Memoria Scrapping Dia

Autores:

Óscar Promio opromio@uoc.edu
Carlos Perez cperezha@uoc.edu

1. Contexto

Para la realización de la práctica valoramos varios escenarios, buscamos un sitio web en que estuviera permitido realizar scrapping y, a la vez, que la información extraída tuviera valor. Uno de los escenarios más habituales en proyectos de scrapping consiste en obtener información en varios sitios web para luego compararlos (esto es muy habitual en buscadores de vuelos como SkyScanner o de seguros como Rastreator), pero eso requiere realizar scrapping sobre varios sitios web con las dificultades que eso conlleva.

Por ese motivo decidimos buscar información que pudiéramos comparar contra sí misma con el paso del tiempo. En este punto, y con el contexto global actual, pensamos que una buena propuesta podría ser monitorizar los precios en un supermercado, ya que se actualizan de forma periódica y son un gran indicador de la inflación.

Ambos hacemos uso de la compra online en el supermercado y nos habíamos percatado de la subida constante de precios en los últimos meses. A partir de la información que hemos recogido en este proyecto podremos medir el cambio en el precio de los productos, compararlo entre diferentes categorías y extraer conclusiones al respecto.

Una vez decidido el enfoque, valoramos varios supermercados según su popularidad y permisos para realizar scrapping como Mercadona, Consum y Dia. Durante las primeras exploraciones advertimos que Mercadona y Consum la web generan dinámicamente el contenido a partir de Javascript, dificultando en gran medida el scrapping, por lo que finalmente decidimos apostar por la web de Dia, que está construida con Hybris (un software de construcción de sites de ecommerce) y esto propicia que esté mucho más estructurada.

