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 aportan 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

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:

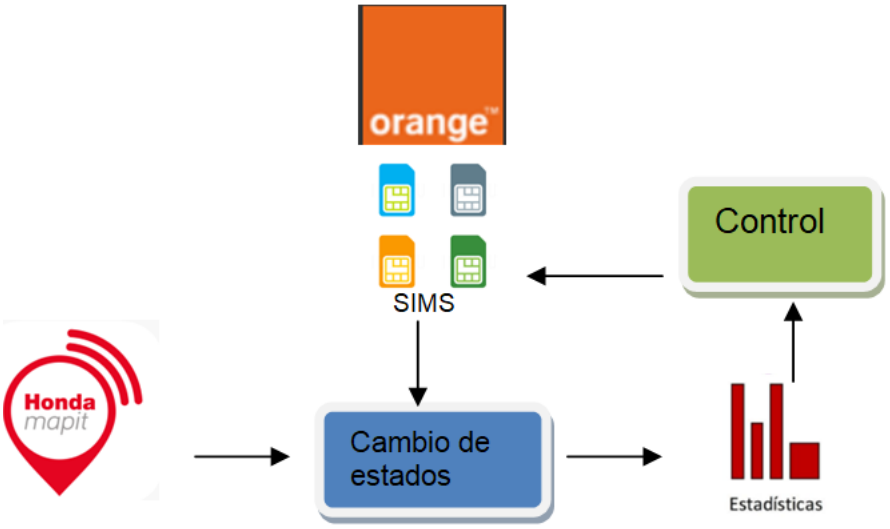
1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.
   Para ponerse en contexto hay que explicar que trabajo para una compañía que realiza dispositivos gps para motos, de esta forma poder trackear sus rutas y el posible robo de la moto. Cada dispositivo cuenta con una tarjeta sim de Orange lo que le permite transmitir los datos del dispositivo a nuestros servidores. Nos era muy importante incorporar a nuestra base de datos la información actualizada de cada una de estas sims para así poder actuar en consecuencia.

   Por lo que en este proyecto me he centrado en desarrollar la extracción de la información necesaria de las sims vía web scraping de la web de Orange (que proporciona información de la flota de líneas contratadas con su servicio).

2. Definir un título para el dataset. Elegir un título que sea descriptivo.
   Estado e información actual de toda la flota de líneas Orange.

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 expresa el título, el dataset está basado en la información actual sobre el estado de todas las sims de nuestra flota. En este dataset, se presentan dichos datos para cada una de las sims. Las fechas de las características extraídas son en formato MM/DD/YYYY. Los datos no han pasado por un proceso de preprocesado o limpieza, por lo que aún pueden existir inconsistencias. En este caso, se extrajo la información para las primeras 300 sims. La descripción de las características extraídas son descritas en las siguientes preguntas. El formato del dataset es un fichero CSV que facilita su visualización y tratamiento.

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 este dataset, se presentan solo los datos actuales para las primeras 100 sims de la lista completa. Las características extraídas son las siguientes:

   A. Line: Identificador único de cada sim.
   B. Status: Estado actual de la sim (dependiendo del estado de la sim esta tendrá comunicación o no)
   C. Date: Fecha en la que se cambió el estado.

   Los datos fueron recogidos a través de web scraping en lenguaje Python sobre la página de la lista de lines dentro del sitio web. Para ello, primero se realizó el login automático del portal web. Luego, de forma automática, se recorre cada una de las filas donde se encuentra la información relevante. Finalmente, se guardan los datos extraídos en un fichero CSV.

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

   El propietario del sitio web y del conjunto de datos es Orange. Orange proporciona está información a través de una API y también a través de ficheros csv descargables. EL método que usamos en nuestra empresa para obtener estos datos es a través de llamadas a la API pero para aprender web scraping he utilizado este método para obtener una parte de los datos y valorar así la idoneidad de este método.

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

El interés detrás de este conjunto de datos es la intención de tener un mayor control de las sims y a la vez obtener diversos datos relevantes a la hora de realizar análisis de comportamiento. Aunque no se ha mostrado también hay datos sobre tráfico de cada sim por lo que se pueden hacer estudios sobre consumo y demás.

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

    Unknown License ya que no se le ha informado a Orange ya que en sí las sims son de nuestra propiedad.

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.
    El código fue hecho en lenguaje Python con la implementación de la librería Selenium. El código fuente se encuentra dentro de la carpeta Git.
10. Dataset. Presentar el dataset en formato CSV
    El dataset en formato CSV se encuentra dentro de la carpeta Git.

Recursos

Los siguientes recursos son de utilidad por 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, 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 del documentos de entrega final.
- Completitud de los documentos requeridos para la entrega final.

Formato y fecha de entrega

Durante la semana del 28 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 (lsubirats@uoc.edu) el enlace al repositorio Github con lo que hayan avanzado.

En referente 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	Integrante 1, Integrante 2, ...
Redacción de las respuestas	Integrante 1, Integrante 2, ...
Desarrollo código	Integrante 1, Integrante 2, ...

3. Una carpeta con el código Python o R generado para obtener los datos.
4. El fichero CSV con 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 11 de noviembre**. No se aceptarán entregas fuera de plazo.