Reinaldo Quintero
Irene Díaz

# PRAC 1: Web Scraping

## Contexto:

En este proyecto se realizará un Web Scrapping a la página de venta online PcComponentes[1]. El propósito del mismo es el de recopilar información sobre los portátiles que se venden en dicho portal con la intención de crear una base de datos. Se recopilará información sobre las características de los portátiles, así como su precio. Antes de comenzar este proyecto, estuvimos buscando y comprobamos que no existía ninguna API de PCComponentes que pudiera hacer frente a los objetivos que nos planteamos con este scrapeo. Para realizar este proyecto se ha hecho uso de librerías como BeautifulSoup con el propósito de parsear el HTML de forma manual. Hemos modificado también el user-agent para que la página no nos bloquee las peticiones al saber que se trata de un acto automático el que está recopilando la información y nos hemos hecho pasar por usuarios. Se ha escogido un user agent de entre esta lista[2]. Se ha incluído un sleep para que el scraping no se realice la recopilación de información muy rápido y no saturar al servidor. Se pasa una lista con diferentes urls para hacer búsquedas en diferentes categorías. La página web no cuenta con paginación per se.

## Título:

Dataset PCComponenetes: análisis sobre los portátiles

## Descripción del dataset:

El dataset contiene datos que hemos considerado relevantes sobre los portátiles que se ofertan en la tienda. Se recopilan diferentes características del ordenador en sí, así como su precio y el número de opiniones sobre ese producto.

## Representación gráfica:

[1] https://www.pccomponentes.com/
[2] http://www.useragentstring.com/pages/useragentstring.php?name=Chrome

## Contenido dataset:

- Product: donde se especifica el nombre del producto y todas sus características:
  - Marca: la marca del producto.
  - Procesador: marca y generación del procesador.
  - RAM: RAM con la que cuenta el producto.
  - Disco duro: capacidad y tipo de disco duro.
  - Tamaño: tamaño del portátil en pulgadas.
- Price: precio del producto en euros.
- Availability: cuándo se puede recibir el producto.
- Number_of_opinions: opiniones totales que ha recibido el producto.

Todas estas variables se han obtenido analizando los elementos de la página web del producto. En cuando a la información contenida en Product, se procedería después a una limpieza de datos para dividir las diferentes características. El periodo de tiempo de los datos es el de los artículos disponibles durante el mes de noviembre.

## Agradecimientos:

Los datos han sido recolectados de la misma web de PCComponentes.

## Inspiración:

Se ha optado por este proyecto ya que la oferta de artículos electrónicos, en este caso más concreto, portátiles, es muy extensa y puede resultar complicado el comparar todo lo que tenemos a nuestra disposición en el mercado puesto que se tienen que valorar los muchos componentes que forman estos productos. Se busca entonces crear una base de datos que permita dar una perspectiva a qué elementos pueden ser los que más influyan sobre el precio de los portátiles que se encuentran ahora en el mercado.

Este tipo de base de datos puede resultar de utilidad para eventos como el Black Friday, Ciberlunes o eventos en los que no se contempla el IVA, ya que permitirá hacer un seguimiento de los precios y valorar la compra.

## Licencia:

La licencia por la que hemos optado para proteger nuestro dataset es CC BY-NA-SA 4.0 (Attribution-Non Commercial-ShareAlike 4.0 International). Los motivos son los siguientes:

- Bajo esta licencia, cualquier otra persona interesada en hacer uso de ella podrá hacerlos libremente. Pudiendo además compartir la misma o transformarla.
- Aunque hagan uso de ella, se nos tiene que dar el crédito correspondiente de la misma. De este modo, nuestro trabajo a la hora de recopilar los datos no se perderá.
- No se puede hacer uso comercial de la misma.
- Si bien el usuario tiene capacidad para cambiar, transformar o crear algo a partir de este juego de datos, se nos tiene que comunicar.

En definitiva, buscamos una licencia que permita a otros usuarios sacarle partido a nuestro juego de datos, pero que no saque rédito económico de ello ni se pierda el hecho de que nosotros somos los autores del mismo.

## Código:

El código se encuentra publicado en el repositorio de GitHub[3].

## Archivo CSV

Se ha subido el archivo .csv que contiene la base de datos a Zenodo[4].

| Contribuciones | firmar |
|---|---|
| Investigación previa | Reinaldo Quintero<br>Irene Díaz |
| Redacción de las respuestas | Reinaldo Quintero<br>Irene Díaz |
| Desarrollo del código | Reinaldo Quintero<br>Irene Díaz |

[3] https://github.com/reinaldoq/Hardware-store-scraper/blob/main/scraper.py
[4] https://zenodo.org/record/4263182