PRACTICA WEB SCRAPING
Sandra Milena Patiño Avella

1.Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

Corabastos es una plaza de abastos situada al occidente de Bogotá. Es una de las mayores plazas mayoristas de la región. La administra la Corporación de Abastos de Bogotá S.A.

Allí es donde se comercializa al por mayor los productos provenientes de todas las zonas de Colombia siendo un referente para el establecimiento de los precios de comercialización.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

Histórico de precios de productos zona Boyacá comercializados en Corabastos.

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

Conjunto de datos que describe los precios en un determinado rango de días de los productos comercializados desde el departamento de Boyacá en la central mayorista de Corabastos.

4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente

5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

- Fecha: Fecha sobre la que se hace la consulta
- Nombre: Nombre de producto a consultar
- Presentacion: Presentación en la que se vende el producto
- Cantidad: Cantidad de acuerdo a la unidad y presentación
- Unidad: Medida de presentación del producto
- PesosCalidadExtra: Valor en pesos del producto cuando su calidad es muy alta.
- PrecioCalidadPrimera: Valor en pesos del producto cuando su calidad es alta.
- PrecioCalidadCorriente: Valor en pesos del producto cuando su calidad es normal.
- GrandesSuperficies: Precio de venta a almacenes de cadena y grandes superficies.

El periodo de tiempo está definido por los argumentos que se pasan al script en Python, siendo el intervalo mayor o igual a 1 día.

Los datos han sido recolectados desde la página de Corabastos mediante técnicas de web scraping y haciendo uso del lenguaje de programación Python.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Se agradece a la Corporación de Abastos de Bogotá S.A. CORABASTOS

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

La principal actividad económica del departamento de Boyacá es la ganadería y la agricultura, en especial el cultivo de tubérculos y hortalizas, los cuales en su gran mayoría son comercializados en la mayor central de abastos de Colombia llamada CORABASTOS. Por esto es importante el comportamiento histórico de los precios de los diferentes productos lo cual servirá como referente para identificar las mejores épocas para vender los productos y así lograr una mayor utilidad.

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

- **Released Under CC0: Public Domain License**
- **Released Under CC BY-NC-SA 4.0 License**
- **Released Under CC BY-SA 4.0 License**
- **Database released under Open Database License, individual contents under Database Contents License**
- **Other (specified above)**
- **Unknown License**

La licencia escogida es la Released Under CC BY-SA 4.0 License dado que se puede:
Compartir — copiar y redistribuir el material en cualquier medio o formato
Adaptar — remezclar, transformar y construir a partir del material para cualquier propósito, incluso comercialmente.

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

El código fuente con el que se ha generado el dataset se encuentra disponible en el repositorio GitHub https://github.com/smpavella/productos/

10. Dataset. Publicación del dataset en formato CSV en Zenodo con una pequeña descripción.

https://zenodo.org/record/3757558#.Xp00oZl7nIU

11. Entrega. Presentar el trabajo con el DOI del dataset en Github.

DOI: 10.5281/zenodo.3757558

https://doi.org/10.5281/zenodo.3757558

ANEXO

Pruebas previas al proceso de web scraping.

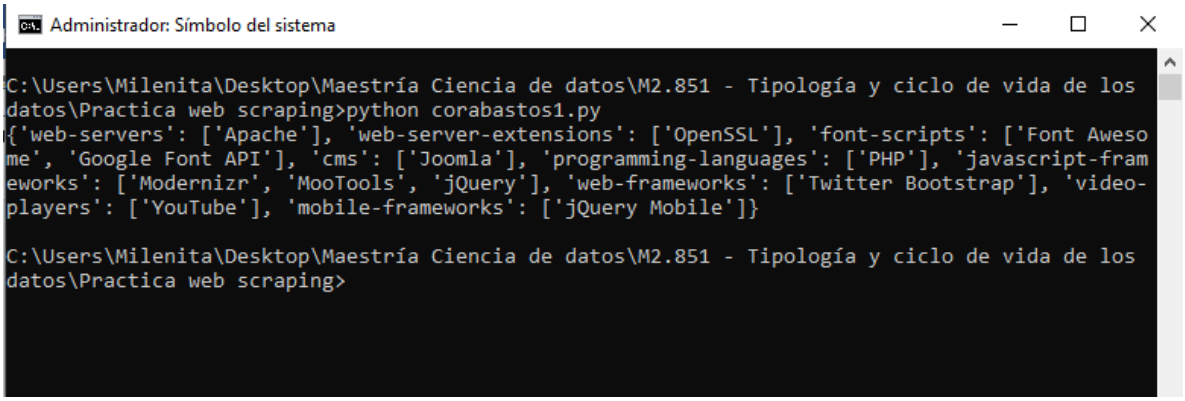
Checking robots.txt

```
# If the Joomla site is installed within a folder
# eg www.example.com/joomla/ then the robots.txt file
# MUST be moved to the site root
# eg www.example.com/robots.txt
# AND the joomla folder name MUST be prefixed to all of the
# paths.
# eg the Disallow rule for the /administrator/ folder MUST
# be changed to read
# Disallow: /joomla/administrator/
#
# For more information about the robots.txt standard, see:
# http://www.robotstxt.org/orig.html
#
# For syntax checking, see:
# http://tool.motoricerca.info/robots-checker.phtml

User-agent: *
Disallow: /administrator/
Disallow: /bin/
Disallow: /cache/
Disallow: /cli/
Disallow: /components/
Disallow: /includes/
Disallow: /installation/
Disallow: /language/
Disallow: /layouts/
Disallow: /libraries/
Disallow: /logs/
Disallow: /modules/
Disallow: /plugins/
Disallow: /tmp/
```

Examining the Sitemap

Identifying the technology used by a website



Finding the owner of a website