Práctica 1 Web Scraping

Alejandra Cristina Marrero Suarez
Luis Miguel Moreno López

Contexto

Nuestro proyecto surge con el objetivo de **aventurar qué géneros y autores de libros están teniendo mayor impacto**, tendencia o éxito en España en un período reciente de tiempo.

Creemos que contar con estos datos puede ser importante para las librerías, tanto para realizar acciones inmediatas (**compras de stock de libros o planificar encuentros con autores del momento**) como para crear modelos de predicción que las ayuden con esto mismo a futuro, creando así una ventaja competitiva para aquellas que los utilicen.

Por tanto, esta información bien utilizada puede suponer desde un **ahorro a la hora de realizar compras a las editoriales** hasta un **aumento de las ventas** si se organizan reuniones con aquellos autores de éxito.

Adentrándonos ahora en detalles más concretos del proyecto. En un primer momento se valoraron varias páginas de donde extraer la información como Amazon, el Corte Inglés ...

Finalmente se decidió seleccionar los libros más vendidos recientemente usando la técnica de *web scraping* sobre la **página web wwwtodos.tuslibros.com** [1], en particular sobre su subdominio *mas_vendido* donde aparecen los 100 libros más vendidos.

No es casualidad elegir esta página como fuente de datos ya que, dentro de esta asociación, se encuentran más de 400 librerías de todo el país; siendo así una muestra realmente representativa. De hecho la iniciativa de esta web está promovida desde la Confederación Española de Gremios y Asociaciones de Librerías (CEGAL), lo que otorga una alta credibilidad a estos datos.

Otro punto importante para optar por esta web es, sin duda, la especialización con la que cuenta el colectivo que se encuentra detrás de ella. Otros vendedores de libros como los mencionados al principio no se dedican solo a este mercado. Nos hemos decantado por aquellos datos que provienen de vendedores que se dedican al mercado del libro de forma exclusiva.

Título del dataset

Titulamos nuestro dataset: ***libros_más_vendidos***.

Descripción del dataset

Ofrecemos la información habitual asociada a un libro (autor, número de páginas, …) de los 100 libros más vendidos en este momento en España.

Cabe decir que en el siguiente apartado se muestra visualmente el conjunto de datos que componen un libro, siendo este numeroso. Posteriormente en el apartado de contenido se explica cada uno de los componentes de nuestro dataset de forma exhaustiva.
Aun así en este punto vamos a argumentar por qué se guardan algunos de estos atributos:

- El puesto del ranking, es fundamental para poder establecer una jerarquía de importancia de los libros de la lista. Aunque todos estén en el top 100, no es lo mismo estar en el puesto 2 que en el 90.

- El título pretende ser un identificador natural, de tal manera que se pueda concebir una referencia consolidada de un atributo clave al que luego se le puedan ligar mentalmente el resto de datos de la misma fila.

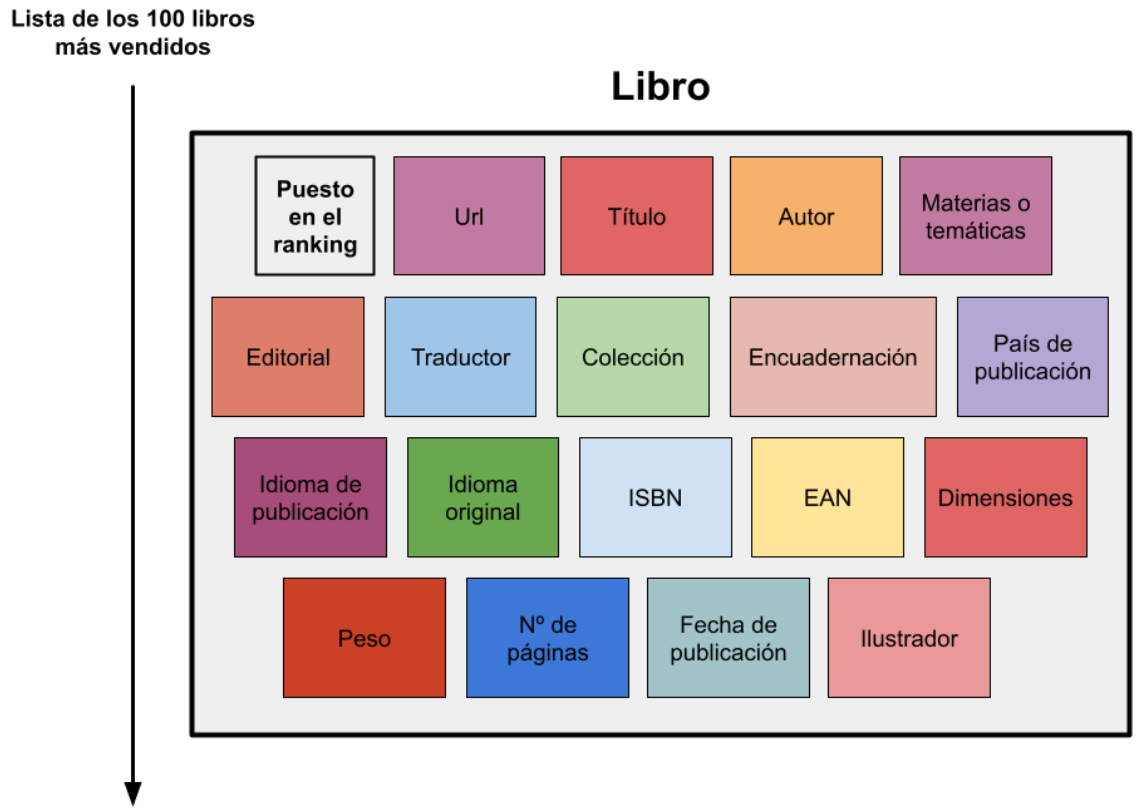
- Tanto el ISBN y el EAN son identificadores técnicos. Permiten diferenciar inequívocamente cada libro y por tanto los datos asociados a estos.

