David Pulido Camacho

Práctica 1 (25% 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.

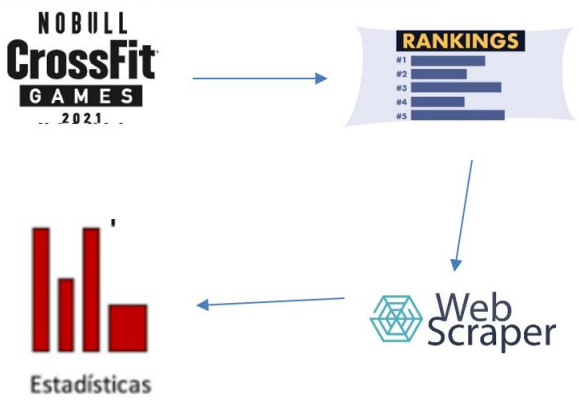
Descripción de la Práctica a realizar

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 uno de mis mayores hobbys es la práctica de Crossfit. Hace un mes se realizó una competición global (Open) en la que pueden participar atletas de todo el mundo vía online y se crea un leaderboard con el ranking. Por curiosidad me gustaría realizar un análisis detallado de esos datos.
   Por lo que en este proyecto me he centrado en desarrollar la extracción de la información necesaria de los atletas vía web scraping de la web de Crossfit (que proporciona información de los atletas inscritos en la competición).

2. Definir un título para el dataset. Elegir un título que sea descriptivo.
   Stats Athletes Crossfit Opengames 2021.

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 todos los atletas inscritos en los Crossfit Opengames de 2021. En este dataset, se presentan datos para cada uno de los atletas. 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 los primeros 1000 atletas. 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 esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.



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 los primeros 1000 atletas de la lista completa. Las características extraídas son las siguientes:
   A. **Country**: Nacionalidad del atleta.
   B. **Fullname**: Nombre y apellidos del atleta.
   C. **Points**: Puntos conseguidos en la competición.
   D. **Rank**: Posición obtenida en la competición.
   C. **Region**: Continente al que pertenece el atleta
   Los datos fueron recogidos a través de web scraping en lenguaje Python con Selenium sobre la página del leaderboard del sitio web. Para ello, 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 análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.
   El propietario del sitio web y del conjunto de datos es Crossfit. Encontré un análisis del ranking del 2018 realizado por Jean-Michel D https://towardsdatascience.com/analysis-of-the-crossfit-open-2018-jean-michel-d-307cbfb06a13 que me ha dado la idea de realizar un análisis más exhaustivo del de este año.

7. Inspiración. Explique 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.
   El interés detrás de este conjunto de datos es la intención de obtener

información relevante sobre la competición. Hacer un perfil de los participantes y analizar las características que pueden llevar a tener un mayor éxito.

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 Crossfit de la creación del dataset.

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. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.
    El dataset en formato CSV se encuentra dentro de la carpeta Git.

Contribuciones	Firma
Investigación previa	DPC
Redacción de las respuestas	DPC
Desarrollo código	DPC

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 y 8 valen 1 punto cada uno.
- Los apartados 6 y 7 valen 1,5 puntos cada uno.
- Los apartados 9 y 10 valen 2 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.
- Seguimiento de recomendaciones para el buen uso del web scraping.

Formato y fecha de entrega

Durante la semana del 29 de marzo al 02 de abril, 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.

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 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 12 de abril**. No se aceptarán entregas fuera de plazo.