PRA1 - Web Scraping.

Tipología y Ciclo de Vida de los Datos.

Autores:

David Herrero Pascual

Andrés Baamonde Lozano

1. Contexto. 3
2. Definir un título para el dataset. 3
3. Descripción del dataset. 3
4. Representación gráfica. 4
5. Contenido. 4
6. Agradecimientos. 5
7. Inspiración. 5
8. Licencia. 6
9. Código. 6
10. Dataset. 6
11. Bibliografía. 7
12. Tabla de contribuciones. 7

1. Contexto.

Actualmente, y en gran parte debido a las restricciones de movilidad y problemas derivados de la pandemia de COVID-19 presente en todo el mundo, el comercio electrónico no ha dejado de crecer en el último año. Según *eMarketer*, una compañía de investigación de mercado que proporciona tendencias relacionadas con el marketing digital, los medios y el comercio, España es el tercer mercado de todo el mundo en el que más ha crecido el comercio electrónico a lo largo del año 2020, solamente superada por Argentina y Singapur. [1]

En este contexto, los supermercados tradicionales han hecho un gran esfuerzo por mejorar sus servicios online, y han invertido en mejorar sus webs para administrar un mayor tráfico y poder asimilar un mayor número de pedidos online.

Así, la gran mayoría de supermercados tradicionales ofrecen la posibilidad de realizar pedidos de forma online, sin necesidad de desplazarse físicamente, y poniendo a disposición de los clientes todo el catálogo de productos como si de los estantes se tratase.

Este proyecto de Web Scraping surge bajo el contexto de la asignatura Tipología y Ciclo de Vida de los datos, perteneciente al Máster en Ciencia de Datos de la Universitat Oberta de Catalunya. En ella, se aplican técnicas de Web Scraping y, en nuestro caso, se desarrolla un proyecto en Python que permite obtener datos de dos supermercados online: Carrefour y Día.

Para la práctica se han obtenido diferentes datos relativos a todas las opciones de leche disponibles en ambos supermercados con el objetivo de poder, por ejemplo, realizar un futuro estudio de mercado para determinar cuál de los dos supermercados es más barato en lo que a este tipo de productos se refiere.

El desarrollo del código se ha llevado a cabo con una filosofía de modularidad y reutilización, de forma que es un código aplicable a cualquier otra sección de cualquiera de los dos supermercados.

2. Definir un título para el dataset.

Para la realización de la práctica hemos optado por obtener los precios de la leche en los supermercados Dia y Carrefour, por lo que hemos decidido llamar al Dataset ***Milk Price Dataset.***

3. Descripción del dataset.

El dataset desarrollado contiene 168 entradas de diferentes tipos de leche vendidas en Carrefour y Dia, con diferentes campos como su precio, precio en oferta, precio por unidad de medida, etc. En el punto 5 de esta práctica se incluye una descripción detallada de la información recogida de cada artículo. Además, los productos van con un timestamp asociado y un identificador único, de

forma que se puede realizar una obtención de datos en diferentes momentos y realizar un estudio de la evolución de los precios en ambos supermercados.

4. Representación gráfica.

En la siguiente tabla, se pueden observar los diferentes campos presentes en el CSV.

description	name	price	offer_price	meassure	pum	size	brand	meassure_description	market	timestamp	identifier

Este CSV es un archivo delimitado por comas, con una cabecera con los nombres de las columnas, y a en la siguiente imagen se puede observar una muestra de algunos valores del csv:

```
description,name,price,offer_price,meassure,pum,size,brand,meassure_description,market,timestamp,identifier
Leche semidesnatada Carrefour brik 1l.,Leche UHT semidesnatada,0.56,0.58,1,0.58,,CARREFOUR,,carrefour,1618235804.9275024,R-521007071
Leche entera Carrefour brik 1l.,Leche UHT entera,0.56,0.58,1,0.58,,CARREFOUR,,carrefour,1618235812.1105704,R-521006992
Leche semidesnatada Central Lechera Asturiana brik 1 l.,Leche UHT SEMIDESNATADA,0.76,0.79,1,0.79,,CENTRAL LECHERA ASTURIANA,,carrefour,1618235819.3329542,R-521007075
Leche entera Central Lechera Asturiana brik 1 l.,Leche UHT entera,0.76,0.79,1,0.79,,CENTRAL LECHERA ASTURIANA,,carrefour,1618235826.4175413,R-521006994
Leche semidesnatada Carrefour sin lactosa brik 1 l.,"Lech UHT semidesnatada sin lactosa, enriquecida con vitamina A D E y ácido fólico.",0.75,None,1,0.75,,CARREFOUR,,carrefour,1618235833.5271351,R-714713105
Leche entera Pascual brik 1 l.,Leche entera UHT,0.8,None,1,0.8,,PASCUAL,,carrefour,1618235840.634083,R-521006986
Leche desnatada Carrefour sin lactosa brik 1 l.,"Lech UHT desnatada sin lactosa, enriquecida con vitamina A D E y ácido fólico.",0.75,None,1,0.75,,CARREFOUR,,carrefour,1618235847.7570148,R-714713109
Leche semidesnatada Central Lechera Asturiana sin lactosa brik 1 l.,Leche UHT semidesnatada sin lactosa,0.93,None,1,0.93,,CENTRAL LECHERA ASTURIANA,,carrefour,1618235854.776314,R-670001999
```

5. Contenido.

A la hora de realizar el desarrollo, se ha definido una clase llamada ***Article***, la cual tiene los siguientes atributos.

