Dataset: Libros ganadores de los GoodReads Awards

José Luis Rodríguez Andreu

Mayo de 2021

Descripción

Este proyecto tiene como objetivo generar un conjunto de datos que reúne información sobre los libros ganadores de los premios Goodreads Awards, del sitio web www.goodreds.com. La aplicación de captura de datos toma como entrada un año determinado, y genera un fichero csv con el ganador de cada categoría para ese año, el cual contiene información sobre el autor, el número de votos o la puntuación del libro en la propia plataforma, además de información del propio libro.

Representación gráfica

El ciclo de vida de este proyecto se corresponde de los siguientes pasos:

- La aplicación toma un valor correspondiente al año indicado del cual se quieren extraer los datos. En este punto, la aplicación realiza scrapping a la url principal de los premios del año indicado:

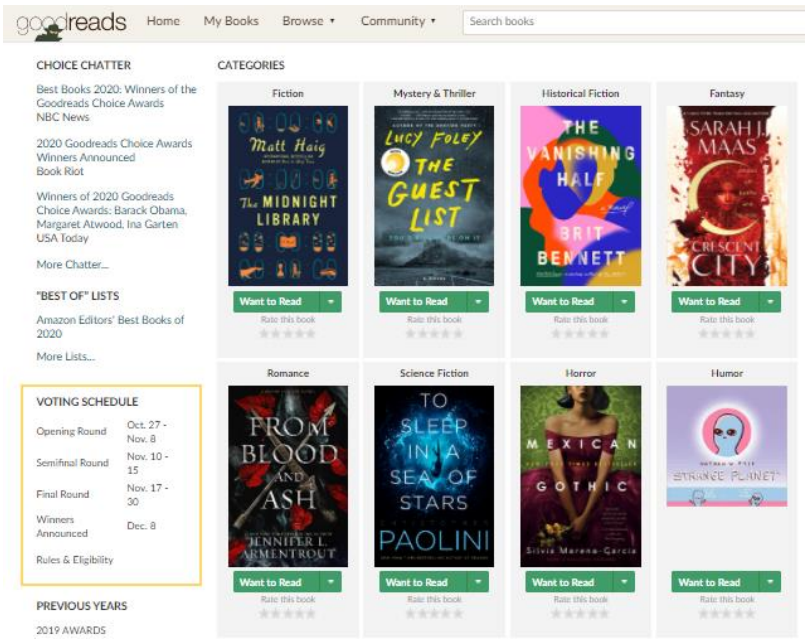
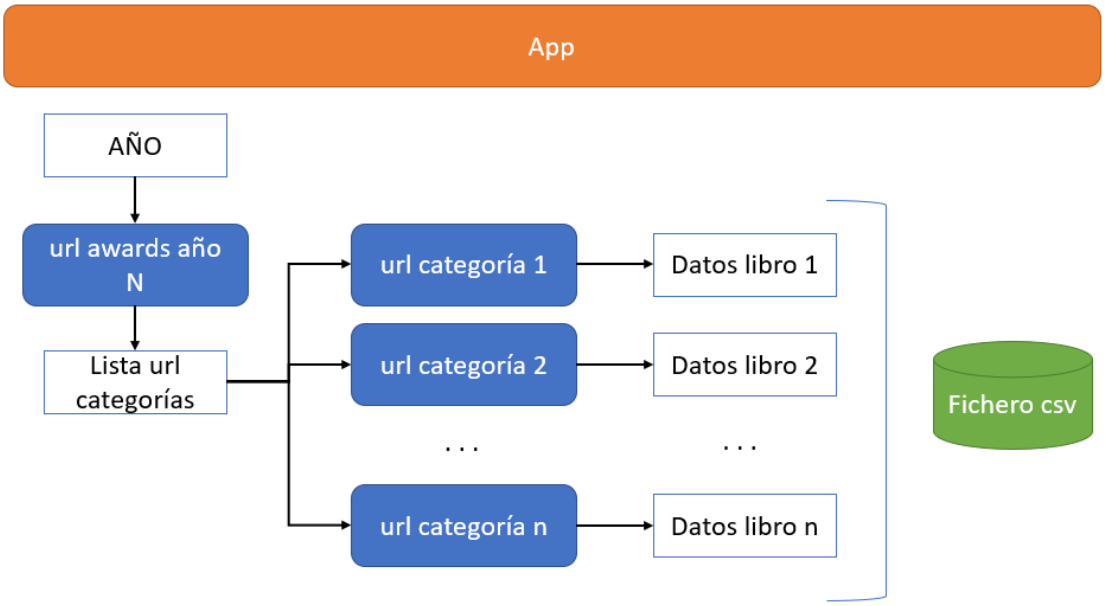


Figura 1: Extracto del sitio web www.goodreds.com con los ganadores del Goodreads Awards de 2020.