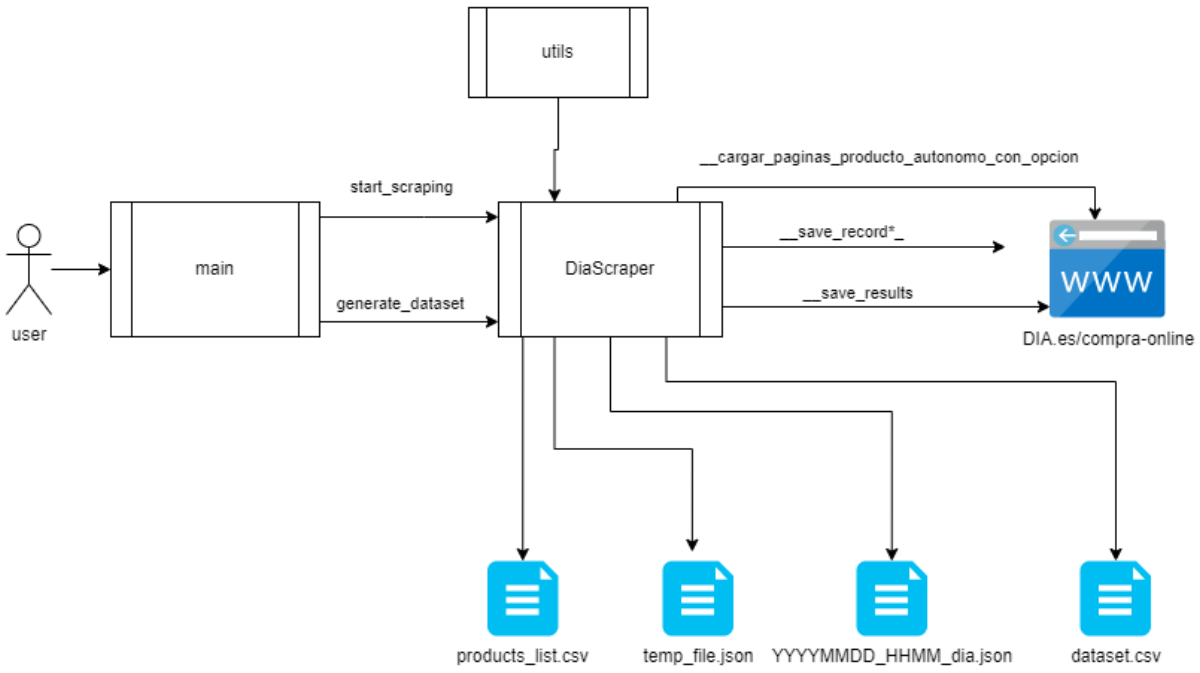
2. Título y descripción del dataset

Hemos titulado el dataset generado como “**Análisis de evolución del precio de productos de gran superficie**”. Nuestro objetivo consiste en recoger toda la información posible de productos, así como su precio para el análisis en un supermercado de gran superficie, en este caso Supermercados DIA:

https://www.dia.es/compra-online/

Hemos recogido toda la información que hemos considerado relevante en la web, en los datos encontramos variables como la categoría del producto, la marca, el descuento y el precio unitario. Estos campos nos ayudarán a realizar comparaciones más precisas y extraer conclusiones más generales como se detalla en el apartado de Inspiración.

3. Representación gráfica



El programa se lanza desde un módulo “main”, creando un objeto de la clase **DiaScraper** e invocando al método “**start_scraping**”. Éste realiza tres funciones:

- **__cargar_paginas_producto_autonomo_con_opcion**: descubre las URLs de los productos que se van a escrapear a partir del site de DIA, persistiéndolas en el archivo “products_list.csv” a modo de caché.

- **__save_record**: recorre cada URL de producto de la lista “products_list.csv”, escaneando la información de los mismos y almacenando cada producto en un fichero temporal.
- **__save_results:** transforma los ficheros temporales en un fichero diario con toda la información de los productos escrapeados.

Por último, ejecuta “**generate_dataset**” y genera el archivo dataset.csv con la información de toda la información diaria generada.

4. Contenido

El dataset está constituido por la información de producto para diferentes fechas de ejecución. La información que se recoge es la siguiente:

- **date:** contiene la fecha de ejecución del scrapping. Corresponde a las fechas en que el precio se recogió. Es una fecha en formato yyyy-mm-dd.
- **product:** nombre del producto. Es una cadena de caracteres.
- **product_id:** id del producto. Es una cadena de caracteres.
- **brand:** nombre de la marca del producto. Es una cadena de caracteres.
- **price:** precio del producto. Es un decimal.
- **categories:** categorías del producto. Es una lista con las categorías más generales al principio de esta.
- **unit_price:** precio por unidad del producto. Es un decimal.
- **units:** unidad en que se mide el ítem individual de producto. Es una cadena de caracteres.
- **discount:** si existe, descuento en porcentaje. Es una cadena de caracteres.

El periodo de tiempo recogido en los datos es del 18 al 22 de noviembre.

5. Propietario

Los datos utilizados para la práctica pertenecen a DIA ESPAÑA, tal y como se indica en el aviso legal del site de compra on-line de DIA:

https://www.dia.es/compra-online/aviso-legal

Tal y como se indica en el aviso legal, no se podrán publicar datos obtenidos del site sin consentimiento de DIA ESPAÑA, de manera que en el apartado “Dataset” publicaremos un dataset simulado.

DIA ESPAÑA es una cadena de supermercados española que pertenece a la multinacional DIA CORPORATE (https://diacorporate.com/), que trabaja en modo franquiciado.

Análisis anteriores:

- Según la OCU los precios suben un 15,2% en un año. Han realizado un estudio con precios en 1.180 supermercados (tanto físicos como on-line), constatando la mayor subida de precios en los últimos 34 años. (https://www.ocu.org/consumo-familia/supermercados/noticias/supermercados-mas-baratos-2022
- El observatorio de Consumo Claro de eldiario.es destaca un incremento del 62,9% en la compra de productos frescos de supermercados online. En este análisis, de pormenorizan exhaustivamente los cambios productos en los precios a nivel de categoría y con un análisis interanual y trimestral. https://www.eldiario.es/consumoclaro/ahorrar_mejor/precio-compra-frescos-supermercados-online-sube-62-9-ano-observatorio-consumoclaro_1_9570953.html

