PRA1: Web Scraping

M2.851 - Tipología y ciclo de vida de los datos

Víctor Blanes Martín
Carlos Allo Latorre

01/04/2021

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

La recolección de datos se ha basado en búsquedas de productos electrónicos en dos sitios web (Mediamarkt y Dominiovirtual.es). Se han seleccionado estos dos sitios web ya que ambos se dedican a la venta de este tipo de artículos y presentan tecnologías distintas en sus servicios web, cosa que permitirá hacer uso de distintas metodologías y técnicas de web scraping.

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

Dado que el programa capturará productos de tres categorías de electrónica (portátiles, monitores y tablets) se presentarán 3 datasets diferentes.

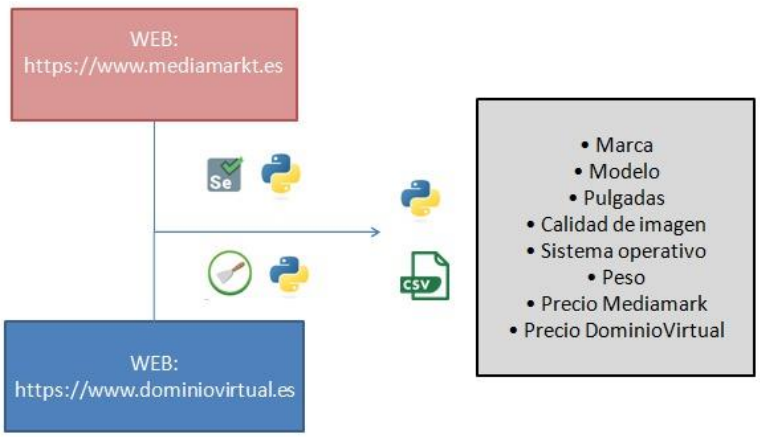
- Especificaciones y precios de dos vendedores de portátiles
- Especificaciones y precios de dos vendedores de monitores
- Especificaciones y precios de dos vendedores de tablets

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 contendrá los precios de los productos de cada categoría en cada una de las paginas web, así como las siguientes especificaciones:

- Marca
- Modelo
- Pulgadas
- Calidad de imagen
- Sistema operativo
- Peso

4. Representación gráfica. Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido



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

Cada uno de los datasets incluirá los siguientes campos:

- Marca: Fabricante del producto
- Modelo: Identificador comercial del producto
- Pulgadas: Tamaño de pantalla en pulgadas del producto
- Calidad de imagen: Resolución de la pantalla
- Sistema operativo: Sistema operativo instalado de fábrica en el producto
- Peso: Peso del producto
- Precio Mediamarkt: Precio del producto en Mediamarkt.es
- Precio Dominiovirtual.es: Precio del producto en dominiovirtual.es

Los datos han sido recogidos, y por lo tanto son válidos, el día XX de abril.

Las tecnologías que se han utilizado para realizar el web scraping han sido distintas para cada uno de los sitios web. Dado que uno de los objetivos planteados en la realización de la práctica ha sido el de estudiar distintas alternativas para la recogida de datos, se escogió en origen una web que tenía carga dinámica como Mediamarkt y otra con carga estática y paginación como dominiovirtual.es.

Para realizar el web scraping de Mediamarkt y poder lidiar con las cargas dinámicas de la página, se ha hecho uso de la librería Selenium y, para hacer la extracción de datos de dominiovirtual.es se ha hecho uso de Scrapy, en ambos casos haciendo uso de Python como lenguaje de programación.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.

Los sitios web de los propietarios de los datos que han sido recogidos en esta práctica son los siguientes:

- https://www.mediamarkt.es/es/
- https://www.dominiovirtual.es/

En ambos casos fueron elegidos por no disponer de una API pública de la cual extraer estos datos, si bien se identificó que Mediamarkt dispone de una API privada al parecer expuesta para sus empleados o colaboradores directos:

- https://developer.mediamarktsaturn.com/apis/car-factfinder-api/index

En cuanto a las restricciones indicadas por ambas páginas respecto a la extracción automatizada de datos de sus webs, se ha podido comprobar que no incorporan restricciones a las búsquedas que han sido planteadas (secciones de productos como portátiles, tablets y monitores). No obstante, dominiovirtual.es establece restricciones al realizar búsquedas por querys en las URL, práctica que se ha evitado en el código generado. Las restricciones de ambos sites están publicadas en:

- https://www.mediamarkt.es/robots.txt
- https://www.dominiovirtual.es/robots.txt

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

El conjunto de datos recogidos dispone de 3 aspectos por los cuales se considera interesante su utilidad:

- Comprobar si un producto específico está disponible en alguno de los portales web analizados para poder realizar su compra.
- Comparativa de precios entre distintos sitios web especializados en estos productos.
- Disponer de capacidades analíticas para analizar si los precios a los que se encuentra cada producto son acordes a las especificaciones y características que tienen (por ejemplo, detectar aumentos de precios en función de marcas o similar).

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

- Released Under CC0: Public Domain License
- Released Under CC BY-NC-SA 4.0 License
- Released Under CC BY-SA 4.0 License
- Database released under Open Database License, individual contents under Database Contents License
- Other (specified above)
- Unknown License

Pendiente

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

https://github.com/vblanesUOC/PRA1-Web-Scraping

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

Pendiente

11. Contribuciones:

Contribuciones	Firma
Investigación previa teórica (lectura material UOC, tutoriales y documentación de web scraping externos, revisión ejemplos anteriores, etc.)	VB, CA
Investigación y elección sitios web (estudio de sus características e idoneidad para llevar a cabo los objetivos).	VB, CA
Desarrollo Web Scraping con Selenium	VB, CA
Desarrollo Web Scraping con Scrapy	VB, CA
Unificación datasets	VB, CA
Elaboración informe	VB, CA