

PRAC1
WEB SCRAPING

Tutora: Meritxell Figueres Boquera

Profesor: José Moreira Sánchez

Alumno: Javier López Calderón y José Maria Cano Hernández

Asignatura: Tipología y ciclo de vida de los datos

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

El contexto en el que se realiza la recolección de datos es bajo la idea de construir un agregador o comparador de productos vinculados al PC, tomando como fuente de datos la información contenida en las tiendas web de dos proveedores conocidos y relevantes de dichos artículos a nivel nacional.

El planteamiento es extensible a tantos proveedores de productos de electrónica como se considere.

Al fin y al cabo, se pretende recolectar información de distintos sitios web que, no solo proporcione información por sí mismos, sino que en última instancia permite comparar entre los distintos proveedores para un mismo producto a nivel de precio.

Debido a lo anterior, se han seleccionado unas cuantas marcas de componentes de ordenador para obtener la información del dataset a elaborar.

En primer lugar, PcComponentes, un proveedor de componentes de ordenador conocido a nivel nacional e internacional cuya estrategia radica en un fuerte posicionamiento online.

En segundo lugar, se ha optado por PcBox, un proveedor de componentes de ordenador conocido a nivel nacional cuya estrategia es contraria al anterior, un fuerte posicionamiento físico pero discreto a nivel online.

Finalmente, destacar que para el tipo de información que se desea obtener, las tiendas web de venta de componentes de ordenador son el target adecuado para realizar las tareas de scraping y obtener información valiosa relacionada con el mercado actual de informática.

2.-Definir un título para el dataset.

El título del dataset elaborado es: Colección de componentes de pc.

Como se ha mencionado en el punto anterior, se trata de recolectar información de los productos o componentes relacionados con la construcción de un PC, tomando como fuente de datos diferentes proveedores de los mismos.

3.- Descripción del dataset.

Para la construcción del dataset se ha realizado un análisis previo consistente en la identificación de las categorías básicas que el proceso de web scraping debería cubrir para ser capaces de recopilar los productos objetivo.

Desde este punto de vista, el dataset tiene los siguientes atributos o grupos de atributos que consideramos relevantes:

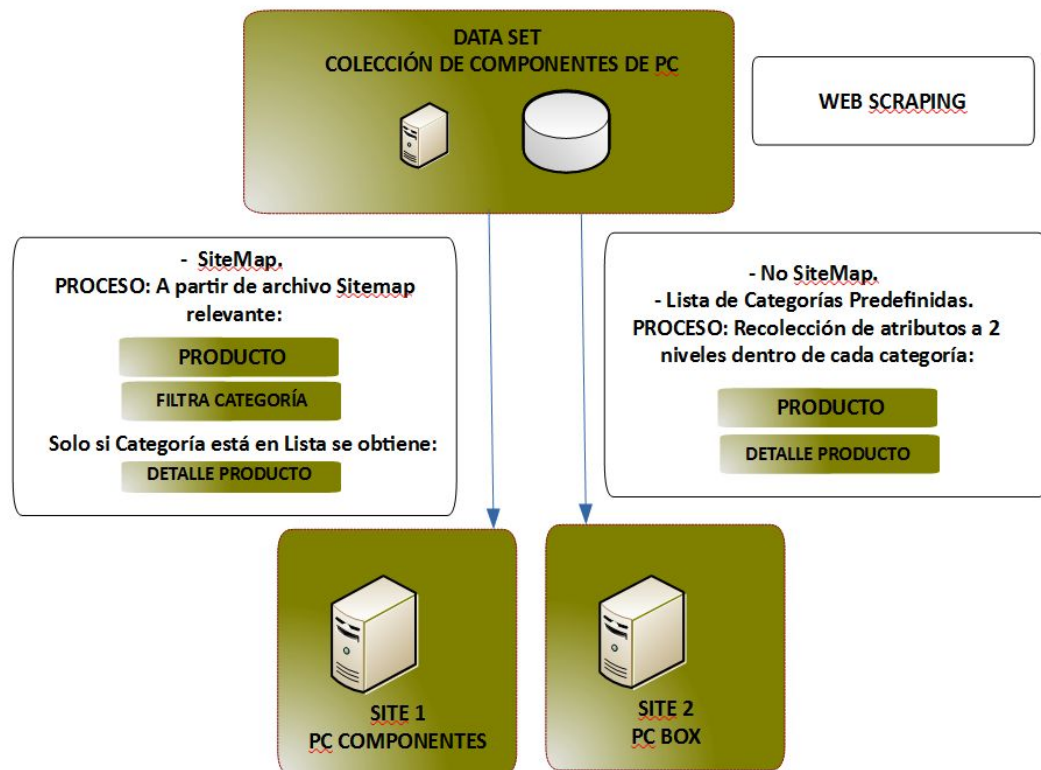
Atributo	Descripción
Timestamp	Indica la fecha en la que se ha llevado a cabo la recolección
Categoría	Grupo de producto al que pertenece el componente. Se han seleccionado los siguientes: <ul style="list-style-type: none">·Placas Base·Procesadores·Memorias RAM·Discos Duros·Torres·Tarjetas Gráficas·Tarjetas de Sonido·Fuentes de Alimentación
Nombre	Nombre del producto/componente
Precio	Precio del producto para esa tienda