Principios éticos:

De cara a mantener una ética en la extracción de los datos y no perjudicar al propietario de los mismos, se han llevado a cabo las siguientes acciones:

- Se ha colocado un “delay” entre llamadas http de 10x el tiempo de una llamada al site para simular una navegación humana y no sobrecargar al servidor.
- Se ha analizado el archivo “robots.txt” para determinar la política del site acerca de los agentes de scraping y seguir sus recomendaciones, no accediendo a ningún contenido no recomendado.
- Se ha estudiado el aviso legal del propietario de los datos y se ha decidido no hacer público el dataset, siguiendo las instrucciones del mismo sobre copyright y publicación de datos sin el consentimiento del propietario.

6. Inspiración

Creemos que el conjunto de datos que generamos puede ser de gran interés para analizar a nivel local el impacto del IPC en los precios del supermercado. Además, la inflación no afecta a todos los productos por igual, durante los últimos meses han aparecido en los medios diversos productos especialmente afectados como el aceite de oliva y derivados del grano, fruto de la sequía y la guerra de Ucrania.

La introducción del árbol de categorías de producto nos permitiría realizar un análisis por categorías para realizar un estudio detallado al respecto. También hemos realizado scrapping sobre el precio unitario del producto, esto nos permite realizar mejores comparaciones entre diferentes versiones de un mismo producto o capturar variaciones en la cantidad de producto manteniendo el precio, que forman parte de las tácticas habituales en productos de supermercado.

En comparación con los análisis que se referencian en el apartado anterior, nuestra información está limitada a un solo supermercado. Eso sí, hemos procurado recabar el mismo tipo de información para poder realizar el mismo análisis. Es decir, la categoría del producto, el precio y precio unitario y las marcas (para diferenciar entre productos de “marca” y “marca blanca”. Será interesante realizar el análisis sobre los datos y compararlos con los obtenidos por los artículos referenciados, ya que el propio Supermercados Dia forma parte del análisis de, por lo menos, el segundo artículo.

7. Licencia

Aunque finalmente no vayamos a publicar el dataset, siguiente las instrucciones legales del propietario de los datos, la licencia elegida en el caso de hacerlo sería:

- Attribution 4.0 International (CC BY 4.0)

Esta licencia de tipo “Approved for free cultural works” permite el uso de los contenidos sin ninguna restricción; lo único que exige es referenciar a los autores de los mismos.. En nuestro caso consideramos la inflación de precios actual un gran problema y creemos que los datos deben ser compartidos para su libre uso y que puedan ser de utilidad.

Para el código del scraper, hemos elegido una licencia de código abierto:

