

Práctica 1 (35% nota final)

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes por un proyecto analítico y usar las herramientas de extracción de datos. Para hacer esta práctica tendréis que trabajar en grupos 2 personas. Tendréis que entregar un solo fichero con el enlace Github (<https://github.com>) donde haya las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos de vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github. Podéis mirar estos ejemplos como guía:

- Ejemplo: <https://github.com/rafoelhonrado/foodPriceScraper>
- Ejemplo complejo: <https://github.com/tteguayco/Web-scraping>

Competencias

En esta PEC se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para resolverlo.
- Capacidad para aplicar las técnicas específicas de web scraping.

Objetivos

Los objetivos concretos de esta práctica son:

- 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 que su 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) y mediante diferentes mecanismos (tales como queries, API y scraping).
- Actuar con 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.

Descripción de la Práctica a realizar

Este proyecto ha sido desarrollado y elaborado por:

- Mónica Alexandra Gómez Martínez
- David Francisco Dávila Ortega

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

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

En el actual proyecto, se ha decidido trabajar sobre la página web de Soriana', la cual es una cadena mexicana de supermercados y almacenes. Dado que en la actualidad se tiene que la cadena Soriana está compitiendo en el mercado principalmente con otras empresas importantes de *retail* como lo son Walmart, Chedraui, Casa Ley y H-E-B, es de gran utilidad el realizar una recolección de información sobre los productos que se venden en dicho establecimiento, la información concerniente a costos ya sea este su precio original como el precio en rebajas (de tenerlo). Este trabajo representa un principal esfuerzo para realizar diferentes estrategias de marketing, como por ejemplo el realizar comparaciones de precios con los establecimientos anteriormente mencionados y a su vez poder atraer mayor clientela o mejorar la segmentación de mercado una vez se procese dicha información. Igualmente estos datos nos pueden ayudar para realizar la publicidad apropiada de aquellos productos que sean de temporada o rebajas.

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

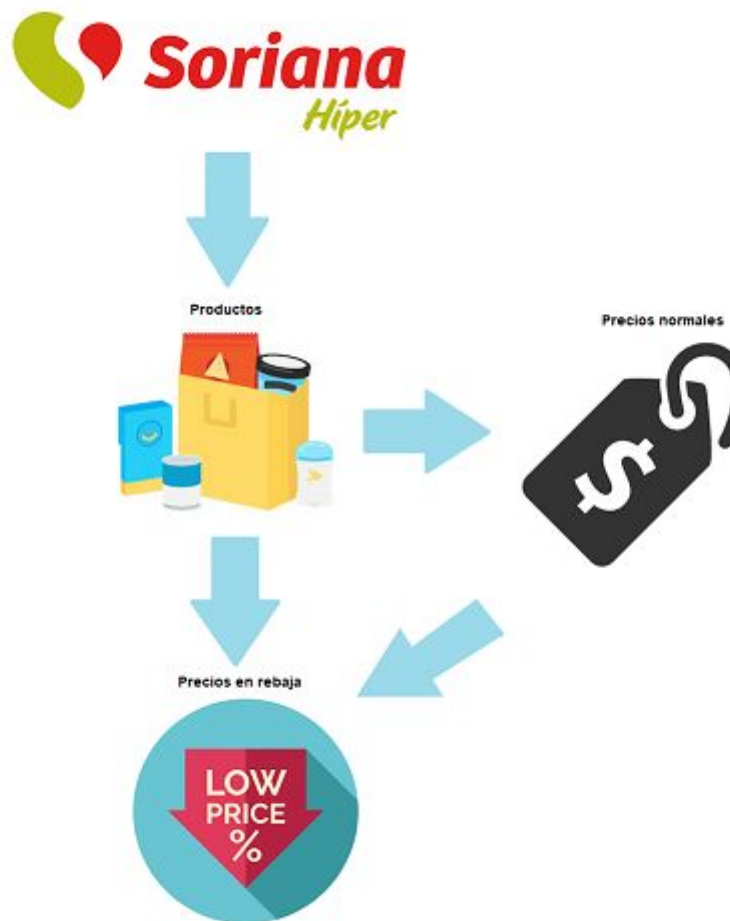
Seguimiento de precios del catálogo de productos que oferta la empresa Soriana.

