Dataset: Libros bestsellers en España

Contexto

Este trabajo se desarrolla en un supuesto contexto empresarial. Una editorial de renombre quiere conocer el comportamiento de sus potenciales clientes a la hora de comprar libros. Para ello, requiere información relativa a los libros más vendidos en una página web de venta de libros online que trabaja con gran cantidad de librerías en toda España.

El objetivo de la editorial es utilizar esta información para tomar decisiones a la hora de seleccionar qué libros editar y distribuir para maximizar sus beneficios.

Descripción

En esta actividad práctica se ha desarrollado un programa que mediante técnicas de Web Scraping, genera un conjunto de datos que contiene información sobre los libros más vendidos en España en el momento de la ejecución de este.

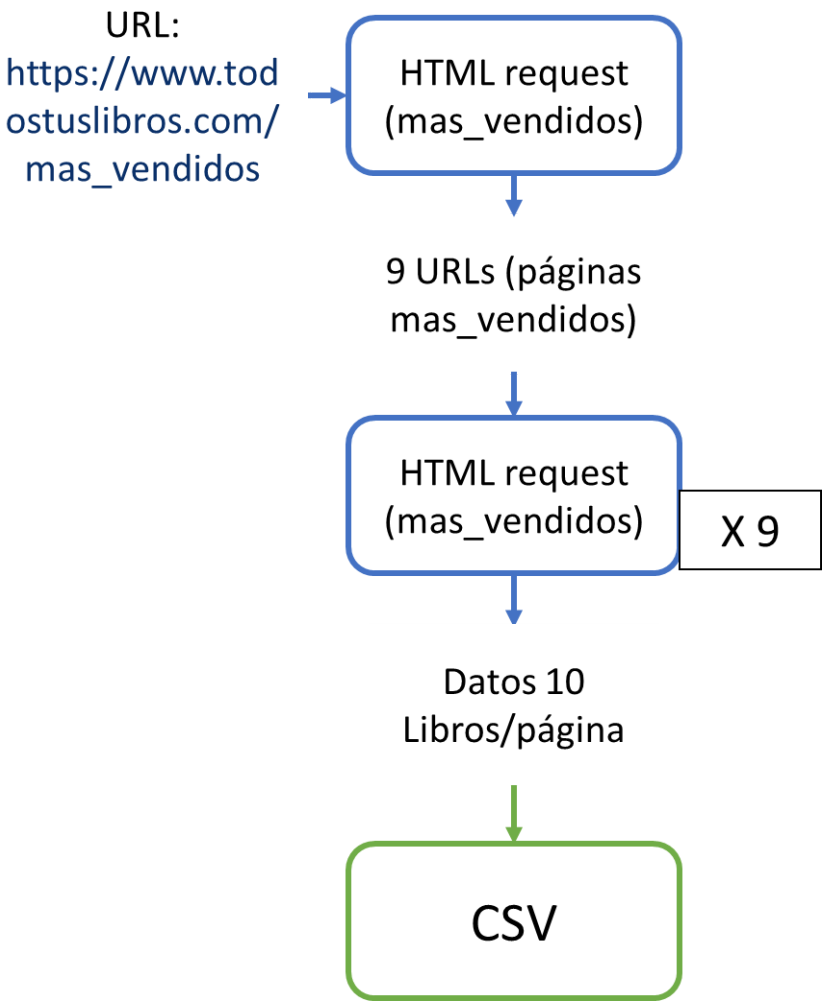
El programa consta de dos modalidades principales de ejecución, la primera nos permite extraer de forma rápida un conjunto de datos que contiene la información más relevante sobre los 100 libros más vendidos, como el título, el autor, la editorial y su precio.

La segunda modalidad nos permite, mediante una extracción más lenta, generar un conjunto de datos más completo con información detallada sobre los libros más vendidos. Algunos de sus parámetros serían el número de páginas o su peso.

