TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS PRÁCTICA 1

PRESENTADO POR: *ANDRES RICARDO SANABRIA GARAY*

Objetivos

- ✓ Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinarios.
- ✓ Saber identificar los datos relevantes cuyo tratamiento aporta valor a una empresa y la identificación de nuevos proyectos analíticos.
- ✓ Saber identificar los datos relevantes para llevar a cabo un proyecto analítico.
- ✓ Capturar datos de diferentes fuentes de datos (tales como redes sociales, web de datos o repositorios).
- ✓ Actuar según los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- ✓ Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Desarrollo de la Práctica

1. Contexto

Vivimos en un mundo que en los últimos años ha presentado cambios significativos para la humanidad, por mencionar solo algunos, podemos referirnos al cambio climático, la pandemia causada por el virus del COVID-19, guerras internacionales y crisis económica. Estos acontecimientos trascienden fronteras y afectan en diferente medida a cada país.

Pero ¿Qué impacto tienen estos acontecimientos sobre la producción y el precio de los alimentos en Colombia?, ¿Cómo afecta esta situación a las personas más vulnerables o con menor capacidad adquisitiva en la capital de este país? Un primer enfoque supone el análisis de la variación del costo de los alimentos.

Para este ejercicio práctico se utilizará como fuente de información el boletín de precios publicado por la Central de Abastos de Bogotá, una plaza de mercado situada en el sur de la ciudad que es la más grande de Colombia y la segunda más grande de Latinoamérica[1]

La URL del sitio web desde el que se tomará la información es:

http://boletin.precioscorabastos.com.co/

Es importante mencionar que este boletín se publica en formato PDF de lunes a viernes excepto los festivos y el sitio cuenta con boletines del último año.

En esta práctica descargaremos estos archivos en formato PDF y extraeremos su contenido para crear un dataset.

2. Título

El título de este conjunto de datos es:

pricefoodbogota22.csv

3. Descripción del dataset

Este dataset está compuesto por alimentos de categorías como verduras, frutas, tubérculos, plátanos, granos y procesados, lácteos, cárnicos y huevos, que se han comercializado en la plaza de mercado de Corabastos en la ciudad de Bogotá D.C. en Colombia durante el último año (noviembre de 2021 - noviembre de 2022).

Los alimentos provienen de diversas regiones del país y al llegar a este lugar que se utiliza como punto principal de distribución en la ciudad, se comercializan al por mayor y al detal.

Muchos de los negocios que venden alimentos en la capital colombiana y sus alrededores adquieren sus productos en este lugar.

4. Representación Gráfica

La Figura 1 presenta un diagrama del proyecto que se adelanta con esta práctica

1 Corabastos. (2022, 8 de agosto). Wikipedia, La enciclopedia libre. Fecha de consulta: 02:44, noviembre 10, 2022 desde https://es.wikipedia.org/w/index.php?title=Corabastos&oldid=145249669.