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).

Tal como se describe en el título del dataset, éste contiene la información principal de cada uno de los productos que se exhiben en la página web de Soriana. Se encuentran tanto el nombre del producto como su precio regular y el de oferta o rebaja (en el caso de ser temporada, ya que no todos los productos se encuentran en rebajas o promoción simultáneamente). Esta información del dataset se descargará en un

archivo .CSV para su posterior limpieza, visualización y/o tratamiento adecuado, por ejemplo utilizando técnicas de minería de datos.

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.

En el actual dataset, se presentan datos de cada uno de los productos de la página web Soriana extraídos el día 25 de octubre del 2020. De donde se obtuvieron las siguientes características:

- A. **Nombre:** Nombre completo del producto.
- B. **Precio original:** Precio original de los productos.
- C. **Precio de rebaja:** Precio en rebaja o promoción de los productos. Esta variable puede tomar valores nulos.

Las tres variables anteriormente mencionadas: Nombre, precio original y precio de rebaja son de tipo *string* o cadena de caracteres, y no numéricas como se pensaría para variables relacionadas con los precios. Esto se debe a que cuando se extrae dicha información, a su vez extraemos el símbolo '\$' de pesos mexicanos. Entonces, en una etapa futura para realizar cualquier tipo de análisis con este dataset (por ejemplo de comparación de precios de mismos productos entre diferentes supermercados de cadena de México) se debe realizar un procesamiento de datos para eliminar dicho símbolo, también se requiere eliminar los espacios vacíos que pueden quedar dentro de algún valor y finalmente convertirlos a un tipo de dato *float* o flotante, esto por la misma característica de los precios.

Buenas prácticas:

Los datos fueron recolectados a través de web scraping usando el lenguaje de programación Python, para lo cual se hizo el respectivo uso de soporte del mapa del sitio (sitemap) de la página web Soriana, e igualmente se realizó la consulta del archivo robots.txt para respetar aquellas rutas que no se puedan contemplar dentro de esta actividad. Realizado esto, primeramente se almacenan cada una de las subcategorías con su respectivo nombre y URL en las cuales se pueden encontrar los productos, y posterior a ello se recorren individualmente cada una de las páginas de las subcategorías para ir almacenando los nombres de los productos, sus respectivos precios e imágenes para la descarga del material audiovisual.

En el caso de bloqueos, se ha utilizado una alternancia de user-agents y se ha agregado un delay de entre 5-8 segundos (a elección del operador) que se genera de manera aleatoria para darle la oportunidad de recuperarse al servidor. Más detalles y comentarios se encuentran incluidos en el archivo de jupyter notebook denominado: [code_PRA1.ipynb](#)

Es importante mencionar que los datos se han extraído en formato .csv como *raw data*, es decir, cualquier proceso de limpieza o procesamiento de datos deberá ser realizado posterior a la carga del archivo y de acuerdo a los requerimientos del proyecto donde se pretenda hacer uso de este conjunto de datos.

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

Se agradece al Director General Ricardo Martín Bringas, representante de D.R. © Tiendas Soriana, S.A. de C.V. Alejandro de Rodas 3102-A Col. Cumbres 8º Sector C.P. 64610 Monterrey, N.L. México

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

Este conjunto de datos es interesante porque nos permite tener un conglomerado de los precios, características y demás información del catálogo de productos que ofrece una tienda en línea en este caso de origen mexicano. Este dataset pretende recopilar toda la información que se muestra en el sitio web dentro de un solo archivo .csv, sobre el cual se puede trabajar, por ejemplo, haciendo inteligencia de negocio.

Concretamente, una vez recolectada la información de este sitio web se pueden contestar preguntas del tipo: ¿Cuántos de los productos se encuentran en temporada de rebajas?, ¿Cuáles son los productos más caros y más baratos de cada categoría determinada?, ¿También se encuentran en rebaja?, entre otras preguntas, a futuro una vez se realice el mismo procedimiento con otras páginas web de supermercados en México, podríamos abordar preguntas comparativas como: ¿Cuáles productos son más baratos en cada uno de los supermercados?, ¿Existe mayor variedad de productos en alguno de ellos?, entre otras.