- El resto de atributos como autor, temática, editorial pretenden ser utilizados como datos de estudios con los que hacer los modelos de predicción y de los que extraer las conclusiones asociadas a los objetivos iniciales del proyecto.

El dataset está formado por todos estos componentes ya que recoge la mayor información que nos proporciona la web siendo así una muestra de datos considerable y descriptiva de los libros más vendidos de este año.

Con esta muestra se pueden realizar otros estudios a parte del que se desarrolla, también se puede analizar que autor vende más, teniendo así varios libros en nuestro conjunto de datos.

Representación gráfica



Contenido

Para cada libro, el cual se corresponde con un conjunto de datos, se recogen las siguientes características:

a. **Título**: El título del libro. Breve texto que identifica o expone de **qué** trata..

b. **Autor**: Persona que ha realizado una obra científica, literaria o artística.

c. **Editorial**: Empresa que se dedica a editar libros y otras publicaciones por medio de la imprenta u otros procedimientos de reproducción.

d. **Url**: Enlace a la web del libro en el dominio de Todostuslibros

e. **Colección**: Conjunto ordenado de libros al que pertenece.

f. **Dimensiones**: Tamaño o extensión de una cosa, en una o varias magnitudes, por las cuales ocupa mayor o menor espacio.Tamaño del libro

g. **EAN**: El código EAN se trata de un tipo de código de barras de origen europeo , numeración que identifica el artículo.

h. **Encuadernación**: Parte exterior de un libro encuadernado, hecha de un material resistente, que cubre y protege el conjunto de las hojas cosidas, pegadas o anilladas en el lomo. En nuestro caso los libros son tapa blanda o bolsillo,o cartoné (tapa dura).

i. **Fecha de publicación**: Es la fecha en que se hace el documento patente a disposición del público.

j. **ISBN**: Sigla de la expresión inglesa *international standard book number*, 'número estándar internacional de libro', número de identificación internacional asignado a los libros.k. **Idioma de publicación**: Idioma del país donde se esté vendiendo.

l. **Idioma original**: Idioma original en que se escribió.

m. **Ilustrador**: Artista gráfico que se especializa en la mejora de la comunicación escrita, a través de representaciones visuales que se corresponden con el contenido.

n. **Nº de páginas**: nº de páginas que contiene cada libro.

o. **País de publicación**: País original donde se publicó el libro.

p. **Peso**: La magnitud del peso de un objeto (libro).

q. **Traductor**: Profesional que partiendo de un texto en un idioma lo convierte en un texto equivalente en otro idioma diferente. Persona encargada de traducir el libro.

r. **Materias**: Temas o géneros que se desarrollan en el libro.

Proceso de extracción

El proceso de extracción de datos se realiza gracias a la **exploración del subdominio mas_vendidos dentro de la web de todostuslibros**.

