COMPARACIÓN DE LOS PRECIOS EN LOS SUPERMERCADOS ESPAÑOLES

Mercadona, Caprabo y Bonpreu

Lucia Blanc Velázquez
Sergio Sanchez Romero

Noviembre 2023

Tipología de datos
Máster en Ciencia de Datos

Universitat Oberta de Catalunya

Lucia Blanc Velázquez
Sergio Sánchez Romero

1. Contexto

En esta actividad se ha llevado a cabo la recopilación de datos con el propósito de comparar los precios de los productos en tres supermercados diferentes: Mercadona, Caprabo y Bonpreu.

Sitios web (enlaces):

- Mercadona : https://tienda.mercadona.es/
- Caprabo: https://www.capraboacasa.com/es
- Bonpreu: https://www.bonpreuesclat.cat/es/home

Esta iniciativa se ha desarrollado en respuesta a la notable subida de precios de diferentes productos, como puede ser el aceite de oliva, un hecho que ha impactado en la economía de los consumidores, y por consiguiente ha generado un creciente interés de identificar las opciones de compra más económicas.

2. Título: Comparación de precios en los supermercados españoles.

El título del dataset "*products.csv*" incluye la fecha de captura de los datos, la palabra "products" la cual hace referencia a la obtención de cada uno de los productos que encontramos en los supermercados de estudio. Es importante remarcar la importancia de documentar la extracción de los datos, debido al hecho que los precios de los productos de los supermercados pueden ir cambiando.
La solución del conjunto de datos es tabular, ya que, como se puede observar se usa la extensión *.CSV.

3. Descripción del dataset

El dataset "*products.csv*" presenta la información de todos los productos disponibles en las páginas webs del supermercado Mercadona, Caprabo y Bonpreu, donde se adjunta las características principales como categoría, subcategoría, producto e información del producto, permitiendo la comparación de precios de productos similares.

Lucia Blanc Velázquez
Sergio Sánchez Romero

4. Representación gráfica

Todos los supermercados incluidos en este trabajo tienen un apartado de tienda online, por lo que, para realizar el scraping correspondiente se ha seguido la estructura de cada una de las páginas web, siguiendo las categorías y subcategorías, y por mediante de la captura de las URLs correspondiente a cada producto. Una vez obtenemos la información completa de los productos se importa a un aplicación "flask" que funciona para buscar y ayudar a clasificar por precios los productos de cada supermercado.

En las siguientes imágenes se pueden ver dos fases el proceso seguido en formato gráfico para la obtención del csv final y la correspondiente aplicación creada.

Fase 1: Web scraping de los productos de los supermercados y creación de un csv conjunto ("products.csv").

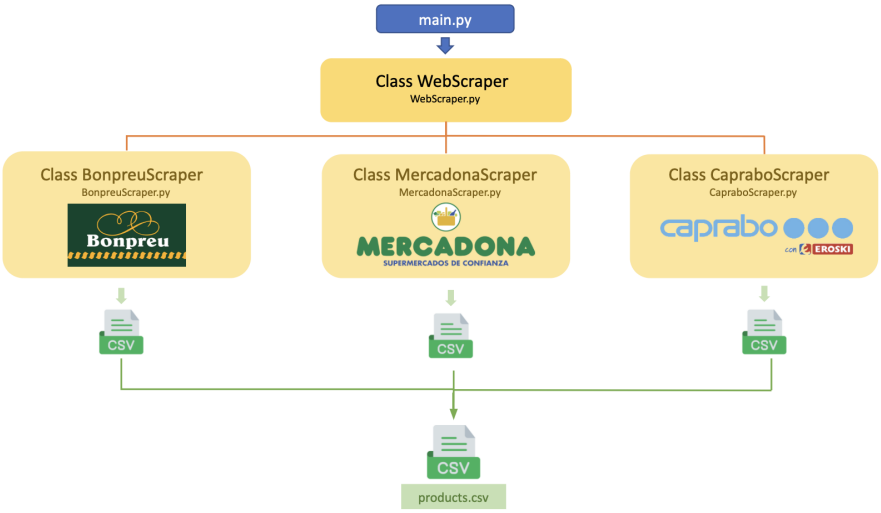

Figura 1. Procedimiento del webscraper del supermercado Mercadona, Caprabo y Bonpreu.

Lucia Blanc Velázquez
Sergio Sánchez Romero

Fase 2: Importación del csv en una app buscador, donde se visualizan los productos ordenados según el precio.

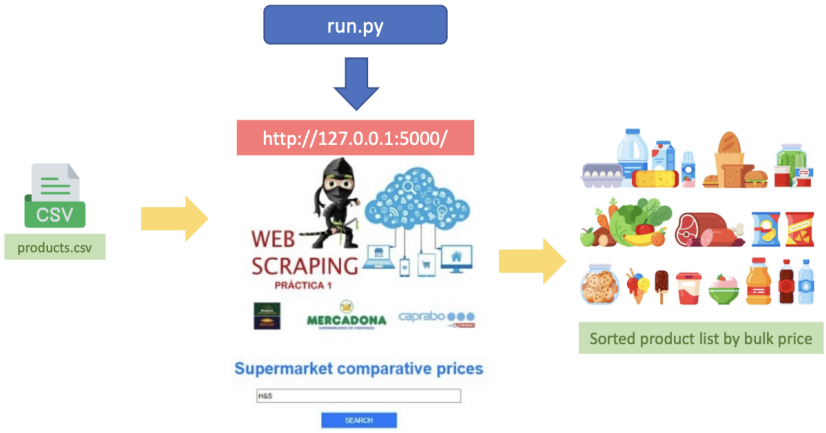

Figura 2. Procedimiento de importación del csv generado a una app buscador.