La idea inicial de dataset es conseguir información de componentes de ordenador para posteriormente ser analizados; por ello, los elementos principales a recopilar han sido la categoría del producto, el nombre del producto, su precio y una marca de tiempo para poder establecer el precio que tenía un determinado producto en una determinada fecha.

Esta información temporal permite hacer un análisis del valor de un producto, encontrar tendencias sobre las diversas categorías y, por supuesto, observar en el tiempo cómo ha ido fluctuando el precio de cada producto.

En el dataset elaborado existe más información a parte de la mencionada anteriormente, como por ejemplo, las valoraciones que los clientes han realizado de cada producto, las reviews que han tenido e incluso un enlace web a una fotografía del artículo. Esta información permite otros análisis más detallados, como por ejemplo, cómo perciben los clientes el precio de un determinado producto o categoría de productos.

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



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

Los campos incluidos en el dataset son:

Campo	Explicación
timestamp	Marca de tiempo que indica cuando se ha obtenido.
company_name	Empresa de la que proviene el dato
name	Nombre del producto
brand_name	Nombre de la marca del producto
category	Categoría o Sección a la que pertenece el producto

product_number	Código que identifica el producto en la tienda
price	Precio del producto
score	Puntuación que tiene el producto
image_url	Dirección donde se encuentra la imagen
image_url_dataset	Dirección donde se encuentra la imagen en local
reviews	Número de opiniones que tiene el producto

****NOTA**:** el campo image_url_dataset se ha eliminado del documento subido a Zenodo debido a que carece de sentido añadir información relativa a una carpeta que no se puede subir a dicha página ni siquiera como fichero adjunto.

La información para el dataset se ha recolectado desde dos fuentes diferentes, por un lado la tienda de PcComponentes y por otro lado la tienda PcBox.

El proceso de recolección de datos para el caso de PcComponentes ha sido:

En primer lugar, realizar un análisis del fichero “robots.txt” que presenta la página, en él se muestran todas las zonas de la web sobre las que no se permite obtener información y aquellas donde sí está permitido, además de la frecuencia de actualización de cada sección.

Dentro del fichero “robots.txt” existen enlaces a diversos sitemaps con información relativa a las categorías de la tienda, los productos, las secciones y otras informaciones. Para el desarrollo de esta práctica se ha empleado uno de los sitemaps que hace referencia a todos los productos presentes en la tienda online.

En dicho sitemap, “https://www.pccomponentes.com/sitemap_articles_components.xml” puede observarse que contiene un enlace a una página de detalle de todos y cada uno de los productos que se encuentran en venta en estos momentos en la tienda y su frecuencia de actualización, informando en la mayoría de casos que puede cambiar a diario.

Se ha procedido a obtener todos y cada uno de estos enlaces empleando para ello la herramienta “Selenium” para abrir un navegador donde ir cargando los enlaces y “Beautifulsoup4” para recorrer la estructura de cada página y obtener la información de cada uno de los campos anteriormente descritos del dataset.

Debido a que en el sitemap aparecen más productos de los relevantes para elaborar este dataset, se ha realizado un proceso de filtrado para restringir los productos obtenidos a solo aquellas categorías que hemos pensado que eran relevantes para la construcción y elaboración de un ordenador.

Una vez se han filtrado todos los productos que no son relevantes, se ha procedido a descargar la imagen principal de cada componente y se ha almacenado como información adjunta.

Finalmente, se ha procedido a almacenar toda la información obtenida en un documento CSV con la estructura de campos antes mencionada.

El proceso de recolección de datos para el caso de PcBox ha sido:

En lo que respecta al proceso de recolección, se parte de una lista de categorías, que llevan asociada una o más páginas del Web site sobre las que se efectuará el Web Scrapping.

A partir de aquí, se extrae la lista de productos bajo dicha categoría y, para cada producto, es necesario acceder a la página de producto detallado, en la que se ha recopilado la Marca (Brand) y el Número de Producto (Product_Number) ya que no están en la página anteriormente mencionada.

Adicionalmente, se obtiene la URL de cada imagen de producto y se almacena bajo la ruta `./imagen/<categoria>/` manteniendo el nombre del fichero.