- A partir de esta url, se obtiene una lista de las url asociadas a cada obra literaria ganadora en cada categoría. De estas url, se obtiene la información indicada en el proyecto para cada libro.

Figura 2: Extracto del sitio web www.goodreds.com con la página principal del libro ganador en 2020 en la categoría de ficción.

La aplicación sigue el diagrama siguiente:



Contexto

Goodreads es una plataforma a modo de red social, donde los usuarios comparten información sobre las obras literarias que han consumido, a la vez que realizan recomendaciones, puntuaciones en sus libros favoritos, u opiniones. La plataforma, a su vez, sirve como base de datos tanto de obras como de autores, pudiendo utilizarse para buscar información sobre las publicaciones de un determinado autor.

Cada año, Goodreads organiza los Goodreads Awards, donde los usuarios votan entre una serie de obras divididas en diferentes géneros literarios que han sido nominadas en base a su popularidad dentro de la plataforma. Son los propios usuarios los que deciden que obra es la ganadora en cada categoría.

Contenido

La aplicación genera un fichero csv de nombre **goodreads_awards_[year].csv**, donde **year** se corresponde al año indicado a la aplicación. En el caso de que se indique un año erróneo o para el cual no haya datos sobre los premios, la aplicación devuelve un mensaje de error.

Cada uno de los registros del conjunto de datos se corresponde al ganador de una determinada categoría para el año escogido, donde se recogen las siguientes características:

- **Category**: Categoría en la que el libro ha resultado ganador.
- **Title**: Título de la obra.
- **Votes**: Número de votos alcanzados por la obra.
- **Autor_name**: Nombre del autor.
- **Book_series**: Serie literaria a la que pertenece el libro, si fuera el caso.
- **Rating_value**: Valoración media (en una escala de 0 a 5) alcanzada en Goodreads.
- **Num_ratings**: Número de valoraciones recibidas por la obra.
- **Num_reviews**: Número de opiniones realizadas por los usuarios para este libro.
- **List_genres**: Lista de géneros asociados al libro según la plataforma y los usuarios. Se corresponde a una cadena de caracteres donde los géneros están separados por el carácter '_'.
- **Book_format**: Formato del libro (tapa dura, audiolibro…).
- **Num_pages**: Número de páginas.
- **Publish_date**: Fecha de publicación.
- **Original_title**: Título de la obra en el idioma original.
- **Isbn**: ISBN, o código identificativo de la edición.
- **Edition_language**: Idioma de la edición.
- **Setting**: Lugar donde transcurre el libro.
- **Num_awards**: Número de premios que la obra ha recibido.
- **Year**: Año de los premios.

Agradecimientos

Como ya se ha comentado, los datos han sido recopilados de la plataforma Goodreads. Para ello, se ha desarrollado una aplicación de Web Scrapping mediante el lenguaje de programación Python, con el objetivo de extraer información de la página web.

Existen una serie de APIs desarrolladas para Goodreads, destacando especialmente el paquete de Python **scrapereads** (https://github.com/arthurdjn/scrape-goodreads), que permite obtener resultados de búsqueda para autores y obras.

Inspiración

La principal motivación para la construcción de este conjunto de datos es el interés por realizar un análisis de datos que nos permita encontrar que patrones o características se pueden encontrar en las obras ganadoras, ya sean asociadas a la popularidad del autor, el género, o el tamaño de la obra.

Una continuación de este proyecto de captura de datos sería mejorar la propia aplicación para generar un conjunto de datos que no solo obtenga información de los ganadores, sino que también obtenga los registros asociados a las obras nominadas para cada género. De esta forma, se plantea un proyecto de minería de datos que tenga como objetivo encontrar que características diferencian las obras ganadoras de las nominadas, y emplearlas para la construcción de un modelo de clasificación.

Licencia

Código fuente y dataset

El código y el conjunto de datos se encuentra en el siguiente repositorio:

https://github.com/josrodand/goodreads_awards_scrapper

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.