2020

PRÁCTICA 1: Web Scraping

MANUEL TABERNER LLORCA Y ANDRÉS PÉREZ SANTANO

TIPOLOGÍA Y CICLO DE LA VIDA DE LOS DATOS

UNIVERSITAT OBERTA DE CATALUNYA | Máster en Ciencia de Datos

Tabla de contenidos

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

En el documento de esta práctica nos solicitan hacer un scraper de un sitio web. Hemos decidido realizar un scraper de una web muy utilizada para comprar productos relacionados con la tecnología PcComponentes. El objetivo que tenemos para realizar el scraper de esta web, es obtener los precios de diferentes productos de componentes informáticos ordenados por categorías y por marcas como placas base, discos duros…

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

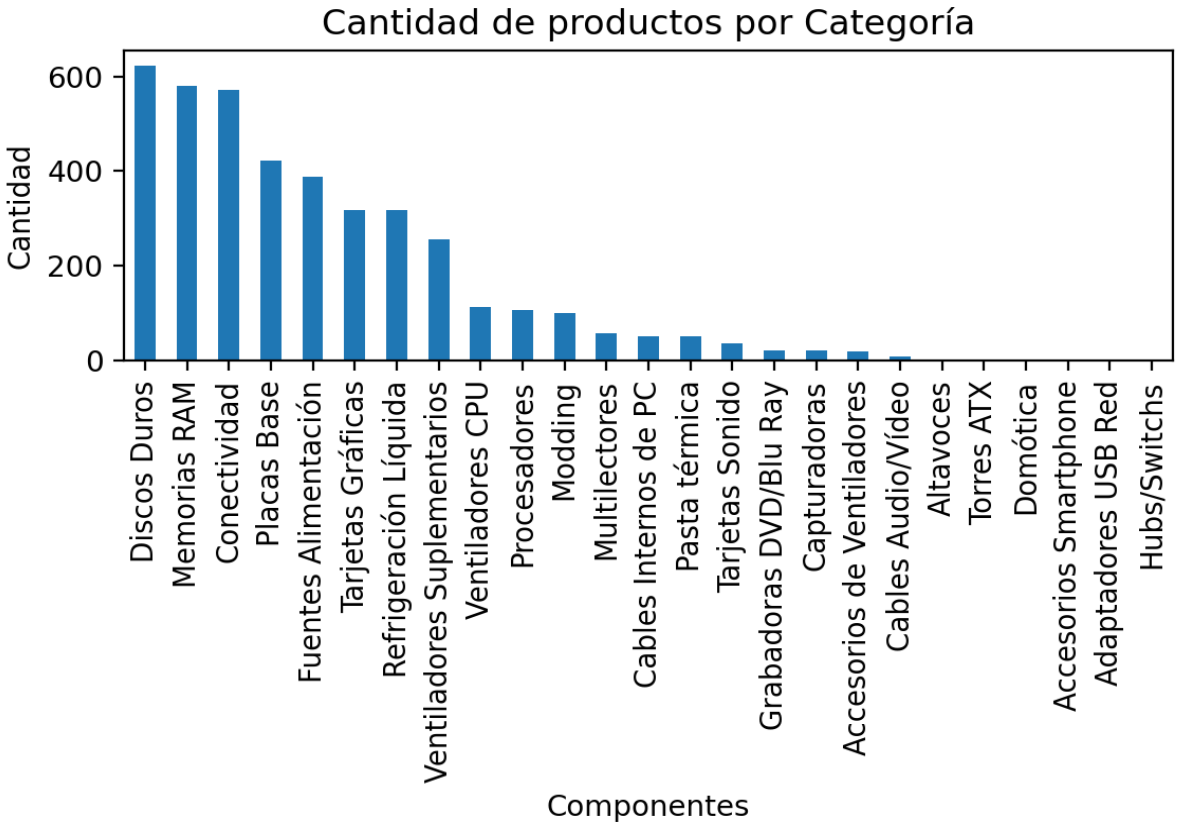
El título que hemos elegido para nuestro dataset es: COMPONENTES PCCOMPONENTES.

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

El dataset obtenido refleja la información de todos los componentes informáticos disponibles en la web de PcComponentes. Los dispositivos reflejan la marca, el precio y la categoría.

4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente

Hacer histograma de los diferentes productos.



5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

Los campos que forman el dataset son precio, nombre, marca, y clasificación en la web

En el dataset presentado se pueden observar los campos:

- **Nombre:** Nombre recoge el nombre del dispositivo, junto con su modelo (AMD Ryzen 7 3700X 3.6GHz BOX).
- **Precio:** incluye el precio del dispositivo.
- **Marca:** indica la marca del dispositivo.
- **Categoría:** indica la sección en la que podemos encontrar el dispositivo.
-

Los datos se han recogido realizando un scraper de la web que ya hemos mencionado anteriormente. Como se puede ver con más detalle en el script adjunto de GitHub, creamos listas vacías con los campos que queremos guardar (los mencionados anteriormente), creamos una lista con todas las categorías que tenemos en la web. De esta manera realizamos un bucle que recorra todas estas categorías y volcamos los datos que nos interesan en nuestro dataset.

Los datos se han recogido en el momento por lo que hemos realizado un snapshot.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Agradecemos a PCComponentes por poder realizar el scraper. El propietario de los datos es PcComponentes.

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

Este conjunto de datos nos ha parecido muy interesante para montar una tienda de dispositivos informáticos. Con este dataset obtenido, podemos comparar precios de esta página web para poder adecuar los de nuestra tienda.

El dataset presentado, nos puede responder a las siguientes preguntas:

- ¿De qué marca hay más dispositivos?
- ¿Cuál es la placa base más cara/barata en la tienda?
- ¿Cuántos marcas discos duros diferentes se pueden observar?
- ¿Comprobar evoluciones de precio si vamos realizando *snapshots* en diferentes días?

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

La licencia escogida es **Released Under CC BY-NC-SA 4.0** License puesto que es la que mejor se adapta a nuestro dataset.

Podemos copiar y distribuir nuestro material en cualquier medio, podemos utilizar los datos para realizar cualquier otro material en base a esos datos, incluso comercialmente. Podemos observar en el siguiente link la licencia completa.

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

https://github.com/manutaberner/scrapper-pccomponentes

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

https://zenodo.org/record/4133874

DOI: 10.5281/zenodo.4133874

11. Contribuciones

Contribuciones	Firma
Investigación Previa	MTL, APS
Redacción de las respuestas	MTL, APS
Desarrollo del código	MTL, APS