**TRABAJO PRÁCTICO DE LA MATERIA PROYECTO INTEGRADOR**

**INFORME**

**LA EXTRACCIÓN DE DATOS MEDIANTE LA METODOLOGÍA DEL WEBSCRAPPING**

**Grupo PI ALFA\_CBBCCC4**

**Bustos, Joni DNI**

**Capiletti, Daniela DNI**

**Catalano P., Diana C. DNI 95701783**

**Fran DNI**

**PROFESORES: Silvia**

**Noviembre, 2022**

**Introducción**

En la actualidad, el web scraping es una técnica muy utilizada y consiste en la extracción de datos significativos de una o varias páginas web, para el análisis o manipulación de los datos, lo cual es de gran importancia para monitorear precios de la competencia, comparación de precios en tiendas, recolectar o detectar cambios en una web, analizar enlaces de una web para buscar links, etc.

Como señalamos anteriormente, el web scraping se enfoca en la transformación de los datos sin estructura (como el formato HTML) en datos estructurados que se pueden almacenar y analizar en una base de datos creada hasta en una hoja de cálculo o cualquier fuente de almacenamiento

Algunas limitaciones del web scraping es que al automatizar un programa para extraer datos de la web y si esta cambia de un día para otro, es posible que dicho programa genere errores, por lo cuál se debe tener esto pendiente.

En el presente informe se mostrará la metodología seguida por el Grupo PI ALFA\_CBBCCC4 en cuánto al scraping realizado a la página www.mercadolibre.com.arg en el rubro de aires acondicionados, posteriormente graficados e interpretados.

**Metodología**

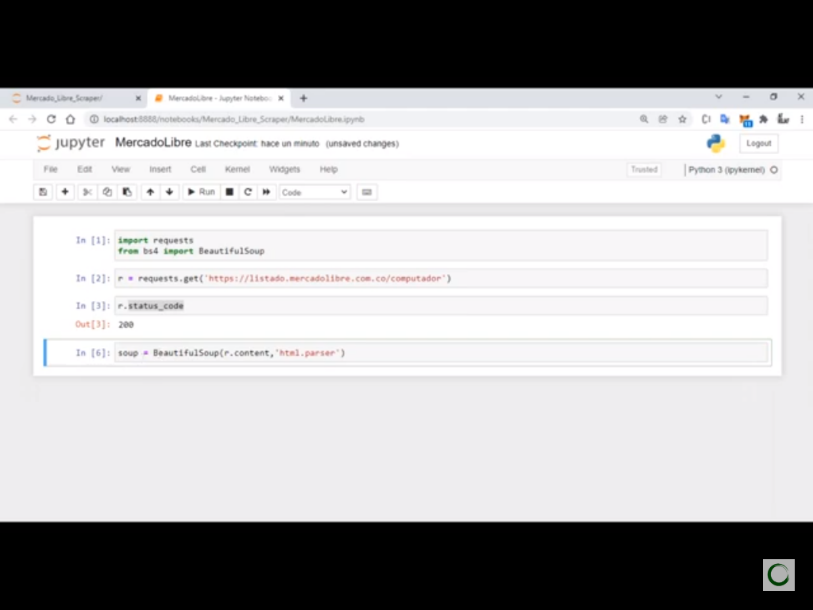
Inicialmente, el grupo se reunió vía zoom para seleccionar el tema a escoger según la consigna dada por los profesores de la materia Proyecto Integrador. Cada integrante hizo una propuesta de tema a desarrollar y quedó seleccionado para el análisis de datos y las comparaciones de los precios de diferentes modelos y marcas de aires acondicionados vendidos por la empresa Mercado Libre Argentina, visto desde el punto de vista de alternativas para la efectivización de la compra.

Para iniciar el trabajo, abrimos cuenta en Trello con sus respectivas tarjetas de acuerdo a la prioridad del proyecto. Y, para que todos los integrantes del grupo tuviésemos acceso a los archivos e ir subiendo información, utilizamos el repositorio de Git y Github.

Paralelamente empezamos a indagar las páginas de Mercado Libre Argentina en el rubro de aires acondicionados para conocer su estructura y principales características, de manera que pudiésemos realizar un trabajo según la consigna dada. La plataforma de Mercado Libre (ML) muestra los productos de profesionales especializados, de pequeñas tiendas y grandes marcas a través de categorías, la publicación muestra las imágenes, precio y la descripción del producto. También los medios de pago que sean convenientes al comprador y el tipo o medio de transporte a utilizar.