Finalmente, se habilita una opción de ejecución que permite descargar las imágenes de las portadas de los libros a partir de la información de sus URLs almacenadas en el conjunto de datos previamente generado.

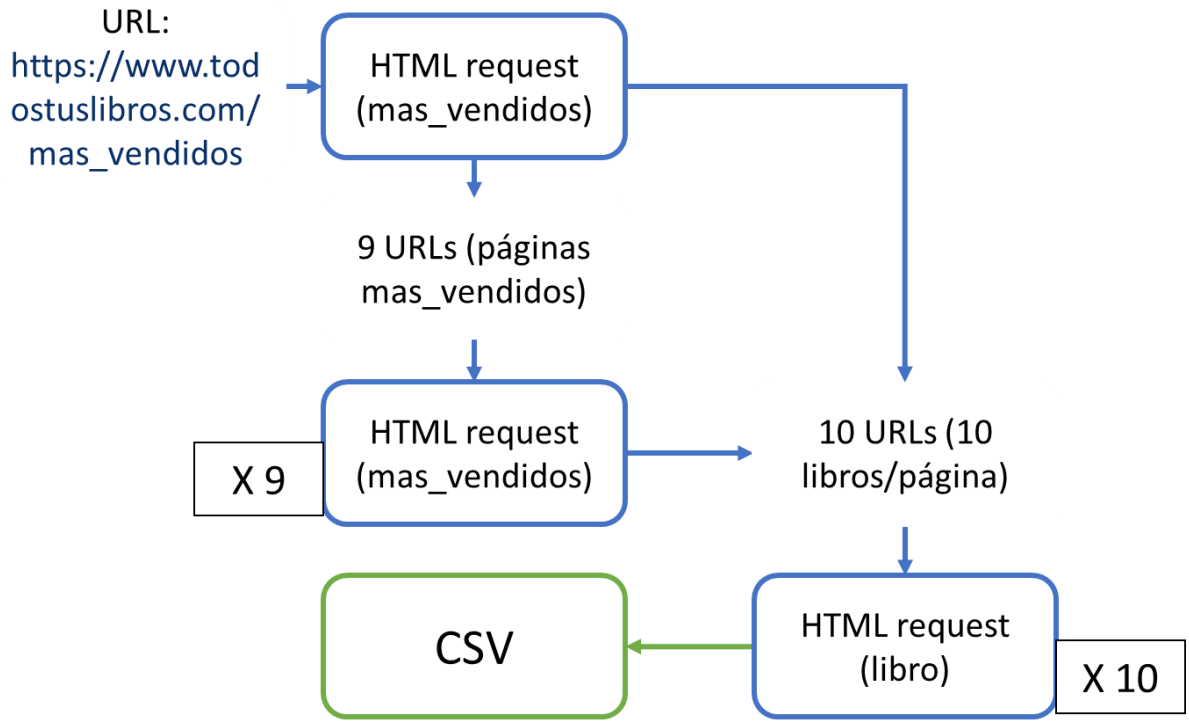
Representación gráfica

Fast_Bestsellers_April.csv



Id	Title	Subtitle	Author	Editorial	ISBN	Price (€)	Untaxed price (€)	Book cover
1	Operación Kazán	Premio Primavera	Vallés, Vicente	Espasa	978-84-670-6368-4	20.9	20.1	https://static.cegal.es/imagenes/m
2	El mentalista		Läckberg, Camilla	Editorial Planeta	978-84-08-25519-2	23.9	22.98	https://static.cegal.es/imagenes/m
3	El mapa de los ar		Kellen, Alice	Editorial Planeta	978-84-08-25595-6	17.9	17.21	https://static.cegal.es/imagenes/m
4	TOKYO REVENGE		WAKUI, KEN	NORMA EDITORIAL	978-84-679-4711-3	16	15.38	https://static.cegal.es/imagenes/m
5	El Libro Negro de		García Sáenz de I	Editorial Planeta	978-84-08-25285-6	20.9	20.1	https://static.cegal.es/imagenes/m
6	El castillo de Barb	Terra Alta III	Cercas, Javier	Tusquets Editores	978-84-1107-084-3	21.9	21.06	https://static.cegal.es/imagenes/m
7	Violeta		Allende, Isabel	PLAZA & JANES	978-84-01-02747-5	22.9	22.02	https://static.cegal.es/imagenes/m

