Práctica 1 (25% nota final)

Autores:

- Adrian Felipe Pinzon Hurtado
- Diego Armando Quintero Quiñones

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar herramientas de extracción de datos. Para hacer esta práctica tendréis que trabajar en grupos de 2 personas. Tendréis que entregar en el REC un solo fichero con el enlace al repositorio Git donde haya las soluciones, incluyendo los nombres de los componentes del grupo. Podéis utilizar la Wiki o README.md del repositorio para describir vuestro grupo y los diferentes archivos de vuestra entrega. Cada miembro del grupo tendrá que contribuir con su usuario del repositorio. 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 universitario en Ciencia de Datos:

- 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 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.

Descripción de la práctica a realizar

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en un sitio web. El idioma del sitio web elegido deberá ser español, inglés o catalán. Se deberán resolver los siguientes apartados:

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

La página seleccionada para recolectar información fue https://www.milanuncios.com/. Esta página brinda a los usuarios la facilidad de comprar y vender distintos tipos de productos de forma segura y rápida desde la comodidad de la casa, la gran ventaja es que cubre gran variedad de productos de los cuales se pide información de acuerdo al tipo, por ejemplo: si se desea vender coches se solicita kilometraje, modelo, marca, número de puertas, tipo de combustibles y demás.

Es un sitio importante ya que permite medir dependiendo del enfoque del análisis cómo ha afectado la pandemia la compra o venta de productos por una categoría específica si se desea. Adicionalmente, puede ser una base para los demás comercios el saber cómo se está comportando el mercado en otros lugares y tomar medidas sobre la arquitectura tecnológica que permita soportar picos de uso.

Es por esta razón, que el sitio milanuncio fue seleccionado, con el fin de empezar a conocer sobre los atributos que este portal tiene y poder a futuro extrapolar la implementación a otros sitios, para esta ocasión se tomó la información de coches.

2. **Título.** Definir un título que sea descriptivo para el dataset.

Detalle de publicaciones de coches en milanuncios

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.

Este dataset es generado de la página de publicaciones milanuncios, en esta ocasión el enfoque fue orientado a retornar la información del anuncio pero sin el procesamiento total de información, por ejemplo:

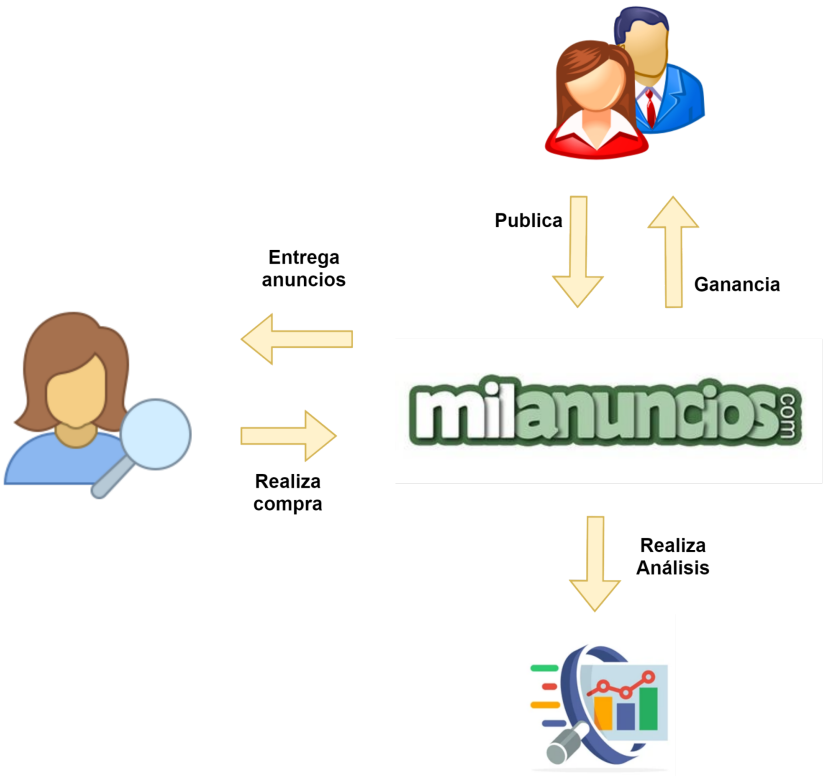
- En los campos se pueden encontrar campos con NaN
- La descripción se encuentra tal como en la publicación, con mayúsculas o minúsculas mezcladas o todo en mayúscula.
- El campo precio contiene el símbolo de la moneda, en este caso “€”
- El kilometraje contiene el número de kilómetros y la abreviación “kms”.