Otros condicionantes que se han tenido en cuenta u otros aspectos relevantes son:

Se ha optado por un fichero de properties para incluir la lista de categorías y url a incluir en el proceso, junto con otros parámetros de configuración, de tal forma que simplifique su invocación por usuario final.

Gestión de intervalos entre peticiones para no saturar la tienda online, se ha definido un intervalo mínimo entre peticiones (*minInterval*) que se establece por configuración en 2 segundos.

Además, un valor calculado variable de espaciado en intervalo entre peticiones (*calcIntervalDelay*) que se añade al min mencionado anteriormente que ayuda a no degradar el comportamiento del site en caso de un aumento en los tiempos de respuesta del mismo.

Por otro lado, también se ha realizado la gestión de timeouts, que se han producido ocasionalmente durante alguna prueba, en la que el proceso se para durante unos minutos para proseguir posteriormente, controlando el error.

Paginación en las páginas HTML de tal modo que la lista de productos se obtiene a partir de extraer todas y cada una de las páginas. Realmente la lista de categorías que conforman el Dataset objetivo no presentan este tipo de paginación, pero el código estaría preparado para hacer Web Scrapping en las mismas (siendo probado en otras páginas con paginación de este Web site).

6.- Agradecimientos.

Los datos han sido obtenidos desde las páginas web de las tiendas online de PcComponentes y PcBox.

Asimismo se han solicitado vía correos electrónicos con ambas empresas su permiso explícito para poder realizar el trabajo y obtener información de sus respectivas tiendas. Además, durante todo el proceso de scraping se ha tenido en cuenta la afluencia de peticiones que se generaban y se han aplicado medios para distribuir en el tiempo todas estas peticiones con la finalidad de causar el menor impacto para estas páginas y poder elaborar los dataset.

En cuanto a investigaciones previas, se valoraron otras tiendas donde venden componentes de ordenador:

Mediamarkt fue una de las tiendas analizadas, se analizó su “robots.txt” para conocer las partes permitidas y privadas de la página, cuenta con un sitemap completo, similar al de PcComponentes.

Otros Sites de otros dominios que también se valoraron en el proceso, revisando la existencia de “robots.txt” y sitemaps, API así como estructura de las páginas web fueron las siguientes:

- Wallapop/Zalando: Productos de segunda mano
- Hanker Sport: Web de ropa deportiva
- Zillow: Web site para alquiler y venta de propiedades inmobiliarias en EEUU.
- Idealista: Web similar a Zillow, alquiler y venta de propiedades pero de España.

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

La idea inicial era realizar una comparativa de componentes de ordenador entre diversas tiendas, de esta forma poder conocer más información sobre precios, valoraciones y opiniones.

A modo de ejemplo, esto es lo que podemos observar en DataSet para un Product Number determinado:

timestamp	company_name	name	brand_name	category	product_number	price
1604577762	pccomponentes	Intel Core i3-10320 3.8 GHz	Intel	Procesadores	BX8070110320	159.9
1604757571	pccomponentes	Intel Core i3-10320 3.8 GHz	Intel	Procesadores	BX8070110320	159.9
1604750058	pcbox	PROCESADOR INTEL CORE I3-10320 3.8 GHZ SK1200 8MB	INTEL	Procesadores	BX8070110320	181.96
1604705335	pcbox	PROCESADOR INTEL CORE I3-10320 3.8 GHZ SK1200 8MB	INTEL	Procesadores	BX8070110320	181.96

Sin embargo, con la evolución del proyecto nos hemos dado cuenta de que obteniendo los datos con cierta periodicidad (por ejemplo, diariamente o semanalmente) se podrían obtener información sobre tendencias de los productos, como los precios, predecir el valor de venta

de un producto en función de sus datos o incluso observar la recepción en base a las opiniones y reviews.

Las preguntas que se pretenden resolver con el dataset son:

¿Dónde se puede obtener más económicamente un determinado producto? ¿Qué otros productos venden a parte de componentes de ordenador? ¿Cómo tienen clasificados los productos? ¿Cómo son las valoraciones de los compradores? ¿Se especializan en alguna marca de componentes? ¿Cómo ha ido evolucionando el precio de un determinado producto en el tiempo? ¿Qué precio están dispuestos a pagar los clientes por un determinado producto de una categoría?

8.- Licencia

La licencia para el dataset que se ha seleccionado ha sido “Release Under CC BY-NC-SA 4.0 License” y el motivo de su elección han sido los tratos que se han realizado con las dos empresas participantes del dataset.

En las comunicaciones con PcComponentes y PCBox, hemos solicitado permiso para poder utilizar los datos dentro del ámbito académico y sin ánimo de lucro ni ninguna finalidad comercial.

