Tipología y ciclo de vida de los datos

PRÁCTICA 1: Web scrapping

Autores: Gabriel Álvarez Morgado y Héctor Castillo Jeria

Marzo 2022

Contents

1. Contexto.

En Chile, últimamente, las farmacias han concentrado su fuerza de venta en tres grandes cadenas: Farmacias Cruz Verde, Farmacias Salco Brand y Farmacias Ahumada. La publicidad de estas cadenas muestra un número reducido de productos con sus precios de promoción, pero no existe una forma de comparar los precios antes de comprar.

Por otro lado, desde una mirada estatal, no existe una forma transparente que permita realizar el control de prevención de colusión entre estas cadenas farmacéuticas.

Es por eso que hemos elegido crear un proyecto que recolecte el precio de los medicamentos y artículos de perfumería que las tres principales cadenas de farmacias ofrecen. Esta información se encuentra disponible en las tres páginas web de estas cadenas **Cruz Verde**, **Salco Brand** y **Ahumada**.

En estas páginas, las farmacias presentan su productos con los respectivos precios normales y precios en oferta, sobre los que hemos aplicado nuestro *web scrapping*.

Potenciales utilidades

- Obtener información comparativa de precios sobre los productos de las diferentes cadenas farmacéuticas.
- Análisis estadístico descriptivo del sector farmacéutico.
- Acceso a información valiosa, consolidada y actualizada de los precios de los medicamentos y artículos de perfumería que ofrecen las farmacias.

2. Definir un título para el dataset.

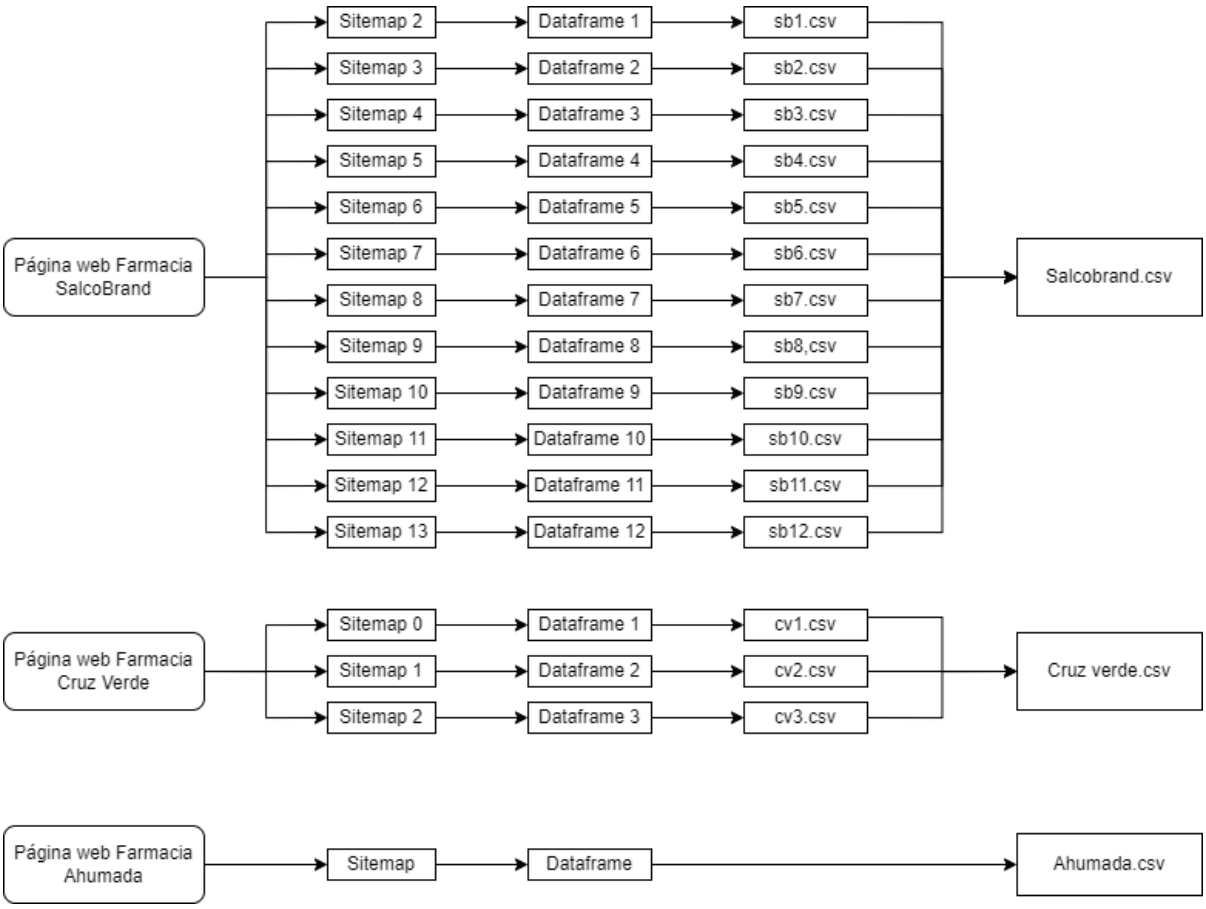
Hemos seleccionado *ProdFarmaCl* como el nombre de nuestro dataset, pues presenta el precio de los productos ofrecidos por las farmacias en Chile.