Bestsellers_April.csv



Id	Title	Subtitle	Author	Matter	Editorial	Traductor	Collection	Binding	Country	Language of publication
1	Operación Kazán	Premio Prima	Vallés, Vicente	Ficción moderna	Espasa		ESPASA NARRATI	Cartoné	España	Castellano
2	El mentalista		Läckberg, Camilla	Obra de misterio	Editorial Pla	Conde Fisa	Planeta Internaci	Cartoné	España	Castellano
3	El mapa de los ar		Kellen, Alice	Ficción moderna	Editorial Pla		(Fuera de colecci	Tapa blan	España	Castellano
4	TOKYO REVENGE		WAKUI, KEN	Novelas gráficas:	NORMA EDI		TOKYO REVENGE	Libro en c	España	Castellano
5	El Libro Negro de		García Sáenz de I	Obra de misterio	Editorial Pla		Autores Españole	Cartoné	España	Castellano

Original language	ISBN	EAN	Dimension	Weight	Number of pages	Publication date	Price (€)	Untaxed price (€	Book cover
Castellano	978-84-670-6368	9788467063684	230 x 150 mm.	624 gramos	424	23-03-2022	20,90	20,10	https://static.ceg
Sueco	978-84-08-25519	9788408255192	230 x 150 mm.	922 gramos	720	16-03-2022	23,90	22,98	https://static.ceg
Castellano	978-84-08-25595	9788408255956	230 x 150 mm.	560 gramos	496	30-03-2022	17,90	17,21	https://static.ceg
Castellano	978-84-679-4711	9788467947113	210 x 140 mm.	200 gramos	374	25-03-2022	16,00	15,38	https://static.ceg
Castellano	978-84-08-25285	9788408252856	230 x 150 mm.	586 gramos	384	02-02-2022	20,90	20,10	https://static.ceg

Contenido

Los datasets contienen los siguientes campos:

Fast_Bestsellers_April.csv

id: Un número identificador único para cada libro. Su rango es del 1 al 100 (int).

title: Título del libro (string).

subtitle: Información ampliatoria relativa al libro como por ejemplo se ha recibido algún premio importante (string).

author: Autor del libro (string).

editorial: Editorial del libro (string).

ISBN: Número que identifica al libro inequívocamente con respecto al resto (International Standard Book Number) (string).

price (€): Precio del libro (num).

untaxed price (€): Precio sin IVA (num).

book cover: URL de la imagen de la portada del libro (string).

Bestsellers_April.csv

id: Un número identificador único para cada libro. Su rango es del 1 al 100 (int).

title: Título del libro (string).

subtitle: Información ampliatoria relativa al libro como por ejemplo se ha recibido algún premio importante (string).

author: Autor del libro (string).

matter: Temática sobre la que trata el libro (string).

editorial: Editorial del libro (string).

traductor: Persona que ha traducido el libro del idioma original al castellano (string).

collection: Colección a la que pertenece el libro dentro de la editorial (string).

binding: Material con el que están hechas las tapas (string).

country: País en el que está editado el libro (string).

lenguage of publication: Lenguaje del libro a la venta (string).

original lenguage: Lenguaje original del libro (string).

ISBN: Número que identifica al libro inequívocamente con respecto al resto (International Standard Book Number) (string).

EAN: Número del código de barras (European Article Number) (int).

dimension: Dimensiones del libro expresado en milímetros de alto por milímetros de ancho (string).

weight: Peso del libro en gramos (string).

number of pages: Número de páginas (int).

publication date: Fecha de publicación con el formato DD-MM-AAAA (date).

price (€): Precio del libro (num).

untaxed price (€): Precio sin IVA (num).

book cover: URL de la imagen de la portada del libro (string).