5. Contenido

A continuación se presentan los campos del dataset generado en una tabla:

Campo		**Descripción**	
Supermarket	supermarket	Nombre del supermercado donde está disponible el producto	String
Main Category	id_cat_principal	Identificador de la categoría principal del producto	Int
	name_cat_principal	Nombre de la categoría principal del producto	String
	url_cat_principal	URL donde se puede encontrar el la categoría principal	String
Category	id_category	Identificador de la categoría del producto	Int
	name_category	Nombre de la categoría del producto	String
	url_category	URL donde se puede encontrar el la categoría	String
Subcategory	id_subcategory	Identificador de la subcategoría del producto	Int

	name_sub_category	Nombre de la subcategoría del producto	String
	url_sub_category	URL donde se puede encontrar el la subcategoría	String
Product	id_product	Identificador del producto	Int
	name_product	Nombre del producto	String
	url_product	URL donde se puede encontrar el producto online	String
Product information	thumbnail	URL a una imagen en miniatura del producto	String
	packaging	Tipo de embalaje utilizado para empaquetar el producto	String
	unit_size	Tamaño o cantidad del producto	Numeric
	unit_price	El precio del producto	Numeric
	size_format	Formato o unidad de medida del producto (“ud”: unidades)	String
	bulk_price	Precio a granel por comprar o descuento por comprar al por mayor	Numeric

Tabla 1: Definición de los campos de los que se compone el dataset

6. Propietario

6.1. Propietario de los datos

- Mercadona: El propietario del sitio web es MERCADONA, S.A., con domicilio en Tavernes Blanques (Valencia), C/ Valencia nº 5. Su NIF es A-46.103.834.
- Caprabo: El propietario del sitio web es CAPRABO, S.A. Av. Estany de la Messeguera, 40-44 El Prat de Llobregat (Barcelona) CIF: A08115032.
- Bonpreu: El propietario del sitio web es Bon Preu, SAU, con domicilio en Les Masies de Voltregà, Ctra. C-17, Km 73 y CIF A-08665838.

6.2. Acceso ético

Durante el proceso de recopilación de los datos , se han seguido los principios éticos y legales, así como las directrices establecidas en los archivos robots.txt

-presentados en el siguiente punto- de cada supermercado con el fin de evitar acceso a aquellas áreas no autorizadas de las páginas web.
El acceso se limitó a las secciones públicas relacionadas con los productos y los precios correspondientes.

6.3. Robots.txt

En relación a protecciones detectades, a continuación se específica el contenido del archivo robots.txt:

- Mercadona: El archivo indica que se permite el acceso a la página raíz (/), al archivo sitemap (/sitemap.xml) y a la sección de productos (/product), pero no a otras partes del sitio web. Es decir, se permite el raspado web a las áreas públicas relacionadas con los productos y los precios.

```
User-agent: *
Allow: /$
Allow: /sitemap.xml
Allow: /product
Disallow: /
Disallow: /api
Disallow: /legal
```

- Caprabo: Según lo que aparece en el archivo robots está prohibido el acceso a páginas con el siguiente URL "*/search/results/?q=" y "*/?zipCode" con el fin de evitar acceso a resultados de búsqueda específicos ni a páginas relacionadas con códigos postales o ubicaciones.

```
User-agent: *
Disallow: */search/results/?q=
Disallow:  */?zipCode
```

- Bonpreu: El archivo de robots prohíbe el acceso a las URL que contiene "/group", "/c/portal" ni "/*?p_p_id=3" con el fin de evitar que motores de búsqueda accedan a páginas relacionadas con el portal y con parámetros específicos como el ID 3.
  Se indica también que tiene un canal de venta virtual (sitemap) diferenciado de la página web específica del supermercado.

```
User-Agent: *
Disallow: /group*/
Disallow: /c/portal/
Disallow: /*?p_p_id=3
Sitemap: https://www.bonpreuesclat.cat/sitemap.xml
```

6.4. Documentación Legal y Política de Privacidad

En cada uno de los sitios webs scrapeados y visitados, se indica un uso respetado de los documentos y se permite el acceso a toda aquella información que se encuentra disponible de cara al público.

Para más información específica, se permite visitar los documentos de Términos y Condiciones de Uso de cada uno de los sitios web.

7. Inspiración

El presente conjunto de datos se podría explotar y aprovechar de diversas maneras, tanto a nivel de la empresa como la de los clientes.

