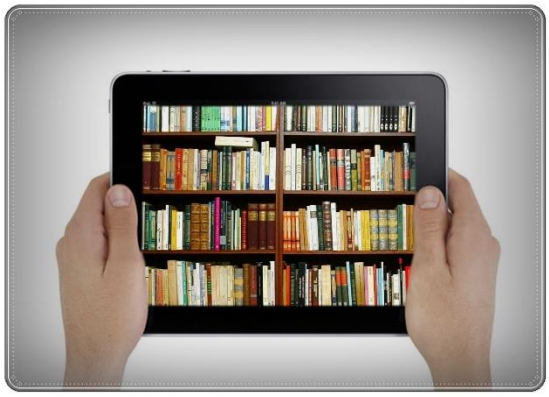
Inmerso en un mar de libros



Contexto

Como parte del curso "Tipología y ciclo de vida de los datos" del *Master in Data Science* de la UOC, se nos solicitó seleccionar una página web y aplicar en ella técnicas de *web scrapping* por medio de herramientas como Python o R y así, generar un conjunto de datos.

El conjunto de datos generado corresponde a libros que han sido publicados en la página BookRix, la cual, es una plataforma en el que las personas pueden leer libros creados por otros usuarios y pueden comentar sobre el libro que están leyendo.

Descripción del conjunto de datos

Como se mencionó en el contexto, el conjunto de datos recolecta información proveniente del sitio BookRix, el cual, cuenta con más de 130.000 libros, los cuales en su mayoría han sido escritos por escritores independientes y que han publicado sus libros ahí para lectura y comentarios de retroalimentación.

Entre la información que podemos encontrar, no solo el nombre del libro y del autor, sino también el número de palabras que tienen estos, lo cual ayuda a dimensionar el tamaño del libro; el idioma en que este fue escrito, la categoría de este, entre otros.

Contenido

La base de datos fue creada utilizando la técnica de *web scrapping* por medio de código de programación en Python. Cada libro corresponde a un registro y este *dataset* incluye todos los libros que han sido publicados en BookRix (https://www.bookrix.com/) al día 9 de noviembre del 2019. Los campos contenidos son:

- **Author Name**: corresponde al nombre del escritor del libro
- **Book Name**: contiene el nombre del libro
- **Categorie**: indica el género literario del libro (aventuras, ciencia ficción, romance, etc.)
- **Language**: especifica el idioma en el que está escrito el libro.
- **Words**: muestra la cantidad de palabras que tiene el libro.
- **Age**: edades a partir de las cuales se recomienda leer el libro.
- **Views**: cantidad de veces que el libro ha sido visto o leído.
- **Favorite**: número de veces que el libro ha sido marcado como favorito.

El contenido de este *website* se rige bajo las leyes de Alemania y no infringe las políticas de copyright.

Agradecimientos

Los datos han sido recolectados desde la página web de BookRix, quienes son los creadores de una plataforma web 2.0 que permite a los usuarios leer libros, publicar (en caso de que tenga) y comentarlos.

Por lo que se agradece no solo a BookRix por la iniciativa de la plataforma, sino también, a todos los usuarios que han publicado sus libros ahí para el disfrute y críticas de muchos lectores.

Inspiración

En el mundo de los lectores, se han venido presentando ciertos perfiles de personas entre los que se podrían mencionar:

- Los amantes de los *best-sellers*.
- Los que siguen a uno o varios autores y leen todo lo que ellos publiquen.
- Los que leen únicamente sobre un género literario.
- Los que leen de todo.

Usualmente, estos lectores suelen ir a librerías a buscar sus libros, o bien, compran *ebooks* por medio de Amazon u otras plataformas que usualmente tienen a disposición los libros de los autores con mayor trayectoria.

Pero ¿qué sucede cuando encontramos una plataforma con miles de libros que no son tan famosos? ¿Qué sucede cuando nadie conoce a los escritores?

Esto es lo que nos viene a brindar BookRix, una plataforma donde los amantes de la lectura, de las cosas nuevas y diferentes pueden leer libros no comunes, enamorarse de ellos, comentarlos, criticarlos y más.

Esta base de datos nos permitirá analizar en qué idioma se escriben la mayoría de los libros y cuál es el idioma que más se lee, qué género literario es el mas publicado y cuál es el mas leído. ¿Son los libros con menos palabras los que más se leen o el tamaño del libro no importa? ¿Los autores suelen escribir solo de un tipo de genero literario o prefieren diversificar y escribir de varios?

Estos datos, podrían ayudarle a escritores y futuros escritores a conocer las tendencias, a ver si lo que ellos tienen pensado tendrá éxito o no, saber sobre lo más leído, hacia donde van las tendencias, etc. Incluso, podría ser utilizada por editores de libros para ver qué escritores independientes han tenido éxito, ayudarles a publicar sus libros y generar ganancias para ambos.

Licencia

La licencia que se seleccionó para esta publicación es la **CC0 1.0 Universal (CC0 1.0),** debido a que la plataforma que esta sirviendo como fuente primaria de datos, busca no solo crear una comunidad donde lectores y escritores puedan interactuar, sino también promover e impulsar escritores independientes.

Al utilizar una licencia dedicada al domino publico y renunciando a los derechos autorales, lo que busca es que las personas tengan acceso a ellos, que puedan copiarlos, modificarlos, distribuirlos e

interpretarlos, independientemente si es para fines educativas o comerciales, sin tener que pedir permiso para el uso de estos.

Es importante mencionar que esta licencia no afecta en ninguna forma los derechos de patentes o de marcas sobre la obra, ni derechos que otras personas puedan tener en la obra o en cómo la obra es usada, como derechos de publicidad o privacidad.

Código y dataset

El dataset fue creado mediando código de programación en Python, utilizando la técnica de web scrapping en la pagina de BookRix. Estos dos pueden ser accedidos en los siguientes links:

- Codigo: https://github.com/jar2015/scific-books-web-scrapping/tree/master/src

- Dataset: https://github.com/jar2015/scific-books-web-scrapping/tree/master/csv

Recursos

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Tutorial de Github https://guides.github.com/activities/hello-world