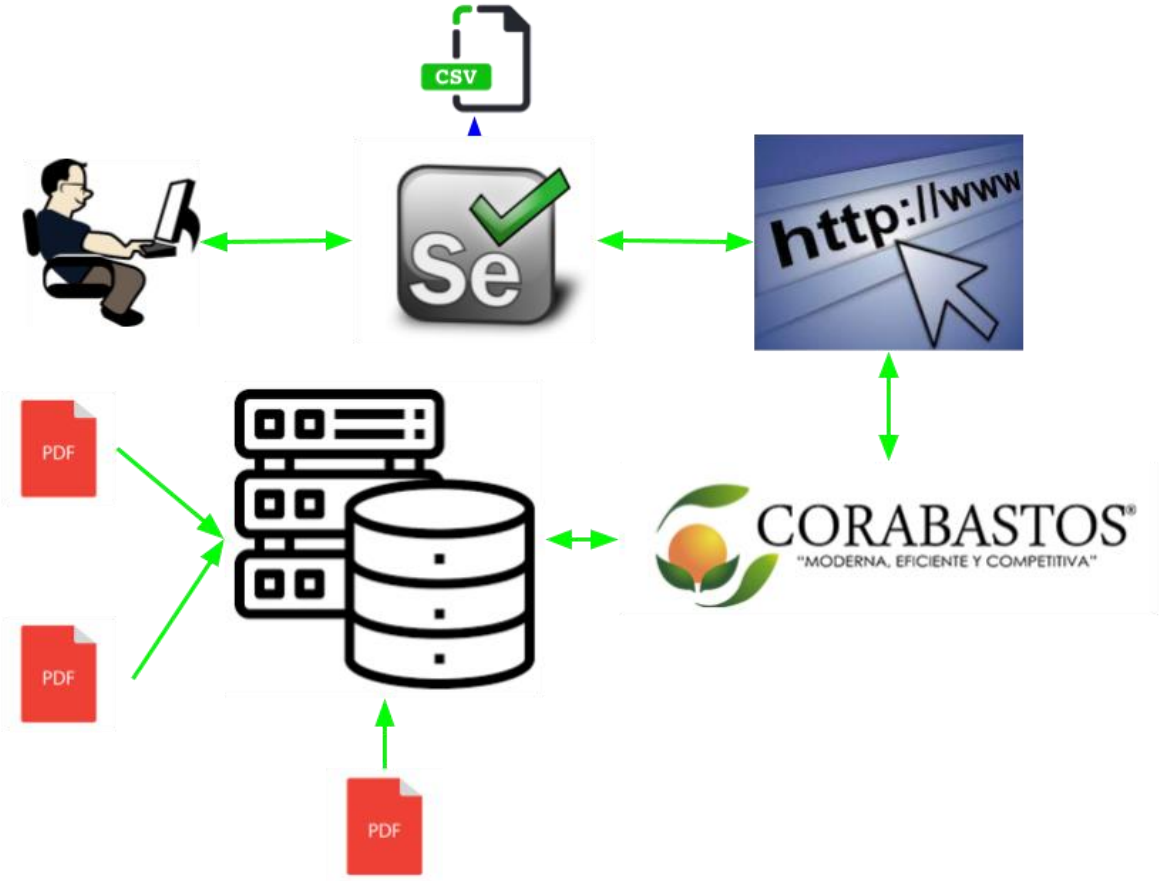

Figura 1. Diagrama del proyecto de Web Scraping de la Práctica 1

Fuente: Elaboración propia

En la Figura 2 se observa una muestra del boletín de precios en formato PDF, a partir del cual se elabora el dataset.

CORABASTOS

miercoles 2 de noviembre de 2022

El presente boletín de precios corresponde a productos de calidad extra y primera en los puntos de venta de la Corporación de Abastos de Bogotá. S.A. Corabastos. La variación en los precios obedece a la oferta y la demanda.

Nombre	Presentación	Cantidad	Unidad	$ Cal. Extra	$ Cal. Primera	Valor x Kilo
ACELGA	ATADO	10,00	KILO	$ 15.000	$ 13.000	$ 1.500
AHUYAMA	KILO	1,00	KILO	$ 2.200	$ 2.000	$ 2.200
AJO ROSADO	ATADO	5,00	KILO	$ 65.000	$ 60.000	$ 13.000
ALCACHOFA	DOCENA	10,00	KILO	$ 40.000	$ 38.000	$ 4.000
APIO	ATADO	10,00	KILO	$ 15.000	$ 13.000	$ 1.500
ARVEJA VERDE	BULTO	50,00	KILO	$ 300.000	$ 295.000	$ 6.000
BERENJENA	KILO	1,00	KILO	$ 2.000	$ 1.900	$ 2.000
BROCOLI	DOCENA	10,00	KILO	$ 40.000	$ 38.000	$ 4.000
CALABACIN	KILO	1,00	KILO	$ 1.000	$ 900	$ 1.000
CALABAZA	KILO	1,00	KILO	$ 1.000	$ 900	$ 1.000
CEBOLLA CABEZONA BLANCA	BULTO	50,00	KILO	$ 140.000	$ 135.000	$ 2.800
CEBOLLA CABEZONA ROJA	BULTO	50,00	KILO	$ 140.000	$ 135.000	$ 2.800
CEBOLLA LARGA	ROLLO	25,00	KILO	$ 65.000	$ 60.000	$ 1.300
CILANTRO	ATADO	10,00	KILO	$ 20.000	$ 18.000	$ 2.000

Figura 2. Muestra del boletín de precios en PDF.

Fuente: http://boletin.precioscorabastos.com.co/wp-content/uploads/2022/11/BOLETIN-DE-PRECIOS-02noviembre2022.pdf

5. Contenido

Este juego de datos se compone de 3200 observaciones y 8 variables en un periodo de tiempo de cerca de un año (noviembre 2021 – noviembre 2022). A continuación, la descripción de cada variable:

date:	Fecha a la que corresponde el precio del alimento
foodname:	Nombre del alimento
salespres:	Se trata de la forma en la que se presenta el alimento para su comercialización
amount:	La forma en la que se agrupa el alimento para su venta
measure:	Unidad de medida en la que se ofrece el alimento (kilo)
ehquality:	Comida de calidad extra, entiéndase como alta calidad
hquality:	Comida de primera calidad
cost:	Costo del alimento por kilo expresado en pesos colombianos

6. Propietario

El sitio web elegido aparece registrado a nombre de la CORPORACION DE ABASTOS DE BOGOTA SA.

La Corporación de Abastos de Bogotá S.A.- CORABASTOS, es una Sociedad del orden nacional, de economía mixta vinculada al Ministerio de Agricultura y Desarrollo Rural, la Gobernación de Cundinamarca y la Alcaldía de Bogotá, entidades que forman parte de los accionistas del sector oficial con un 47.92% del total de las acciones, correspondiendo el 52.08% al sector del comercio[2].