3. Descripción del dataset.

El conjunto de datos está compuesto por 3 archivos *SalcoBrand.csv*, *CruzVerde.csv* y *Ahumada.csv*, los que recogen información sobre el nombre de los productos y precios ofrecidos por estas tres cadenas de farmacia en Chile. En total, el dataset consta de 33.701 registros

4. Representación gráfica.

El mapa del sitio de la farmacia SalcoBrand consta de 15 submapas, de los cuales, desde el sitemap 2 al 13 incluyen enlaces a los artículos que la famrmacia ofrece. Los Enlaces de estos 12 sitemaps fueron visitados, recolectando la información descrita en el apartado anterior y almacenándola en dataframes. Cada dataframe fue exportado a un archivo csv para posteriormente unir todos los csv en uno solo, tal como muestra la siguiente figura.



En la figura anterior también se describe la manera en que fue recolectada la información en las otras dos cadenas de farmacias: En el caso de Cruz verde, la información de los productos se dispone en 3 submapas del sitio, donde se operó de manera muy similar que lo descrito anteriormente. En Farmacias Ahumada, por el contrario, la página dispone de un solo sitemap desde donde fueron visitados los enlaces de los productos para obtener su información.

5. Contenido.

Tal como se explicó anteriormente, el dataset incluye 3 archivos csv los que en total componen 33.701 registros observados y recolectados durante la semana del 04 al 10 de abril de 2022. La manera en que fue recolectada esta información fue descrita en el apartado número 4.

Cada uno de estos registros se corresponde a un producto ofertado por alguna cadena de farmacia. Cada registro está conformado por 5 variables diferentes indicadas a continuación:

variable	**descripción**
Farmacia	Identificador de la farmacia.
Producto	Nombre del producto.
SKU	Código del producto en la farmacia.
Normal	Precio normal del producto.
Oferta	Precio oferta del producto.

La estructura del sitio de la farmacia SalcoBrand y la farmacia Cruz Verde son bastante similares, por lo que se generaron funciones bastantes parecidas para rescatar los datos de ambas páginas. Estas funciones

fueron escritas en un archivo aparte.

El script principal importa las funciones del archivo anterior, además de otras librerias, y ejecuta comandos que permiten recolectar la información de los sitios web de ambas farmacias y los guarda en archivo.csv

En la página de Cruz verde, la información es cargada a través de Javascript, lo que no permite rescatarla a través del comando request. Para sortear este problema, se utilizó la librería Selenium.

La estructura del sitio de Farmacias Ahumada es bastante distinta a las otras dos páginas web. Por este motivo, se decidió generar un script secundario que recogiera la información de los productos ofrecidos por esta cadena.

Los archivos csv generados contienen la información en bruto de los sitios web. Es necesario, en una segunda etapa, que al dataset se le aplique una limpieza y técnicas de preprocesamiento de datos, con el objetivo de obtener una base de datos óptima para ser explotada.