La duración de ambos datasets es de un mes, ya que mensualmente se actualiza la sección de libros más vendidos de la página web todostuslibros.

Para el almacenaje de los datos se ha procesado el formato en aquellos que son cadenas de caracteres y se han extraído los números de los precios.

Agradecimientos

El conjunto de datos ha sido extraído mediante el uso de las técnicas de Web Scraping presentadas en el temario de la asignatura de Tipología y ciclo de vida de los datos.

El conjunto de datos resultante, contiene información obtenida a partir de la página web https://www.todostuslibros.com/mas_vendidos. Se trata de una página web creada por la Confederación Española de Gremios y Asociaciones de Librerías (CEGAL) para facilitar información sobre libros que se comercializan en España o Latino América. Proporcionan todo tipo de información sobre los libros, permiten realizar búsquedas e indican en qué librerías se pueden encontrar o cuales son los libros actualmente más vendidos.

Se trata de una página con información sobre más de cuatro millones de libros y cada uno de ellos tiene una página HTML dedicada. De esta información, se podían esperar ciertos límites en cuanto al acceso a ciertos directorios del dominio, para evitar un abuso en número de peticiones HTML por parte de programas informáticos, que pudieran saturar el servidor. Por esta razón, nuestra primera tarea fue inspeccionar el archivo *robots.txt.*

En el archivo *robots.txt*, se restringía el acceso a los directorios de búsquedas y a los directorios relacionados con las librerías. Por esta razón, centramos el proyecto en torno al directorio de *Best Sellers*, es decir, los libros más vendidos, debido a que este no se encontraba restringido.

Para extraer el conjunto de datos *Bestsellers_April.csv*, que contiene el total de la información disponible extraíble sobre los libros más vendidos, se requería acceder individualmente a la URL de cada libro del listado. Para prevenir saturar el servidor con demasiadas peticiones HTML y hacer un uso responsable, introdujimos retrasos al realizar consecutivamente este tipo de peticiones.

En el momento de publicar en Zenodo los conjuntos de datos, optamos por eliminar los campos *Book cover* ya que estos contenían las URLs que apuntaban a las imágenes de las cubiertas de los libros, y consideramos que no era apropiado para un uso público y puede prevenir consecuencias legales.

Para realizar la actividad, no hemos partido de ningún proyecto previo de recolección de datos de libros, en vez de eso, hemos consultado guías sobre *Web Scraping* disponibles en Youtube [1][2], los ejemplos proporcionados por Ricardo Moya, en su guía de ejemplos del uso de *BeautifulSoup* con Python [3], y los ejemplos proporcionados como recursos de la práctica [4][5].

Para resolver dudas concretas acerca del código se consultó el foro StackOverFlow [6].

Inspiración

Gracias a este trabajo se podrán evaluar los comportamientos de los lectores de cara a conocer cuáles son sus gustos. De esta información se podrán extraer patrones de conducta con la finalidad de poder facilitar la toma de decisiones. Se podría llegar a crear un modelo predictivo con técnicas de Machine Learning.

La realización de este proyecto se ha llevado a cabo partiendo desde cero sin utilizar un estudio previo. Por ello, realizar una comparación con los ejemplos consultados no aporta información demasiado valiosa. Consultando el ejemplo de Teguayco Gutiérrez González [5], se ha estructurado el código creando una clase para cada extractor de contenido, *scraper,* y se dispone de un archivo *main* que ejecuta la función principal del programa. Como mejora respecto de los proyectos consultados, se han añadido tres opciones de ejecución, que permiten escoger si se desea realizar una extracción rápida de la información principal de los libros más vendidos, si se quiere realizar una extracción completa, y por ello más lenta, y finalmente, se ha añadido una opción que permite escoger si se desea descargar las imágenes de las cubiertas de los libros o no.