El enfoque en esta extracción es depositar los datos tal como se encuentran en la

publicación y ya en un trabajo previo de acuerdo a la finalidad será necesario realizar una fase posterior de limpieza.

4. **Representación gráfica.** Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

A continuación se realiza una representación gráfica del dominio utilizado en la practica, como ya se había dicho anteriormente, se utiliza la página milanuncia empleada para la compra, venta y alquiler de productos de distintos tipos.



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

Los datos contenidos en el dataset son los siguientes:

- **Name**: Hace referencia al nombre de la publicación.
- **Descripción**: Descripción del anunció.
- **Categoría**: Categoría en la que se publicó el anuncio.
- **Subcategoría**: Subcategoría en la que se publicó el anuncio.
- **Link**: Link directo del anuncio.
- **Ubicación**: ubicación donde se encuentra el producto publicado

- **Oferta_Demanda**: Indica si es Oferta o Demanda.
- **Price**: Precio del coche
- **Kms/Type_vehicle**: Kilometraje del coche
- **Other_skills**: Otras características.

Se obtuvieron mediante web scraping con python desde Jupyter Notebook, fueron almacenados en un objeto directorio y posteriormente convertidos en dataframe con la librería pandas. Para finalizar el resultado es almacenado en formato .CSV.

Los datos aquí registrados se encuentran tal como están en la publicación y para un análisis posterior será necesario realizar la limpieza correspondiente, en el apartado 3 se describe cómo se encuentra el dataset.

6. **Agradecimientos.** Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

La propiedad intelectual de las publicaciones es del sitio Milanuncios debido a que el usuario al subir las fotos y publicaciones cede gratuitamente a Milanuncios los derechos de explotación de propiedad intelectual sobre las mismas. cito textualmente:

"Al subir fotografías al Portal, el usuario cede gratuitamente a Milanuncios los derechos de explotación de propiedad intelectual sobre las mismas, por lo que Milanuncios podrá reproducirlas, transformarlas (incluyendo, sin limitación, la inclusión de marcas de agua u otros mecanismos que impidan el aprovechamiento inconsentido por parte de terceros), distribuirlas y comunicarlas al público (incluida la puesta a disposición del público) a través de cualquier modalidad de explotación y utilizando cualquier formato o soporte o medio de explotación o comunicación. Dicha cesión de derechos no está sujeta a ninguna limitación de carácter temporal ni territorial, esto es, se realiza para todo el mundo y por todo el tiempo de vigencia legal de los mismos. Milanuncios podrá ejercer los derechos de explotación de las fotografías en la forma que estime más conveniente, y podrá incluso transmitirlos o cederlos con carácter exclusivo o no a terceros en los términos y condiciones que considere oportunos."

Esta información se puede encontrar en el apartado **4. Propiedad intelectual e industrial** de la página de milanuncios.

Basado en lo anterior, se envió un correo a la dependencia que maneja la privacidad de los datos al correo privacidad@milanuncios.com en busca de tener una autorización para el uso de los datos de manera experimental y académica.

Con lo anterior damos respeto al uso de los datos y es por ellos que damos agradecimientos a la página milanuncios a las personas encargadas de brindar una experiencia magnífica para el uso del portal, agradecemos al equipo que está detrás del portal dando soporte y permitiendo

que el sitio se mantenga en optimas condiciones, tambien agradecemos a los propietarios que publican sus anuncios y mantiene la comunidad.

Para entender un poco mejor la metodología de trabajo y la manera de realizarlo pudimos encontrar el siguiente trabajo:

1. Web Scraping, Programming - 🐍 Python - [[MilAnuncios.com]]
   https://www.youtube.com/watch?v=EuuVow9Lx6o
   Este es un video donde se realiza un desarrollo orientado a la página milanuncios que que permite tener una guía de implementación.

2. Comparador de componentes de hardware [1]. Este trabajo realiza una comparación de precios en los comercios en línea con el fin de mejorar la cuota de mercado de los pequeños y medianos comercios permitiéndole a los consumidores comparar productos entre las diferentes tiendas.

Existen otros portales de anuncios como:

1. https://www.amazon.com/
2. https://www.mercadolibre.com.co/

7. **Inspiración.** Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

La pandemia ha llegado consigo un crecimiento exponencial del uso distintas tecnologías, uso del internet, aplicativos digitales y demás, eso ha llevado consigo la aparición de muchos startups pero también a que los distintos sitios deban robustecer su arquitectura, milanuncios al ser un sitio de publicaciones, compra y venta de distintos elementos se convierte en un sitio interesante de análisis, ver su crecimiento, cantidad de publicaciones por categoria, picos y demás.