6. Agradecimientos.

Agradecemos el trabajo de los equipos TI de cada cadena farmacéutica por mantener actualizados los precios de sus productos en sus respectivas páginas web y por definir políticas en su archivos robots.txt que permiten realizar este tipo de proyectos.

Cabe destacar que al extraer la información, hemos aplicado las buenas prácticas recogidas en el material de la asignatura:

- Visualización del archivo *robots.txt* para verificar que el propietario de los datos permite el acceso a la información.
- Revisión del *Sitemap* indicado en el archivo *robots.txt*, para analizar la estructura e identificar la ubicación de la información deseada.
- Análisis del tamaño y tecnología de la página web y evaluación de la cantidad total de información a recoger.
- Utilización de libreria BeautifulSoup para realizar un parseado de las páginas en forma automática, evitando errores que pueden suceder al parsear en forma manual.
- No saturar de peticiones el servidor web. Para esto, establecimos un tiempo de espera de al menos 3 segundos entre peticiones.
- La extracción del contenido de las variables fue recolectada directamente con la herramienta de Web Scraping.

No se han incluido citas sobre análisis anteriores, pues no hemos encontrado poryectos similares basados en web scrapping. De todas maneras, agradecemos el esfuerzo del Servicio Nacional del Consumidor, quienes hasta el año 2019 mantuvieron un servicio web de comparador de precios de medicamentos. La falta de mantención de este servicio ha sido un elemento motivador para generar este proyecto.

7. Inspiración.

El aumento de precios de los fármacos, el relajo del Estado en el control de precios, la aparición de farmacias populares con medicamentos más baratos que las grandes cadenas y una historia de colusión perpretada por las cadenas de farmacia alrededor del año 2010 han sido elementos que nos han motivado e inspirado para encontrar una solución que pueda aliviar la economía familiar de las personas que deben invertir cuantiosas sumas de dinero en comprar medicamentos.

Por este motivo, hemos decidido proveer una herramienta que semanalmente pueda observar la variación de los precios de los productos farmacéuticos, entregando a la comunidad información de los precios de las grandes cadenas, de forma independiente, transparente y actualizada.

8. Licencia.

Hemos creado este proyecto bajo licencia GNU General Public License v3.0. Creemos en la creación de comunidades en torno al desarrollo de código abierto y software libre. Al publicar nuestro trabajo bajo licencia GNU-GPL permitimos que otras personas tengan la posibilidad de utilizar y mejorar nuestro trabajo en beneficio de la comunidad, evitando que este derecho sea restringido por terceras personas.

9. Código.

El código utilizado para la extracción de la información requerida mediante Web Scrapping se encuentra en este repositorio Git.

Se sugiere leer el archivo README.md que ofrece una descripción del contenido y estructura del repositorio.

Tal como se mencionó anteriormente, para rescatar la información de las páginas web de la farmacia Cruz Verde, fue necesario utilizar la librería Selenium. Para ejecutar correctamente esa parte del script, es necesario contar con el navegador Firefox y el driver Geckodriver.

Este último debe estar dispuesto en la carpeta *bin* (Linux) o *scripts* (Windows) dentro de la carpeta Venv creada para el environment.

10. Dataset.

El dataset se ha publicado en Zenodo bajo el DOI: 10.5281/zenodo.6441879 con la siguiente descripción: *The dataset contains information extracted from three pharmacy websites, on different medical products, with the intention of analyzing potential collusion issues and allowing price comparisons.*

Además, se puede acceder a ellos en la carpeta **data** del repositorio GitHub, en los siguientes enlaces:

Cruz Verde. Salco Brand. Ahumada.

11. Video

Un video explicativo de la práctica ha sido grabado y compartido para su reproducción en el siguiente enlace

Tabla contribuciones

Contribuciones	**Firma**
Investigación Previa	Gabriel Álvarez, Héctor Castillo
Redacción de las respuestas	Gabriel Álvarez, Héctor Castillo
Desarrollo del código	Gabriel Álvarez, Héctor Castillo