PRA1: Web Scraping

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

Víctor Blanes Martín
Carlos Allo Latorre

01/04/2021

Tabla de contenido

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.
Hoy en día, la compra de dispositivos electrónicos está a la orden del día y es más frecuente que nunca. A raíz del desarrollo de nuevas tecnologías, cada vez es más frecuente disponer de una oferta de productos más amplia, excesiva en ocasiones, que dificulta la toma de decisiones en cuanto a la compra de dichos productos.

En función del perfil del comprador, el desconocimiento de la gama de productos y sus características puede que los lleve a tomar una decisión de compra poco adecuada o ajustada en precio. Por este motivo, disponer de un repositorio centralizado con datos de productos, especificaciones y precios de productos de distintos sitios web, podría facilitar la compra tanto para la gente inexperta como para los usuarios más exigentes que requieran maximizar la relación especificaciones/precio.

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.

La construcción del dataset permite añadir futuros datos a partir de nuevos scraping de otros sitios web. Dado que el modelo y especificaciones técnicas son las mismas, esto nos permitiría ampliar el espectro de precios disponibles para la compra del producto.

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

Para identificar de forma general la publicación, se hará uso del siguiente título:

- Especificaciones y precios de dos vendedores de productos electrónicos

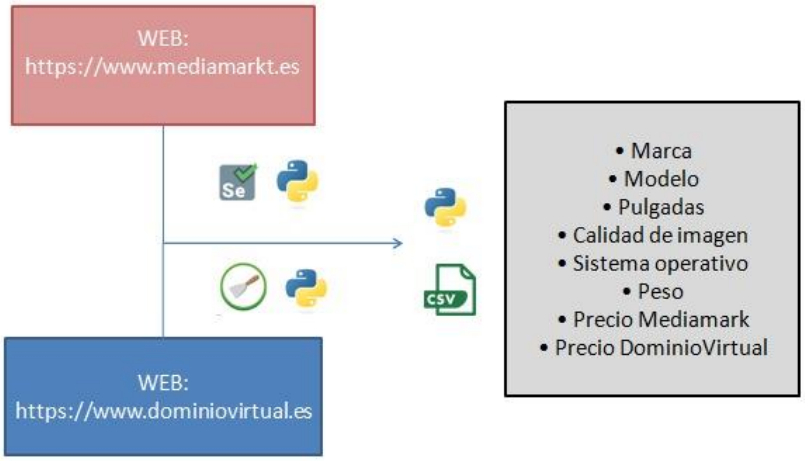
3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).
Cada uno de los tres datasets que se han comentado en la anterior pregunta está formado por siete columnas, que incluyen la marca y el modelo del producto para poder identificarlo, las pulgadas, calidad de imagen, sistema operativo y peso del producto, así como su precio en cada uno de los sitios web escogidos.

Dado que se han extraído tres datasets independientes en función del tipo de producto, cada uno de ellos tiene una cantidad de elementos distinta. Aproximadamente, los dataset de portátiles y monitores tienen un volumen aproximado de 100 productos, mientras que el de tablets (donde hay menos variedad de productos) dispone de unos 50 elementos.

Tal y como se ha mencionado anteriormente, el dataset resultante puede ser fácilmente ampliable (de requerirse en futuros proyectos) con campos adicionales que incluyan el precio en

otros sitios web para ampliar la variedad de modelos y mejorar la comparación entre modelos iguales.

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 11 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.

En la web de DominioVirtual, se ha iniciado el procedimiento de extracción a partir de la URL de cada una de las categorías y se ha ido rastreando, accediendo a cada uno de los productos de donde se extraía la información, a partir de los HTML proporcionados.

Además, se han tenido en cuenta técnicas para que no se identificara al programa como robot. Por ejemplo, el hecho de que espere un segundo entre cada una de las peticiones, o que se cambie la cabecera donde se indica quién realiza la petición, poniendo en ésta navegadores diferentes en peticiones contiguas, evitando de esta forma que el servidor nos etiquete como 'no humanos' o robots.

En la web de Mediamarkt, de forma similar a DominioVirtual, también se partía de la URL base de cada categoría, a partir de la cual se extraía un listado de links que apuntaban a cada uno de los productos de la categoría (iterando las páginas de cada categoría para sacar una mayor cantidad de productos).

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 al 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

Respecto a posibles análisis que se hayan hecho con anterioridad sobre el mismo tema, se ha identificado una muestra de dataset publicado en Kaggle por el usuario DataMarket, y que está disponible en el siguiente enlace:

- https://www.kaggle.com/datamarket/productos-de-electrnica

El contenido del dataset, si bien no es el mismo que se ha planteado en esta práctica, tiene cierta similitud ya que plantea una distinción de productos por sitios web, categorías, modelos, precios y valoración otorgada. No obstante, en Kaggle únicamente hay publicada una muestra del conjunto de datos y no se ha podido tener acceso a la muestra completa ya que ha dejado de estar disponible el vínculo a la misma (https://datamarket.es/).

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.

Aunque en un primer momento se planteó la utilidad del conjunto de datos para ofrecer una mayor facilidad a un público objetivo poco familiarizado con los productos electrónicos, a lo largo del diseño y construcción del dataset han ido surgiendo diferentes aspectos que lo hacen interesante.