El código ha sido escrito siguiendo la guía de estilo de programación PEP8, que permite una mejor visualización y lectura del código. Además, se han añadido recursos para prevenir ser bloqueados por la página web, como añadir retrasos entre peticiones HTML consecutivas o modificar las cabeceras.

Otra dificultad afrontada durante el proyecto ha sido que la información que se ha tenido que extraer, a veces no tenía un formato único, si no que variaba en función del HTML que se estuviera consultando. Esto ha complicado en gran parte la programación del extractor

de contenido, pero también lo ha hecho más robusto ante posibles cambios del contenido de la web.

Licencia

Se selecciona una licencia Released Under CC BY-NC-SA 4.0 License para no permitir el uso comercial de los datasets. Se permite compartir, copiar y distribuir el material en cualquier medio o formato.

Siempre que se utilice esta licencia hay que proveer el nombre del creador y de los contribuyentes, un aviso de copyright, una aviso de licencia, un aviso de exención de responsabilidad, un link al material, un link a la licencia original CC BY-NC-SA 4.0 license – https://creativecommons.org/licenses/by-nc-sa/4.0/ y hay que indicar si se ha modificado el material y mantener un historial de modificaciones previas.

Toda modificación deberá ser distribuida con la misma licencia.

No se pueden añadir términos legales que restrinjan a otros hacer cosas que la licencia permita.

Se ha escogido esta licencia ya que vemos necesario restringir el uso comercial puesto que los datos se han obtenido con fines académicos.

Código

La actividad se ha realizado utilizando el lenguaje de programación Python 3.10.3, y está disponible en el siguiente enlace de Github: https://github.com/MartaCollPol/WebScraping

Dataset

Los dos conjuntos de datos generados, *Fast_Bestsellers_April.csv* y *Bestsellers_April.csv* han sido publicados en Zenodo y pueden ser encontrados en el siguiente enlace: https://zenodo.org/record/6435770#.YlRKvshBzIU

Cabe destacar que la columna *Book cover* del conjunto de datos, ha sido eliminada antes de su publicación debido a que contenía URLs que consideramos que no deben publicarse en un conjunto de datos de dominio público para evitar consecuencias legales.

Por las mismas razones, no se han publicado las imágenes de las cubiertas de los libros, las cuales han sido descargadas de las correspondientes URLs eliminadas. En el repositorio de Github proporcionado anteriormente, se puede encontrar un ejemplo de una descarga de estas imágenes.

Referencias

[1] NICOLAS MARIN TORRES. *Como hacer Web Scraping con Python y Beautiful Soup*. [Video en línea]. 2021 [consulta: 07 de abril de 2022]. Disponible en: https://www.youtube.com/watch?v=RjfqdJEwWyU&t=432s&ab_channel=NicolasMarinTorres

[2] PYTHON SIMPLIFIED. *Web Scraping with Beautiful Soup - Make Databases from Scratch.* [Video en línea]. 2020 [consulta: 08 de abril de 2022]. Disponible en: https://www.youtube.com/watch?v=ySNSY7iiBDY&ab_channel=PythonSimplified

[3] *Scraping en Python (BeautifulSoup), con ejemplos* [en línea] [consulta: 07 de abril de 2022]. Disponible en: https://jarroba.com/scraping-python-beautifulsoup-ejemplos/

[4] RAFAEL REYNA CAMONES. *Repositorio de GitHub foodPriceScraper* [en línea] [consulta: 07 de abril de 2022]. Disponible en: https://github.com/rafoelhonrado/foodPriceScraper

[5] TEGUAYCO GUTIÉRREZ GONZÁLEZ. *Repositorio de GitHub Web-scraping-aviation-accidents* [en línea] [consulta: 09 de abril de 2022]. Disponible en: https://github.com/tteguayco/Web-scraping-aviation-accidents

[6] *Stack Overflow* [en línea] [consulta: 07 de abril de 2022]. Disponible en: https://stackoverflow.com/