guide to Crafting Your Destiny	Sadhguru	4,51	1804	4,999	50
8	7	The Rose Code	Kate Quinn	4,48	67146	3,618	37
9	8	Klara and the Sun	Kazuo Ishiguro	3,81	102001	3,609	37
10	9	The Lost Apothecary	Sarah Penner	3,78	118164	3,329	34
11	10	The Push	Ashley Audrain	4,13	119521	2,947	30
12	11	When Strivings Cease: Replacing the Gospel of Self-Improvement with the Gospel of Life-Transfo	Ruth Chou Simons	4,86	196	2,598	26
13	12	Rule of Wolves (King of Scars, #2)	Leigh Bardugo	4,41	59026	2,561	26
14	13	The Soulmate Equation	Christina Lauren	4,07	44763	2,428	26
15	14	The Zero Signal (Science Crimes Division #1)	Rick Wayne	4,58	40	2,393	24
16	15	My Wife Jodie	V.A. Rudys	4,09	393	2,392	24
17	16	The Mary Shelley Club	Goldy Moldavsky	3,97	3733	1,867	19
18	17	Fugitive Telemetry (The Murderbot Diaries, #6)	Martha Wells	4,33	23517	1,784	18
19	18	Any Way the Wind Blows (Simon Snow, #3)	Rainbow Rowell	4,22	22179	1,765	18
20	19	The Last Thing He Told Me	Laura Dave	3,97	199352	1,718	18
21	20	Chain of Iron (The Last Hours, #2)	Cassandra Clare	4,46	37764	1,677	17
22	21	The Wife Upstairs	Rachel Hawkins	3,77	9125	1,659	17
23	22	Concrete Rose (The Hate U Give, #0)	Angie Thomas	4,44	35357	1,583	16
24	23	Mina and the Undead (Mina and the Undead #1)	Amy McCaw	4,08	754	1,582	16
25	24	Last One to Die	Cynthia Murphy	3,59	697	1,577	16

4. REPRESENTACIÓN GRÁFICA.

Flujo del programa:



Representación gráfica de ejemplo del dataset:



5. CONTENIDO.

El dataset extraído se contiene los siguientes campos:

- **ID**: identificador único del registro, también sirve como indicador del puesto de los mejores libros.
- **Título**: Título del libro
- **Autor**: Autor/es del libro
- **Avg_rating**: puntuación media que han dado los usuarios de GoodReads al libro.
- **Ratings**: Cantidad de usuarios de GoodReads que han puntuado el libro
- **Score**: puntuación acumulada específica para este listado de libros de GoodReads y que marca el puesto dentro de este listado. Un mismo libro puede tener diferentes puntuaciones dependiendo del listado en el que se muestra, por ejemplo, el mejor libro de ciencia ficción del 2021 podría ser el décimo puesto en los mejores libros de ciencia ficción del siglo XXI.
- **People_voted**: cantidad de usuarios de GoodReads que han votado a ese libro en este listado en concreto.

Los datos se han extraído a fecha 1 de noviembre de 2021.
Para extraerlos se ha utilizado Python con Jupyter Notebooks, de manera más específica con la librería Selenium para hacer el web scrap y chromedriver para poder navegar e interactuar con la página web GoodReads y alcanzar el apartado HTML que contiene el dataset que hemos marcado como objetivo para esta práctica.

6. AGRADECIMIENTOS.

Se agradece a Good Reads por poner a disposición los listados de libros en su web www.goodreads.com.
Igualmente se agradece a Jason Huggins[1] por ser el creador de la librería Selenium (2004), la cual se ha utilizado en el código Python para acceder al contenido de la web de Good Reads.

7. INSPIRACIÓN.

El conjunto de datos nos ha parecido interesante por los siguientes motivos:

- La web permite actuar conforme a los principios éticos y legales del presente proyecto, en los que se busca información pública accesible a técnicas de web scraping.
- Son datos culturales de interés general, específicamente una lista de los libros mejor valorados del año es algo relevante para los amantes de la lectura.
- Se ha querido utilizar Selenium en vez de otras librerías más sencillas como BeautifulSoup con el objeto de la mejora personal de los participantes de la práctica.
- La idea es poder utilizar el dataset para posteriores trabajos de análisis y minería de datos.

8. LICENCIA.

Se ha optado por *"Creative Commons Attribution 4.0 International"* por ser una de las licencias más utilizadas que permiten compartir, adaptar y modificar a partir del material publicado, siempre y cuando se de crédito al creador original y no se apliquen restricciones adicionales ya sean tecnológicas o legales a los posteriores proyectos derivados del original.

9. CÓDIGO.

El código, dataset y documentación se encuentran en GitHub en el siguiente enlace: https://github.com/eliaspaez/MDS-TYCDD-PRA1

[1] https://en.wikipedia.org/wiki/Selenium_(software)

10. **DATASET.**

    El dataset ha sido publicado en Zenodo en el siguiente enlace: https://zenodo.org/record/5636151

11. **TABLA DE CONTRIBUCIONES.**

CONTRIBUCIONES	FIRMA
Investigación previa	Juan Miguel Iglesias Labraca Elías Páez de la Rosa
Redacción de las respuestas	Juan Miguel Iglesias Labraca Elías Páez de la Rosa
Desarrollo del código	Juan Miguel Iglesias Labraca Elías Páez de la Rosa

BIBLIOGRAFÍA

- ✓ Subirats, L., Pérez, D., Calvo, M. (2019). Introducción al ciclo de vida de los datos. Editorial UOC.
- ✓ Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.
- ✓ Tutorial de Github (GitHub Guides website) https://guides.github.com/activities/hello-world.
- ✓ Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- ✓ Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.
- ✓ *Chrome Driver. Chromium.org (2021). Version 94.0.4606.61* *https://chromedriver.chromium.org*
- ✓ *Zenodo website (2021)* https://zenodo.org