Práctica 1: ¿Cómo podemos capturar los datos de la web?

Trabajo realizado por Pablo Monforte Izquierdo

En esta práctica vamos a realizar un proyecto de web scraping, concretamente crearemos un dataset con información sobre libros y para conseguirlo scrapearemos la web de la casa del libro. A continuación se explicarán los detalles del proyecto.

Contexto

Los datos que vamos a scrapear en este proyecto son información sobre libros. El contexto de este proyecto responde a varios escenarios en los que podemos utilizar esta información.

El primer escenario sería suponer que tenemos una librería y queremos espiar a la competencia viendo que precios ponen, en que categorías clasifican los libros o incluso cuál es su catalogo de libros.

Otro escenario sería poder utilizar estos datos para realizar un estudio de mercado y ver como son los libros que se están escribiendo fijándonos en detalles como la categoría, el número de páginas, el tipo de tapa que utilizar si banda o dura…

Por último, esta información también podría ser utilizada para crear una pagina web que fuera un directorio de libros en el que estarían clasificados y las personas los pudieran encontrar con facilidad.

Por todos estos estos motivos he elegido https://www.casadellibro.com/ para obtener toda esta información, ya que tenemos libros clasificados con sus especificaciones y precios.

He de mencionar que para simplificar la tarea no scrapearemos todos los libros que hay en esta web porque el programa tardaría días en ejecutarse así que solamente scrapearemos las primeras 150 páginas de la sección política, esto supone intentar scrapear unos 3000 libros y el tiempo de ejecución del programa es de unos 30-40 minutos.

https://www.casadellibro.com/libros/ciencias-politicas-y-sociales/politica/105002000

Título

El título del dataset será “**Libros sección política casa del libro**”

Descripción de dataset

Libros sección política casa del libro

Este dataset contiene la información más relevante sobre algunos de los libros de la sección política de la casa del libro.

Representación gráfica

Esta imagen es ideal para representar el dataset ya que aparecen libros en la fachada de una casa y yo he extraído datos sobre libros de la casa del libro.

Contenido

A continuación se explican las variables que contiene el dataset.

- Title: titulo del libro
- Author: autor del libro
- Image: URL de la imagen de la portada
- Tags: Categorías en las que se clasifica el libro
- Sinopsis: sinopsis del libro
- Price: precio del libro
- Real_price: precio actual del libro que a veces cuenta con descuentos
- N_pages: número de páginas del libro
- Editorial: editorial del libro
- Lang: idioma del libro
- Encuadernación: tipo de encuadernación del libro
- ISBN: ISBN del libro
- Year: año en el que se publicó el libro
- Date: fecha en la que se publicó el libro

Propietario

El propietario del dataset es Pablo Monforte Izquierdo y los datos que se han extraído son datos que cualquier persona puede obtener solamente que se utilizan bots para automatizar la tarea.

Inspiración

La inspiración de este dataset viene tras pensar en algún elemento que cumpliera los requisitos de la práctica. Al ser aficionado a los libros pensé que era una buena opción crear un dataset con la información de los libros.

Licencia

La licencia es Creative Commons Zero (CC0), se puede utilizar libremente este dataset con sin ningún tipo de restricción. He elegido este tipo de licencia para que todas las personas puedan utilizar el dataset sin restricciones para los usos que consideren oportunos.

Código

Se encuentra en la carpeta libros y será explicado en detalle en el video.

Dataset

El dataset lo podemos encontrar en el siguiente enlace de Zenodo

https://zenodo.org/record/7307801#.Y2ubh4LMJH0

Video

El video se encuentra en el siguiente enlace:

https://drive.google.com/file/d/1gCGpSrtuTwmp7eocs4O--gOkrwlr_zM7/view?usp=sharing