- **description**: String que representa la descripción del producto, por ejemplo, “Leche entera Carrefour brik 1l”.
- **name**: String que representa el nombre del producto, por ejemplo, “Leche UHT semidesnatada”.
- **price**: Float que representa el precio en euros del producto.
- **offer_price**: Float que representa el precio del producto cuando está en oferta.
- **meassure**: String que representa la unidad de medida del producto, por ejemplo, l (litros).
- **pum**: Float que representa el precio por unidad de medida, por ejemplo, 0.56€/L
- **size**: Float que representa el tamaño del artículo, por ejemplo 1 (litro).
- **brand**: String que representa la marca del producto, por ejemplo, Central Lechera Asturiana.
- **meassure_description**: String que representa el formato de la medida, por ejemplo, brick 1 l.
- **market**: String que representa el supermercado que vende el producto, en nuestro caso, será Carrefour o Dia.
- **timestamp**: Float, Unix Timestamp que representa la marca temporal en la que se ha realizado el scraping del producto.
- **identifier**: String, identificador único del producto en cada supermercado.

Con esto en mente, se ha guardado la información de 168 tipos de leche en un archivo CSV, en el que cada columna representa uno de estos atributos.

6. Agradecimientos.

En nuestro caso, existen dos propietarios de los datos obtenidos para este Dataset: Carrefour y Dia. Toda la información referente a sus productos se puede encontrar en sus correspondientes supermercados online, disponibles desde las siguientes URLs:

- Dia: https://www.dia.es/compra-online/
- Carrefour: https://www.carrefour.es/supermercado/

Actualmente, existen algunas aplicaciones como Radarprice (http://www.radarprice.com/es/) o Minderest (https://www.minderest.com/es/software-comparacion-precios) que realizan esta función, pero de forma distinta a cómo lo hemos planteado en esta práctica.

Para la realización de esta práctica hemos utilizado la librería *Selenium* para Python y su documentación oficial, disponible desde el siguiente enlace: https://www.selenium.dev/documentation/en/

El proyecto con mayor semejanza a lo que se ha realizado en esta práctica es el del usuario tonsmets en Github, disponible desde el siguiente enlace: https://github.com/tonsmets/SupermarketScraper

Mediante este proyecto, se puede hacer scraping de diversos supermercados holandeses. Es un desarrollo más extenso y que incluye más supermercados, pero lo cierto es que no hemos encontrado ningún proyecto similar con supermercados españoles.

También es importante tener en cuenta que hasta hace poco algunos supermercados españoles como Mercadona no tenían opción de compra online, o si la tenían era tremendamente rudimentaria, y que a raíz de la pandemia de COVID-19 todo lo referente a venta online ha crecido enormemente, lo que abre las puertas para la realización de proyectos como este.

7. Inspiración.

La obtención de este dataset se ha realizado buscando un objetivo: el análisis de precios de productos semejantes en dos grandes superficies diferentes. Mediante la obtención de datasets como este, se podrían desarrollar aplicaciones web que permitieran al usuario escoger el supermercado más barato para hacer la compra en función de los productos que quiera comprar.

Con todas las opciones que existen actualmente para hacer la compra online, sería muy positivo para los consumidores contar con una herramienta que les permitiera saber dónde hacer su compra online y ahorrarse dinero sin tener que estar entrando en cada página y comparando de forma manual el precio. A partir de este proyecto, se podría ampliar la aplicación de muchas formas, como por ejemplo teniendo en cuenta ofertas del tipo 2x1 o semejantes, que actualmente no se contemplan en el dataset.

Aunque existen numerosos proyectos de Web Scraping para supermercados, lo cierto es que no hemos encontrado prácticamente ninguno que tome datos de dos supermercados distintos, y creemos que es muy positivo contar con dos webs distintas para poder realizar un estudio de mercado y una comparativa de precios más profunda, incluso pudiendo detectar si un determinado supermercado está subiendo su precio por encima de los demás, o si es el producto el que está subiendo y lo está haciendo en todos.

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

Atendiendo a las restricciones que deseamos imponer al conjunto de datos podríamos escoger las licencias menos restrictivas.
Las licencias que obligan a un uso no comercial, en este ámbito no tienen demasiado sentido, puesto que su propósito, sería elaborar comparativas entre supermercados. Esto nos descarta CC BY-NC-SA 4.0.

Una licencia interesante en caso de querer asegurarse que los datos no cambian de licencia sería CC BY-SA 4.0, con ella mantendremos nuestra autoría de los datos y nos aseguraremos que datos derivados de ellos seguirán bajo la misma licencia. O incluso ODbL, que permite utilizar licencias en caso de ser más restrictivas aún.

Pero debemos elegir CCO, que no tiene derechos autorales ya que nos encontramos en el marco de una práctica de un master, el conjunto de datos es extraído de una web sin ánimo de lucro y no se realiza ningún tipo de procesado o cálculo de atributos nuevos del conjunto de datos. Por lo que no es relevante proteger la autoría de los datos.

9. Código.

El código está disponible en la plataforma Github, en el siguiente repositorio: https://github.com/Deividhp13/uoc-datascrapping-supermarket.

Existe un README.md que explica como instalar el entorno. La totalidad del código está escrito en Python, y serán compatibles versiones de python de la 3.8 en adelante (debido al uso de las dataclasses).

10. Dataset.

El dataset está disponible en zenodo y su DOI es: "10.5281/zenodo.4681691", este csv también está disponible en el código de git.

11. Bibliografía.

[1]. *España es el tercer mercado de todo el mundo en el que más ha crecido el comercio electrónico en 2020 - Marta Pachón Díaz, 15/21/2020 08:00, disponible online en:* https://www.businessinsider.es/espana-tercer-mercado-donde-crecio-ecommerce-2020-774071

12. Tabla de contribuciones.

Contribuciones	Firma
Investigación Previa	ABL, DHP
Redacción de las respuestas	ABL, DHP
Desarrollo código	ABL, DHP