Obtención del *html* de 10 páginas con una lista de 10 libros cada una

Analizando este subdominio, podemos advertir que **las peticiones *GET* a esta *url* deben ir acompañadas de un parámetro *page***. Este parámetro sirve para mostrar los diferentes libros del *ranking* ya que este se encuentran separados en páginas (*pages*) de 10 libros, teniendo así 10 páginas de 10 libros cada una.

Es decir, **si queremos obtener la lista de 100 libros más vendidos debemos recorrer el subdominio de mas_vendidos sustituyendo el valor del parámetro *page***, que tomará valores entre **1 y 10** .

Un punto extra que hemos aplicado en estas peticiones *GET* gracias a que no tienen ningún tipo de dependencia entre ellas (no se necesita información de una para usar en otra) es **realizarlas de forma simultánea o en paralelo**. Algunos de los puntos a considerar en esta técnica:

- Se ha hecho usando la librería *threading* de *python*.
- Se tuvo en cuenta que el servidor pudiera denegar alguna de estas peticiones al realizarse varias en un corto espacio de tiempo, pero tras varias pruebas se validó que **10 peticiones casi inmediatas desde la misma ip no son problema para el servidor.**
- La ventaja es una **reducción** aproximada del **tiempo de ejecución** del programa en diez veces menos. Pasando en algunos test de unos 250 segundos a solo 25.

Tras aplicar este primer paso, **tenemos a nuestra disposición 10 páginas html cada una con información básica de 10 de los libros del *ranking*.**

En cada una de estas páginas vemos el autor y el título de cada uno de los 10 libros que guarda, se nos permite de esta forma obtenerlos para nuestro dataset. Aun así, primero debemos separar el contenido de cada libro. Para eso nos fijamos en que la información de cada libro viene agrupada en *divs* con la clase *book-details*. Obtenemos una lista con los textos *html* dentro de cada *div*.

Después, para seleccionar el título y el autor aplicamos selectores *css*. Para el título tomamos la primera instancia de un elemento *h2* y para el autor lo hacemos con la clase *author*.

Obtención de información específica de cada libro

Esta información básica como el título o autor **no es en nuestro caso del todo completa** ya que, **si nos quedamos solo con esto, perdemos atributos** tan importantes como las materias o temáticas del libro. **Estos atributos solo aparecen en la página específica para cada libro.**

Si repasamos una de las páginas con la información de 10 libros, podemos ver que **cada título de un libro tiene asociado un elemento html *a*. Este es un enlace a la página específica del libro**. Realizamos una petición *GET* para cada enlace que hay asociado a los títulos.

Analizamos ahora la página específica de un libro. **De esta página se obtiene información como materias, editorial, fecha de publicación, colección …** Para obtener esta información de forma adecuada observamos la estructura *html*.

La estructura *html* de estas páginas tiene la dificultad de que la información de cada campo no está debidamente identificada con un atributo *id* o de clase. En este caso la información se estructura dentro de un elemento de lista descriptiva *dl* de clase *row* hijo de otros elementos *div* con clases vinculadas a la librería de javascript bootstrap. Por tanto, para extraer en un primer momento el código donde se encuentra la información específica del libro se hace necesario aplicar selectores css más complejos.

No se puede usar solo como selector la clase row ya que es utilizada en más elementos que son ajenos a contener información del libro.

Una vez seleccionado todo el código abarcado por nuestro selector, el problema está en que el nombre de un atributo del libro, por ejemplo temática, se sitúa en un elemento *dt* y el

valor del mismo en un elemento *dd* a continuación; de esta forma se alterna nombre de atributo y valor hasta estar desplegada toda la información específica del libro.

Para superar esto **obtenemos dos listas, una con los nombres de los atributos** usando como selector *dt* **y otra con los valores de los atributos** usando como selector *dd*. **Después se recorren y se almacena la información en el dataset.**

Recordamos que se realiza este proceso para cada libro y finalmente obtenemos nuestro dataset de los 100 libros más vendidos en españa.

Como detalles finales decir que, **para cada petición**, se ha ejecutado una función que le otorga un **user-agent aleatorio de una lista previamente fijada**.

Agradecimientos

Tal como expone la web de todostuslibros en su apartado de “quienes somos”, **los datos de los libros más vendidos son proporcionados directamente por todas las librerías españolas asociadas a su proyecto. Por tanto, nuestro dataset es directamente obtenido de estas librerías.**

Se optó por este enfoque de recoger los libros más vendidos del año basándose en análisis similares, como son la lista de las canciones más reclamadas/sonadas de otros años [2], o el estudio de los libros más famosos/comprados de la historia. Estos estudios y artículos sirvieron de inspiración para el desarrollo de nuestro proyecto.