9.- Código

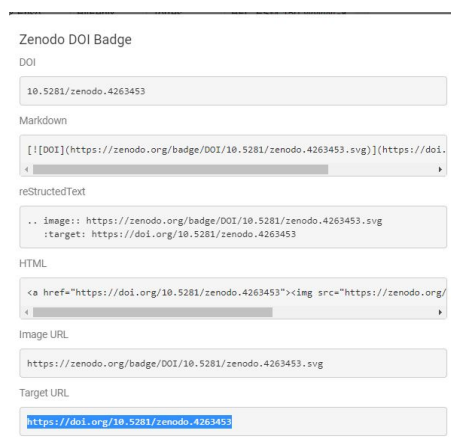
A continuación se adjuntan los enlaces a los repositorios de código del proyecto de cada estudiante.

Github Javier López: https://github.com/jlcldev/components_pc_scrap

Github José María Cano: https://github.com/jcanoh/ComponentesPC_Scraping

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

DOI: 10.5281/zenodo.4263453



The image shows a web interface for generating a Zenodo DOI badge. It contains several input fields and checkboxes:

- Zenodo DOI Badge**: The main title of the section.
- DOI**: A text input field containing the DOI `10.5281/zenodo.4263453`.
- Markdown**: A text area containing the markdown code `[](https://doi.org/10.5281/zenodo.4263453)`.
- reStructuredText**: A text area containing the reStructuredText code `.. image:: https://zenodo.org/badge/DOI/10.5281/zenodo.4263453.svg :target: https://doi.org/10.5281/zenodo.4263453`.
- HTML**: A text area containing the HTML code ``.
- Image URL**: A text input field containing the URL `https://zenodo.org/badge/DOI/10.5281/zenodo.4263453.svg`.
- Target URL**: A text input field containing the URL `https://doi.org/10.5281/zenodo.4263453`.

URL DATASET: <https://zenodo.org/record/4263453>

November 8, 2020 Dataset Open Access

ComponentesPC_Scraper

Javier Lopez Calderon; Jose M Cano Hernandez

This Dataset is the result of applying Web Scrapping techniques to a couple of Web sites (PC Componentes <https://www.pccomponentes.com/> & PC Box(<https://www.pcbox.com/>) in order to collect information regarding a set of basic categories of PC Parts and Computer Components.

This project has been conducted as a WEB SCRAPING practical assignment under the scope of "Data Tipology" subject which is part of Data Science Master in "Universitat Oberta de Catalunya (UOC)" in Spain.

timestamp	company_name	name	brand_name	category	product_number
1604570838	pccomponentes	Gigabyte SSD M.2 512GB 2280 PCIe x2 NVMe	Gigabyte	Discos Duros	GP-GSM2NE8512GNTD
1604570839	pccomponentes	BitFenix Enso Cristal Templado USB 3.0 RGB Blanca	BitFenix	Torres	BFC-ENS-150-WWWGK
1604570846	pccomponentes	PNY XLR8 CS3030 250GB M.2 3D TLC NVMe PCI-Express	PNY	Discos Duros	M280CS3030-250-RB
1604570849	pccomponentes	PNY XLR8 CS3030 500GB M.2 3D TLC NVMe	PNY	Discos Duros	M280CS3030-500-RB

Files (1.6 MB)

Name	Size	Preview	Download
ComponentesPC_Scraper_DataSet.csv	1.6 MB		

md5:57e177620dcee7cc3bd8554fb455762e

Descripción aportada en la web de Zenodo:

ESP-

Este Dataset es el resultado de aplicar técnicas de Web Scrapping a un par de sitios Web (PC Componentes <https://www.pccomponentes.com/> & PC Box (<https://www.pcbox.com/>) con el fin de recopilar información sobre un conjunto de categorías básicas de piezas y componentes informáticos.

Este proyecto se ha realizado como un trabajo práctico WEB SCRAPING en el ámbito de la asignatura "Tipología de datos" que forma parte del Máster en Ciencia de Datos de la "Universitat Oberta de Catalunya (UOC)" en España.

ENGLISH-

This Dataset is the result of applying Web Scrapping techniques to a couple of Web sites (PC Componentes <https://www.pccomponentes.com/> & PC Box(<https://www.pcbox.com/>) in order to collect information regarding a set of basic categories of PC Parts and Computer Components.

This project has been conducted as a WEB SCRAPING practical assignment under the scope of "Data Tipology" subject which is part of Data Science Master in "Universitat Oberta de Catalunya (UOC)" in Spain.

CONTRIBUCIONES

Contribuciones	Firma
Investigación Previa	Javier López Calderón, José M Cano Hernández
Redacción de las respuestas	Javier López Calderón, José M Cano Hernández
Desarrollo de Código	Javier López Calderón, José M Cano Hernández