Los pasos que se siguieron para actuar de acuerdo con los principios éticos y legales en el contexto del proyecto implicaron la revisión detallada del sitio web en busca del apartado de Términos y Condiciones. Sin embargo, el sitio web que se relacionan con esta sociedad no cuentan con este item, pero si con una marca de Copyright © 2022 Precios Corabastos Powered by Corabastos S.A. que aparece en la parte inferior central de la página y que básicamente representa derechos de autor sobre la imagen y el contenido publicado en el sitio web.

Tampoco existe el archivo robots.txt lo que supone que no hay restricción para hacer scraping sobre este sitio web.

2 Corabastos. (12 de noviembre de 2022). Nuestra Historia. https://corabastos.com.co/inicio/nosotros/

7. Inspiración

Colombia desde hace varios meses enfrenta una fuerte ola invernal que ha provocado la pérdida de vidas humanas, animales, cosechas y derrumbes de vías en diferentes partes del país.

Lo anterior sumado a la escasez de productos agroquímicos por la guerra entre Rusia y Ucrania, una inflación que en octubre fue del 12.22%[3] y una fuerte devaluación del peso colombiano que para el 25 de octubre de este año ya rondaba el 24.8%[4] han traído serias consecuencias sobre la canasta familiar de los hogares colombianos, encareciendo los alimentos en un país que a nivel económico aún no se termina de recuperar de los efectos de la pandemia por el COVID-19.

Por esta razón y con la amenaza de la desnutrición de la población más vulnerable o que vive en pobreza nos planteamos la siguiente pregunta ¿Cuál ha sido la variación que ha presentado el precio de los alimentos de la canasta familiar durante el último año en la ciudad de Bogotá D.C.?

El resultado de esta práctica es la construcción de un dataset a partir del boletín de precios publicado por La Corporación de Abastos de Bogotá S.A.- CORABASTOS, con el cual se intentará responder esta pregunta.

Es importante mencionar que CORABASTOS con la publicación de un boletín diario en el que se fijan los precios de los principales productos agroalimentarios del país contribuye en orientar de manera adecuada las operaciones comerciales en la capital e influenciar las de toda Colombia.

8. Licencia

Para el dataset generado se ha escogido la licencia: **Atribución, Compartir igual: (CC-BY-SA v4.0)**

Las razones por las cuales se escoge esta licencia de Creative Commons son las siguientes:

3 Portafolio. (15 de noviembre de 2022). Inflación anual en Colombia fue de 12,22% en octubre. https://www.portafolio.co/economia/inflacion-anual-en-colombia-octubre-2022-573688

4 La República. (15 de noviembre de 2022). La devaluación de la TRM durante 2022 es la tercera más alta de este siglo XXI. https://rb.gy/jso23z

- ✓ Los divulgadores de datos abiertos deben proporcionar un acceso fácil a la licencia para todos los conjuntos de datos a los que se pueda acceder, que puedan utilizarse y compartirse[5].

- ✓ Se da libertad de compartir – copiar y redistribuir el material en cualquier medio o formato así como de adaptar — remezclar, transformar y construir a partir del material para cualquier propósito, incluso comercialmente.

- ✓ La libertad para compartir y adaptar se da bajo los términos siguientes:

> *"Atribución — Usted debe dar crédito de manera adecuada, brindar un enlace a la licencia, e indicar si se han realizado cambios. Puede hacerlo en cualquier forma razonable, pero no de forma tal que sugiera que usted o su uso tienen el apoyo de la licenciante.*
>
> *CompartirIgual — Si remezcla, transforma o crea a partir del material, debe distribuir su contribución bajo la misma licencia del original"[6].*

Estas características permiten que el conjunto de datos generado se ubique en la categoría de datos abiertos, ideal para la publicación en una plataforma como Zenodo.

9. Código

Esta práctica de Web Scraping se desarrolló en Python trabajando en conjunto con una serie de librerías como Selenium.

En la primera parte, el programa accede al sitio web del boletín de precios de Corabastos S.A., este es un sitio dinámico que emplea tecnología de Javascript (razón por la cual utilizamos Selenium), y luego busca los elementos que contienen los enlaces al boletín de precios que se publican los días no feriados, desde allí descarga el enlace al respectivo pdf. Estos enlaces se almacenan en una lista. Para hacer esto es importante destacar que se utilizan condiciones esperadas como click para acceder a dichos elementos.

Con la lista de las URL, se procede a realizar la descarga de los respectivos archivos, que constituyen la base para la creación del dataset.

El código de esta práctica se encuentra disponible en:

https://github.com/Sanabriaga/pra1_topologia_y_ciclo_de_vida_uoc.git

5 Data Europa EU. (19 de noviembre de 2022). Como encontrar una licencia. https://data.europa.eu/elearning/es/module4/#/id/co-01
6 Creative Commons. (19 de noviembre de 2022). Atribución-CompartirIgual 4.0 Internacional (CC BY-SA 4.0). https://creativecommons.org/licenses/by-sa/4.0/deed.es