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

De acuerdo a las cláusulas indicadas dentro de la web: <https://creativecommons.org/about/cclicenses/> se eligió la licencia **CC BY-NC-SA** que cubre los aspectos que deben estar regulados para un trabajo de esta magnitud. La(s) persona(s) interesadas en replicar o distribuir esta obra deberán hacerlo únicamente para fines no comerciales y en el caso de hacer alguna adaptación de la misma se deberá siempre dar el crédito respectivo a sus creadores.




- **CC BY-NC-SA:** This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator. If you remix, adapt, or build upon the material, you must license the modified material under identical terms.

CC BY-NC-SA includes the following elements:

BY  – Credit must be given to the creator

NC  – Only noncommercial uses of the work are permitted

SA  – Adaptations must be shared under the same terms

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

Project Link: https://github.com/dfdavila2/Webscraping_PRA1

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

10.5281/zenodo.4151260

Recursos

Los siguientes recursos son de utilidad para la realización de la PEC:

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Masip, D. El lenguaje Python. Editorial UOC.
 - Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
 - Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.
 - Tutorial de Github <https://guides.github.com/activities/hello-world>.

Criterios de valoración

Todos los apartados son obligatorios. La ponderación de los ejercicios es la siguiente:

- Los apartados 1, 2, 3 y 4 valen 0,25 puntos cada uno.
- Los apartados 5, 6, 7 y 8 valen 1 punto cada uno.
- Los apartados 9 y 10 valen 2,5 puntos cada uno.

Otros criterios que se tomarán en cuenta para la evaluación son:

- Idoneidad de las respuestas (deberán ser claras y completas).
- Complejidad del sitio web elegido para la extracción.
- Síntesis y claridad, a través del uso de comentarios, del código resultante.
- Presentación adecuada de los datos.
- Organización y claridad de los documentos de entrega final.
- Completitud de los documentos requeridos para la entrega final.

Quisiera destacar también que la nota final dependerá de la complejidad de la entrega realizada. Algunos aspectos que incrementan la dificultad son:

- Si se gestiona contenido audiovisual
- Si se utilizan tecnologías avanzadas como Selenium
- Si se gestiona código Javascript
- Si se utilizan métodos avanzados para saltarse la prevención de web scraping
- Si se gestionan usuarios y contraseñas

Formato y fecha de entrega

Durante la semana del 26 al 30 de octubre, el grupo podrá entregar al profesor una entrega parcial opcional. Esta entrega parcial es muy recomendable para recibir asesoramiento sobre la práctica y verificar que la dirección tomada es la correcta. Se entregarán comentarios a los estudiantes que hayan efectuado la entrega parcial pero no contará para la nota de la práctica. En la entrega parcial los estudiantes deberán entregar por correo electrónico, al profesor encargado del aula, el enlace al repositorio Github con lo que hayan avanzado. mcalvogonza@uoc.edu

En referencia a la entrega final, hay que entregar un único fichero que contenga el enlace a Github donde haya:

1. Una Wiki donde estén los nombres de los componentes del grupo y una descripción de los ficheros.
2. Un documento PDF con las respuestas a las preguntas y los nombres de los componentes del grupo. Además, al final del documento, debe aparecer la siguiente tabla de contribuciones al trabajo, la cual debe firmar cada integrante del grupo con sus iniciales. Las iniciales representan la confirmación por parte del grupo que el integrante ha participado en dicho apartado. Todos los integrantes deben participar en cada apartado, por lo que, idealmente, los apartados deberían estar firmados por todos los integrantes.

Contribuciones	Firma
Investigación previa	DFDO (dfdavila) Mónica Alexandra Gómez Martínez
Redacción de las respuestas	DFDO (dfdavila) Mónica Alexandra Gómez Martínez
Desarrollo código	DFDO (dfdavila) Mónica Alexandra Gómez Martínez

3. Una carpeta con el código Python o R generado para obtener los datos.
4. El DOI a los datos.

Este documento de la entrega final se tiene que entregar en el espacio de Entrega y Registro de AC del aula antes de las **23:59 del día 9 de noviembre**. No se aceptarán entregas fuera de plazo.