En nuestra misma línea de trabajo y tema encontramos algún dataset de años anteriores. **La empresa Statista** [3] **ya ha realizado algún dataset relacionado con los libros más vendidos**. De forma particular cuenta con el dataset de libros más vendidos en España de 2020.

También podemos valorar en este espacio a **otras empresas con sus propias listas de “best sellers”** basadas en sus propias ventas como son Fnac [4], El corte Inglés [5], Amazon [6], Grupo Planeta [7], Carrefour [8] …

Inspiración

Volviendo a centrarnos en **el dataset de Statista** [3], que es **el más representativo encontrado**. El **breve análisis** que incorpora centra su atención **sobre el autor que más vendió** aunque simplemente a título de “primero de la lista”, sin contar posibles agregaciones o un análisis más amplio. Como factor distintivo destacamos que **el dataset de Statista añade el número de ejemplares vendidos de cada libro** lo cual otorga mayor información respecto a la diferencia de ventas entre puestos. Si algo se le puede reprochar a este dataset es que **solo dispone los datos de un ranking de diez libros**, en comparación con nuestro dataset de cien posiciones.

Respondiendo a la pregunta de qué nos aportan estos datos. Podemos ver su importancia en entender un mercado como el del libro cuya facturación interior en España ronda los 2400 millones de euros. Gracias a estos datos y estudios similares como los mencionados podemos comprender al consumidor de este mercado, el lector. Estos datos nos permiten ver sus inquietudes, sus temas favoritos, sus autores consagrados… y así adaptarnos a él. Un claro ejemplo práctico es poder establecer mayor stock de libros de determinados autores, organizar encuentros con estos autores e incluso descubrir potenciales nuevos intereses de los lectores tanto en temas como en escritores.

Licencia

Para nuestro dataset optariamos por una **licencia de base de datos abiertas (Database released under Open Database License)** con el fin de que otros puedan obtener nuestros datos o contribuir en ellos. Para ello mantendremos nuestra base de datos abierta y sin ningún medio de pago con el fin de que nuestros datos se puedan usar como un componente más de un programa diseñado para la distribución comercial.

Dataset

Se ha publicado el DataSet en Zenodo.
El enlace de acceso es: https://zenodo.org/record/4664583

Bibliografía

[1] «Todos tus libros». https://www.todostuslibros.com/ (accedido abr. 06, 2021).
[2] «Las canciones más escuchadas de Spotify en 2020: ni Bad Bunny ni The Weeknd, la lista española la lidera el tema que mejor nos representa», *GQ España*. https://www.revistagq.com/noticias/articulo/canciones-mas-escuchadas-spotify-verano-2020 (accedido abr. 06, 2021).
[3] «Libros más vendidos España 2020», *Statista*. https://es.statista.com/estadisticas/808007/ranking-de-los-libros-mas-vendidos-espana/ (accedido abr. 06, 2021).
[4] «Libros más vendidos: Libros y ebooks ▷ FNAC, la mejor selección de Libros y ebooks». https://www.fnac.es/n710/Libros-mas-vendidos (accedido abr. 06, 2021).
[5] «Más vendidos - Libros - Librería · Libros · El Corte Inglés (4.892)». https://www.elcorteingles.es/libros-mas-vendidos/libros/ (accedido abr. 06, 2021).
[6] «Amazon.es Los más vendidos: Los productos más populares en Libros». https://www.amazon.es/gp/bestsellers/books (accedido abr. 06, 2021).
[7] «Libros más vendidos del 2021 | Planeta de Libros», *PlanetadeLibros*. https://www.planetadelibros.com/libros-mas-vendidos (accedido abr. 06, 2021).
[8] «Libros de lectura - Carrefour.es». https://www.carrefour.es/libros-de-lectura/cat9830092/c (accedido abr. 06, 2021).

Contribuciones	Firma
Investigación previa	Luis Miguel Moreno Lopez, Alejandra Cristina Marrero Suárez.
Redacción de las respuestas	Luis Miguel Moreno Lopez, Alejandra Cristina Marrero Suárez.
Desarrollo del código	Luis Miguel Moreno Lopez, Alejandra Cristina Marrero Suárez.