- GNU GENERAL PUBLIC LICENSE

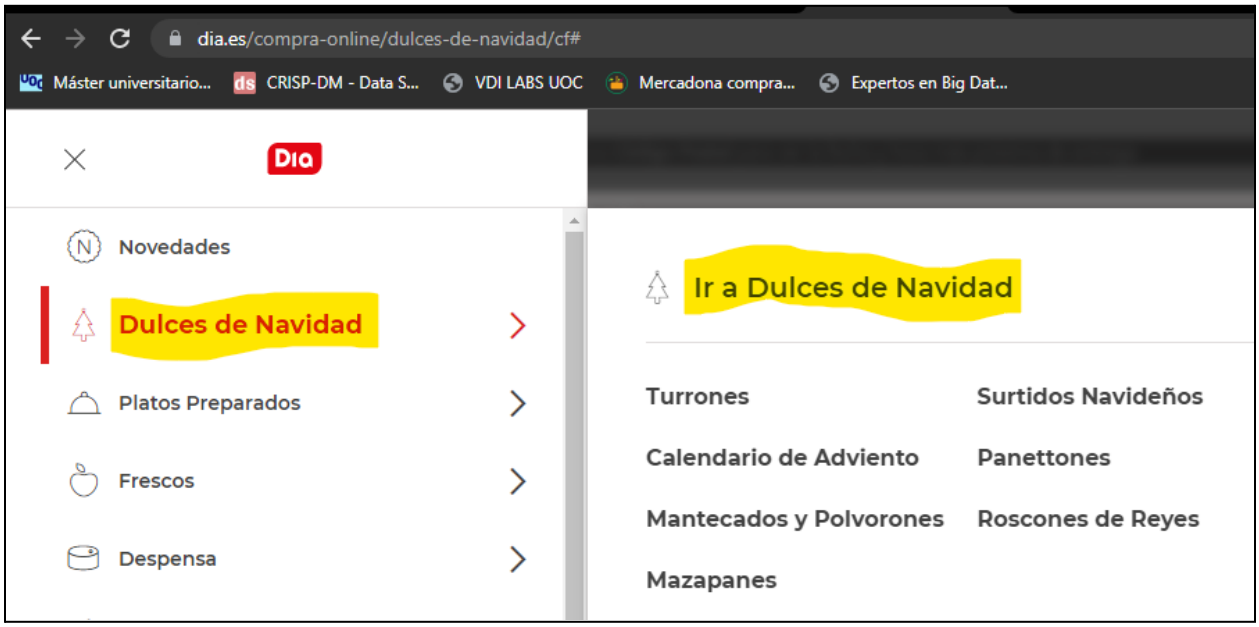
De manera que el software pueda ser utilizado libremente.

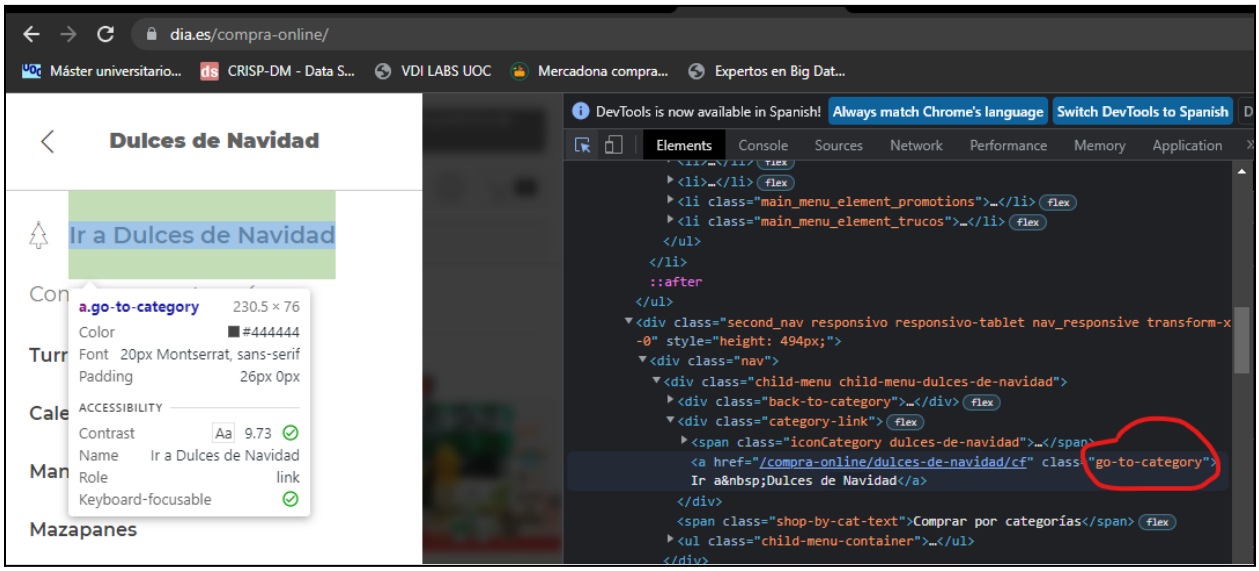
8. Código

La estructura del crawler está dividida en tres principales funcionalidades que se ejecutan de forma sucesiva, todas ellas forman parte de una misma clase DiaScraper, que se invoca desde el main. La primera realiza la navegación autónoma y se encarga de ir almacenando las urls disponibles en el site. En la segunda fase se procesa la información de cada una de las urls para, finalmente, procesarlas y almacenarlas en un único fichero csv de salida. A continuación detallamos con más profundidad la estructura de cada una de las tres fases:

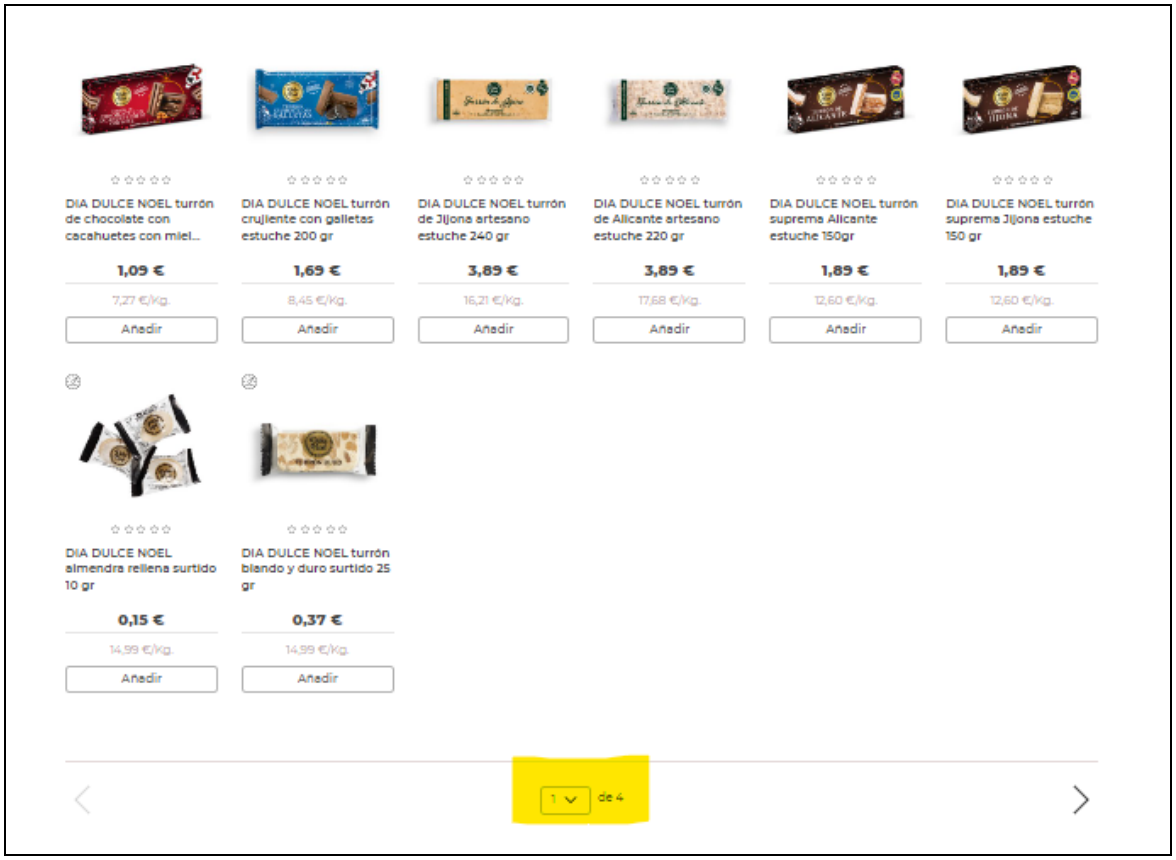
Navegación autónoma

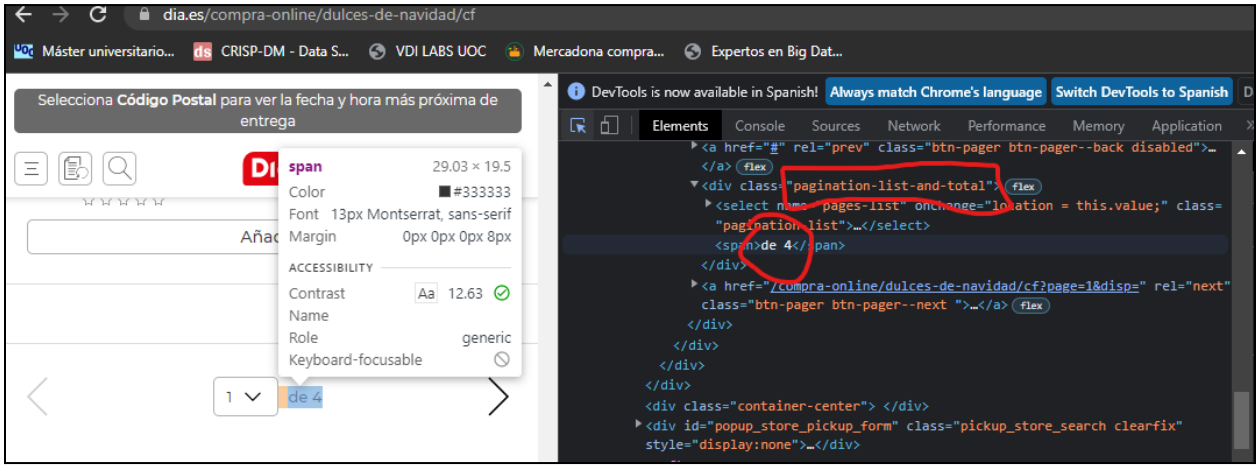
En esa fase, el programa accederá al site de DIA y buscará la lista de enlaces a las categorías principales (“Despensa”, “Bebidas”, etc.), etiquetadas con la clase “go-to-category”.





Una vez obtenidas, accederá a cada una de ellas para determinar cuántas páginas de productos hay en cada página, ya que las categorías están paginadas:



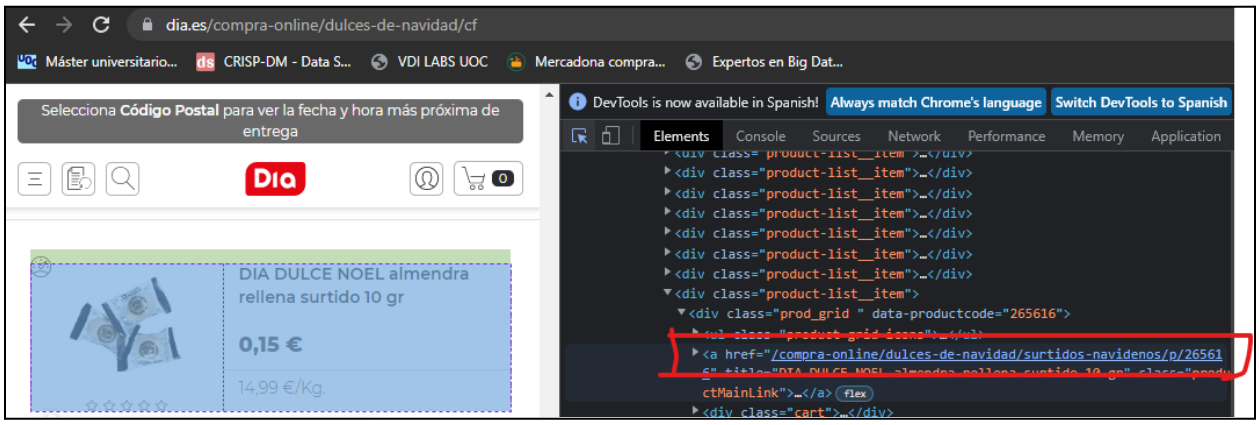


El programa buscaré el bloque de pie de página y extraerá el número de páginas que hay en la categoría, generando una lista de páginas que habrá que recorrer para escrapear los productos de la categoría:

/compra-online/frescos/cf
/compra-online/frescos/cf?page=1
/compra-online/frescos/cf?page=2
/compra-online/frescos/cf?page=3

Etc.