De momento es un análisis experimental y está bien conocer la estructura y cómo acceder a los datos, más adelante y quizás de uso profesional pueda estar orientado a buscar categorías específicas como coches filtrado por una marca y año y determinar el comportamiento del precios y dichos vehículos, esto además ayuda a que esta implementación pueda ser orientada a otras página similares de regiones especificar en otros países.
.

8. **Licencia.** Seleccionar una de estas licencias para el dataset resultante y justificar 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.

En este apartado hemos seleccionado la licencia **Released Under CC BY-SA 4.0 License** basado en:

- Este trabajo se ha realizado de manera exploratoria y no afectará directamente la ejecución de scraper a la página.
- Se han presentado los autores y propietarios de la información
- Se puede enriquecer la investigación realizada y posteriormente aplicarla a otros ámbitos.
- Permite realizar trabajos sobre los resultados de datos sin solicitar permiso.

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

Este código fue realizado en Júpiter Notebook con python y la librería Selenium, se instalaron algunos paquetes como:

- pip install selenium
- pip install chromium-chromedriver

El código se encuentra en el repositorio GitHub en la siguiente ruta:

10. **Dataset.** Publicar el dataset obtenido*(*)* en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

    (*) Si existe algún impedimento para publicar el dataset real, se deberá justificar esta situación y realizar y publicar en Zenodo un dataset simulado. En este caso, el dataset real se comunicará al profesor de forma privada (p.ej., enlace de Google Drive).

Referencias

[1] Pedreño Marine, J. (2019). *Buildy: comparador de components de hardware*. http://openaccess.uoc.edu/webapps/o2/bitstream/10609/95367/9/jpedrenomTFG0619mem%c3%b2ria.pdf

Recursos

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

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Masip, D. (2019) 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:

Apartado	1	2	3	4	5	6	7	8	9	10
Puntos	0,25	0,25	0,25	0,5	1	1,5	1,25	1	2	2

Criterios que se tomarán en cuenta para la valoración de la práctica:

- 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.
- Seguimiento de recomendaciones para el buen uso del web scraping.

Formato y fecha de entrega

Durante la semana **del 25 al 29 de octubre**, el grupo podrá realizar 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 enviar por correo electrónico, al profesor colaborador del aula, el enlace al repositorio Git con lo que hayan avanzado.

En referencia a la entrega final, se pide:

a. **Un único documento** (.txt, .pdf, .docx) que contenga **el enlace al repositorio Git** del proyecto (apartado b) y **el enlace al video del proyecto** (apartado c).

Este documento se entregará en el espacio de Entrega y Registro de EC del aula.

b. Un **repositorio Git** con las soluciones de la práctica. El repositorio Git se creará en Github (https://github.com/), y podrá ser un repositorio público o privado, a elección del grupo. Si se utiliza un repositorio privado, se deberá facilitar acceso al profesor, mediante el nombre de usuario que indicará en el Tablón del aula o por email. **El repositorio no se podrá modificar pasada la fecha de entrega**, y deberá contener:
   b.1. Una **Wiki** o **README.md** donde estén los nombres de los componentes del grupo, una descripción de los ficheros y el DOI de Zenodo del dataset generado.
   b.2. Un **documento PDF** con las respuestas a los apartados 1-10 y los nombres de los componentes del grupo. **La extensión de este documento no debe superar las 20 páginas.** 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 de 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	Integrante 1, Integrante 2
Redacción de las respuestas	Integrante 1, Integrante 2
Desarrollo del código	Integrante 1, Integrante 2

   b.3. Una carpeta con el **código Python** o **R** generado para obtener los datos.

c. Un **breve vídeo** con la participación de los dos componentes del grupo, donde se realizará una presentación del proyecto, destacando los puntos más relevantes. El video se deberá compartir mediante un enlace del Google Drive de la UOC o incluirse en el repositorio Git. **La duración de este vídeo no debe superar los 10 minutos.**

El documento de la entrega final se tiene que subir al espacio de Entrega y Registro de EC del aula antes de las **23:59h CET del día 8 de noviembre**. No se aceptarán entregas fuera de plazo.

Si se estima oportuno, el profesor solicitará a los integrantes del grupo una entrevista remota (de forma conjunta o individual) mediante Google Meet, en referencia a la práctica realizada, en un día y hora acordados.