Una vez entendida la estructura de la página, Como se muestra en la figura 1, copiamos la URL al Jupyter y mediante un import request hicimos la solicitud de tipo Get para que la página devolviera el código HTML. Seguidamente, desde bs4 BeautifulSoup para posteriormente hacer el scraping del HTML e ir elemento por elemento de lo que deseábamos encontrar en la página de ML. Se hizo la respectiva comprobación con status\_code que iba todo bien.

**Figura 1. URL e import request del HTML de ML.**



Luego se definió la variable ‘aire acondicionado’ para al buscar por consola / teclado nos arrojara el contenido. Realizamos el arreglo pertinente (array) para contener todos los productos que deseábamos (nombre y precio del producto con su respectiva URL). Además, se importó la librería Pandas que nos permitió exportar los datos requeridos y se guardaron en formato CSV. Se realizó el scraping a 40 páginas con un contenido de 50 productos cada una.

Para el análisis de los datos consideramos las variables tipo de aire acondicionado según la publicación y precio promedio según el título del artículo. Para lo cual utilizamos herramientas estadísticas que permitieron realizar una comparativa de precios y de los tipos de aires acondicionados sin considerar marca o modelo porque no lo consideramos relevante.

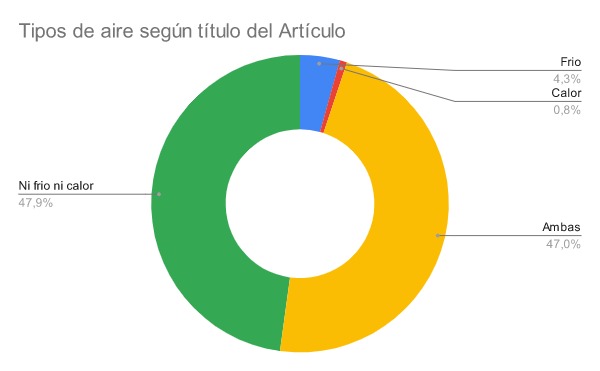
**Análisis de los datos obtenidos de ML**

1. **Tipos de aire acondicionado según título del artículo**

Estos datos fueron de los más sencillos, como se señaló el scraping se realizó en 40 páginas con un total de 50 productos por página. Como puede observarse en la tabla 1 se obtuvieron 4 tipos de aires acondicionados según el título del artículo ofrecido frío (4.3%), calor (0.8), frío-calor (47.9%) y ambos ($&%). Posiblemente estos resultados obedezcan al cambio de estación en Argentina en el que aún el calor no es muy fuerte (al momento del scraping estábamos en primavera).

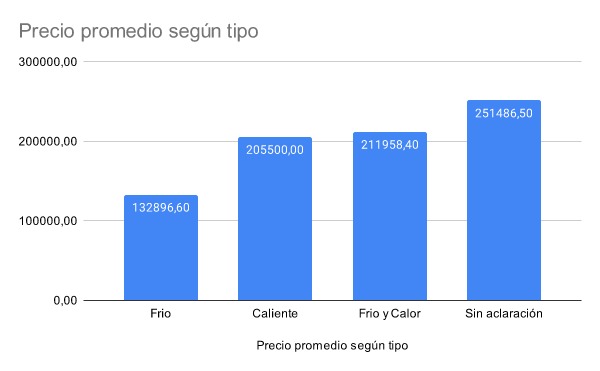
**Tabla 1.** **Tipos de aire acondicionado según título del artículo**

|  |  |
| --- | --- |
| **Tipo aire acondicionado** | **Según título del artículo (%)** |
| Frío | 4.3 |
| Calor | 0.8 |
| Ni frío ni calor | 47.9 |
| Ambos | 47.0 |
| **Total 100.0** | |



1. **Precio promedio de los aires acondicionados según tipo**

Según el scraping realizado a las páginas de productos aire acondicionados de ML el precio de estos equipos estaban ubicados en un rango entre $132000 y $300000, siendo la categoría Frío más baja en precios promedio de $ 132896 y los aires de un mayor precio promedio son los denominados “sin aclaración” .



**CONCLUSIONES**

A la hora de adquirir un equipo frío / calor hay que tener en cuenta no solamente el precio sino también el consumo promedio que tendrá el mismo, ya que este impactará en la factura de electricidad, es por ello que es imperativo leer la etiqueta de eficiencia energética, definir el tipo de tecnología es más conveniente y el tamaño del espacio donde se colocará el aparato.

Razón por la que, nos pareció importante considerar también, una pequeña información obtenida de la página https://www.argentina.gob.ar/enre/uso-eficiente-y-seguro/consumo-basico-electrodomesticos, en donde se obtienen los consumos básicos promedio, según la potencia y tipo de tecnología, que no debería de ser un dato menor a la hora de elegir la mejor opción de compra en función de los costos y beneficios.