Una vez obtenidas todas las páginas que contienen productos de todas las categorías, entrará en cada una de ellas, para obtener las URLs de cada producto, que serán posteriormente escaneadas para obtener la información de cada producto:



https://www.dia.es/compra-online/dulces-de-navidad/panettones/p/224028
https://www.dia.es/compra-online/dulces-de-navidad/surtidos-navidenos/p/222602
https://www.dia.es/compra-online/dulces-de-navidad/mantecados-y-polvorones/p/236265
https://www.dia.es/compra-online/dulces-de-navidad/surtidos-navidenos/p/24248
https://www.dia.es/compra-online/dulces-de-navidad/mazapanes/p/25251

Por ultimo, se genera una lista de productos en CSV que se utilizará como caché para no volver a realizar este proceso a menos que se le indique al programa.

Scrapping

Una vez se han almacenado las urls a procesar, el método de scrapping realiza una petición a cada url y recoge la siguiente información a través de los métodos de la clase BeautifulSoup. El id del producto no aparece en la web y por lo tanto se obtiene de la misma url de acceso. En la imagen siguiente se muestra la información que se recoge de cada producto:
El método de scrapping incorpora control de errores para, en el caso que no se pueda recoger algún campo, sí se recoja el resto de información disponible. En algunas ocasiones podemos encontrar una página de producto vacía, en ese caso no se almacena ninguna información, se almacena la url vacía en los logs y el proceso continúa.

Una vez realizado el scrapping, la información se agrega en un diccionario y se almacena en un directorio local de forma temporal para no saturar la memoria. De este modo, en el caso de que se interrumpa el proceso por causa de una caída de red o algún cambio en la web que produzca un error (algo habitual en scrapping), la información que ya se ha recogido no se pierde y permite continuar el scrapping donde se dejó. Este proceso se repite para cada url obtenida en la sección anterior.

Guardado información

Una vez se han recorrido todas las urls se agrega toda la información que se ha ido almacenando en el directorio temporal en un DataFrame. Dado que el propósito del proyecto es almacenar la información de los precios día a día, se permite la existencia de un archivo dataset.csv en el directorio “dataset“ para agregar la información recogida en el mismo archivo. En caso de que no exista el archivo, lo genera el propio programa.

Este componente verifica que no existan entradas ni productos duplicados antes de almacenar la información. De esta forma se puede ejecutar directamente el scrapping día a día y el fichero dataset.csv va almacenando la información que se va procesando.

9. Dataset

Debido a la imposición legal del propietario de los datos, no podemos publicar los mismos. Por lo tanto, hemos publicado un dataset simulado. Se adjunta la URL:

ZENODO Doi URL: https://doi.org/10.5281/zenodo.7334808

En el aviso legal del site de DIA se indica:

https://www.dia.es/compra-online/aviso-legal

"Quedan reservados todos los derechos de Propiedad Intelectual e Industrial sobre los contenidos y/o servicios y, en particular, queda prohibido modificar, copiar, reproducir, **comunicar públicamente**, transformar o distribuir de cualquier forma la totalidad o parte de los contenidos y/o servicio incluidos en el Portal, para propósitos públicos o comerciales**, si no se cuenta con la autorización previa, expresa y por escrito de DIA ESPAÑA** o, en su caso, del titular de los derechos correspondientes."

El dataset final privado, generado con el scraper es este:

https://drive.google.com/file/d/1HOAYLUxMIIbCPm_8CVwrRuL50ggzGWyn/view?usp=share_link

10. Video

Enlace al vídeo:
https://drive.google.com/file/d/1m6VW0_t3ZkuPBH-UbuDek3B1YaCv2UUW/view?usp=sharing

11. Contribuciones

Contribuciones	Firma
Investigación previa	CPH, OPM
Redacción de las respuestas	CPH, OPM
Desarrollo del código	CPH, OPM
Participación en el vídeo	CPH, OPM