7.1 Nivel de la Empresa

Uno de los posibles usos de los datos de "products.csv" en la aplicación Flask a nivel de empresa, podría ser la de optimizar el inventario y la estrategia de compra, usando los datos de cada venta y las preferencias que tiene cada cliente para estudiar y analizar las tendencias de compra, la demanda de los productos y las fluctuaciones. De esta manera, se optimizaría el inventario, asegurando que los productos están disponibles cuando los clientes los quieran sin producir excesos en el stock, al igual que se podrían identificar los productos populares y promoverlos en momentos estratégicos con el fin de aumentar las ventas de estos.

7.2 Nivel del Cliente

Siguiendo alguno de los ejemplos encontrados en otras páginas web de comparador de productos de supermercados[4], una de las maneras que consideramos más importante de aplicar sería el hecho de que los clientes pudieran usar estos datos generados, mediante el uso de una aplicación móvil, para comparar la calidad y el precio de los productos entre varios supermercados, y así tomar mejores decisiones y ahorrar dinero. La aplicación podría permitir que el cliente tuviera la posibilidad de crear listas de compra personalizadas y estudiar donde saldría más económico y factible comprar los productos.

Los datos recopilados de los supermercados de estudio, podrían usar tanto por cada una de las empresas para mejorar la eficiencia y aumentar la fidelización de los clientes, como por los clientes para mejores tomar decisiones ahorrando tiempo y dinero, es decir, la explotación de estos datos conlleva beneficios significativos para ambas partes.

8. Licencia

La licencia seleccionada para el dataset resultante es: Licence **Commons Attribution-Non-Commercial-ShareAlike 4.0 International (CC BY-NC-SA)**.

Esta licencia es la más adecuada para este dataset al combinar información de diferentes fuentes, y por lo tanto se necesita que respete los derechos de autor y los términos de uso de las páginas web de donde hemos obtenido la información.

Las características de esta licencia permite que otros usuarios utilicen y compartan el conjunto de datos siempre y cuando den crédito al creador por la recopilación y creación de los datos, es decir, se hace un reconocimiento del trabajo y esfuerzo en la obtención. Por otro lado, es NonCommercial, es decir que no se pueden usar los datos con un objetivo comercial sin el permiso del creador. La licencia CC BY-NC-SA está aceptada en la comunidad de datos abiertos, permitiendo crear obras derivadas bajo la misma licencia, hecho que fomenta la colaboración y el enriquecimiento de los datos. Por último, el uso de esta licencia fomenta la transparencia en la recopilación y la comparación de los precios de los productos de los diferentes supermercados, lo que puede generar un beneficio directo a los consumidores y otros investigadores.

9. Código

El código fuente de extracción y el dataset creado son accesibles mediante la plataforma GitHub. Para acceder se puede visitar el siguiente enlace: https://github.com/ssanchezromer/supers

El programa usado para realizar la codificación de esta práctica ha sido PyCharm Community, donde se han usado scripts con el formato .py.

Los principales paquetes que se han usado durante toda la práctica han sido los siguientes:

- BeautifulSoap
- Selenium
- CSV
- Requests
- Time
- Pandas

Los archivos .py creados para está práctica son:

- *WebScraper.py:* Proporciona métodos y funciones para realizar web scraping en las páginas de Mercadona, Caprabo y Bonpreu, y administrar los datos extraídos
- *MercadonaScraper.py:* Recopila datos de los productos del supermercado "Mercadona" a través de la API web.
- *CapraboScraper.py:* Recopilan datos de los productos del supermercado "Caprabo".
- *BonpreuScraper.py:* Recopilan datos de los productos del supermercado"Bonpreu
- *main.py y run.py:* Con el fin de juntar todos los procesos de scraping de los tres supermercados, y cargarlo en la aplicación de Flask creada para realizar las búsquedas de productos pertinentes.

10. Dataset

El dataset generado en está práctica se encuentra almacenado en el repositorio público Zenodo, accesible de manera pública sin restricciones.
El nombre y la descripción del conjunto de datos coinciden con las características especificadas anteriormente en el documento y por lo tanto, el nombre del fichero CSV es "products.csv".

El DOI del dataset publicado en Zenodo se encuentra en el siguiente enlace: https://zenodo.org/records/10086087

11. Contribuciones

La tabla de contribuciones es la siguiente:

Contribuciones	**Firma**
Investigación prèvia	Sergi Sánchez Romero, Lucia Blanc Velázquez
Redacción de contenido	Sergi Sánchez Romero, Lucia Blanc Velázquez
Desarrollo del código	Sergi Sánchez Romero, Lucia Blanc Velázquez

Referencias

1. Lawson, R. (2015). *Web Scraping with Python*. Packt Publishing Ltd. Chapter 2. Scraping the Data.
2. Mitchel, R. (2015). *Web Scraping with Python: Collecting Data from the Modern Web*. O'Reilly Media, Inc. Chapter 1. Your First Web Scraper.
3. Open Data Commons. https://creativecommons.org/licenses/by-nc-sa/4.0/
4. Soysuper. (s.f.). Aceite de oliva. https://soysuper.com/search?q=Aceite%20de%20oliva&p=2E26F150-7DA6-11EE-821B-B23648A4FFC4