Primero, nos da la posibilidad de disponer de información actualizada (precios y características) acerca del stock de productos de las principales páginas web del mercado de productos electrónicos. Aunque esta práctica se ha restringido al scraping de dos sitios web, el concepto bajo el que se ha planteado para el dataset nos permitiría nutrirlo de nueva información de otros sitios web, cosa que enriquecería la información proporcionada por el mismo.

En base a lo comentado, esto nos permite cubrir varias necesidades que consideramos útiles hoy en día:

- Verificar el stock de productos en varios sitios web. Por ejemplo, posibilidad de buscar un producto concreto que tengamos claro que vamos a comprar, para verificar su disponibilidad y precio.
- Realizar una comparativa de precios entre los distintos sitios web. Puede servir tanto a modo genérico, por ejemplo, para saber qué web suele tener los precios más asequibles, como para realizar un análisis más en detalle, por ejemplo, para ver dónde venden más barato el producto que se desee.
- Realizar búsquedas en base a características preferidas. Por ejemplo, si la preferencia principal de una compra fuera la calidad de imagen de un dispositivo, podríamos identificar dónde venden los productos con mejor calidad de imagen y más baratos.
- Capacidades analíticas derivadas de las características de los productos. Por ejemplo, estos análisis podrían servir para determinar si el precio de un producto está sobrevalorado respecto a las características técnicas que presenta, o si un sitio web tiende más a vender productos de alta gama que de gama baja.

En cuanto al análisis que se ha mencionado en el ejercicio 6 (DataMarket), podemos ver que tiene una capacidad más limitada para responder algunas de las cuestiones que nos hemos planteado. Por ejemplo, al no disponer de información acerca de las especificaciones de los productos, no se podrían realizar las búsquedas en base a características preferidas y se reducirían en gran parte las capacidades analíticas que proporciona el conjunto de los datos.

Por último cabe mencionar tres limitaciones o dificultades que se han encontrado al realizar el trabajo que nos han hecho vivir en primera persona las diferentes dificultades y barreras del mundo del Web Sraping. La primera de ellas ha aparecido al realizar el rastreo de dominio virtual. Una vez planteada la paginación para la misma, la cual funcionaba por adición de parámetros en la url (‘?p=NumPag’), en el archivo robots.txt se encontró que las peticiones con parámetros no están permitidas por lo que se dejó comentada esta parte de código. Además, muy pocos días antes de entregar la práctica se produjo un cambio en la web en este sentido, en donde la paginación no iba por parámetro sino por la adición a la sección de ‘/page-NumPag/’. Sin embargo, por falta de tiempo y por tener la práctica muy desarrollada, no se cambió esta parte. Cabe decir que esto no se considera un problema porque cada página de producto contiene otros enlaces a productos de la categoría que permitirá recorrer un gran abanico de los productos.

Por otro lado, el segundo punto a comentar fueron las complicaciones en la parte de unión de ambas webs, ya que no para ambas páginas el código de referencia o modelo hacían mención al mismo identificador. Con un estudio exhaustivo de ambas webs y con una adición de lógica en la parte del limpiado de datos del Dominio Virtual, se logró que este identificador fuera el mismo para de esta forma, poder hacer unión con los productos. Es importante decir que esta unión resultó producirse para pocos productos por categoría ya que en los csv obtenidos de cada web los productos eran muy diferentes unos de otros (hablando de referencias de los mismos).

Finalmente, la última limitación (y que podría ser objeto de mejora para próximos estudios), fue que en alguna ocasión se contaban con mismos productos (misma marca-modelo) con diferentes precios. Investigando un poco más, esto era debido a que poseían diferentes procesadores, almacenamiento… que eran variables no tomadas en cuenta en este análisis y que complicaban la unión anterior. Por ello, la decisión tomada ha sido seleccionar, en el caso de que esto ocurra, el elemento con el precio más bajo, siendo el consumidor del dataSet el responsable de verificar si el producto cumple con sus necesidades o requiere uno con mejores características.

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

La licencia escogida para la publicación de este conjunto de datos es: CC BY-NC-SA 4.0 License. El motivo de la elección de esta licencia es el restringir los usos comerciales que se le pueda dar al dataset generado, puesto que proviene de unos sitios web comerciales con sus propias licencias que también restringen este tipo de derechos.

Por lo tanto, la licencia que se aplica al dataset conlleva:

- Atribución: Se debe dar crédito de manera adecuada, facilitando un enlace a la licencia e indicar si se han realizado cambios sobre el conjunto de datos.
- No comercial: No se puede hacer uso del conjunto de datos con propósitos comerciales.
- Compartir igual: Si se modifica o complementa el material con más información, se debe distribuir la contribución bajo la misma licencia.

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.
El código utilizado para la extracción de los datos puede consultarse en el siguiente repositorio de Github, dentro de la carpeta 'src':

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

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.
Los CSV obtenidos, tanto los intermedios como el resultado de su unión se encuentran disponibles Github en el siguiente enlace:

- https://github.com/vblanesUOC/PRA1-Web-Scraping/tree/main/csv

Asimismo, están publicados en Zenodo y son consultables desde el siguiente enlace:

- https://zenodo.org/record/4679647#.YHMxCugzYuU

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