Tipología y ciclo de vida de los datos

Práctica 1

Universitat Oberta de Catalunya

¿Cómo podemos capturar los datos de la web?
25% nota final

Fecha de entrega
12 de noviembre de 2024

Presentación

En esta práctica se elaborará un caso práctico orientado a identificar y extraer datos relevantes para un proyecto analítico, empleando herramientas específicas de *web scraping*. Para realizar esta práctica se requiere trabajar en **grupos de dos personas**.

De forma orientativa, se pueden consultar los siguientes ejemplos, teniendo en cuenta que las respuestas pueden no ser las más adecuadas para la práctica que se plantea este semestre:

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

La entrega de esta práctica se ha de realizar según lo especificado en el apartado Formato y fecha de entrega. Se debe entregar una memoria PDF con las respuestas a los diferentes apartados, un repositorio con el código fuente y un vídeo explicativo, en el que ambos integrantes del grupo comenten los aspectos más relevantes del proyecto.

Es importante tener en cuenta las siguientes consideraciones a la hora de entregar la práctica:

- Es obligatorio y **queda como responsabilidad de cada estudiante revisar que el archivo entregado es el correcto**. Un archivo vacío o no pertinente se considerará como no entregado.
- Para que la práctica se considere como entregada, se debe completar al menos el 25% de toda la actividad.
- No podrá modificarse ningún elemento de la práctica pasada la fecha de entrega (repositorio, archivos de Google Drive, etc.).
- Asimismo, también es responsabilidad del estudiante asegurarse de que, en el momento de entregar la práctica, **se haya dado acceso al profesor a los diferentes elementos privados que se entreguen** (p. ej., repositorio GitHub privado o archivos restringidos de Google Drive). El profesor indicará en los Anuncios del aula su nombre de usuario en estas plataformas.
- No se puede hacer grupos con alumnos de aulas diferentes.

Competencias

En esta práctica se desarrollan las siguientes competencias del máster universitario de 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 la 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. Deben tenerse en cuenta las consideraciones sobre el sitio web elegido, el código y el dataset que se indican más adelante. Se deberá presentar una memoria en PDF **(máximo 20 páginas)** en la que se resuelvan los siguientes apartados:

1. **Contexto.** Explicar en qué contexto específico se han recolectado los datos y argumentar por qué el sitio web seleccionado es una fuente pertinente y fiable de esa información. Indicar la dirección del sitio web.
2. **Título.** Definir un título conciso y que sea descriptivo para el dataset.
3. **Descripción del dataset.** Desarrollar una breve descripción del conjunto de datos que se ha extraído. Es necesario que esta descripción sea coherente con el título elegido.
4. **Representación gráfica.** Dibujar un esquema o diagrama que refleje visualmente el dataset y el proyecto elegido.
5. **Contenido.** Explicar los campos que se incluyen en el dataset y el período de tiempo al que pertenecen los datos.
6. **Propietario.** Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en su defecto, justificar esta búsqueda con análisis similares. Indicar qué pasos se han seguido para actuar de acuerdo con los principios éticos y legales en el contexto del proyecto elegido.
7. **Inspiración.** Explicar por qué puede ser interesante este conjunto de datos y qué preguntas se pretenden responder con ellos. Es necesario comparar con los análisis anteriores o análisis similares presentados en el apartado 6.
8. **Licencia.** Seleccionar una licencia adecuada para el dataset resultante y justificar el motivo de su elección. Ejemplos de licencias que pueden considerarse:
    - 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.
- Otra (especificar cuál).

9. **Código.** Código implementado para la obtención del dataset, preferiblemente en Python o, alternativamente, en R.
    - El código deberá ubicarse en la carpeta **/source** del repositorio.
    - Se deben indicar las librerías y versiones utilizadas. P. ej., en Python pueden obtenerse mediante el comando
      `pip3 freeze > requirements.txt`
    - En la memoria en PDF, se deben comentar los aspectos más relevantes sobre cómo el código realiza el proceso de recolección de datos, qué dificultades presenta el sitio web elegido, y cómo se han resuelto.
10. **Dataset.** Publicar el dataset obtenido en formato CSV en Zenodo, incluyendo una breve descripción del mismo. Obtener y adjuntar el enlace del DOI del dataset (https://doi.org/…). El dataset también deberá incluirse en la carpeta **/dataset** del repositorio. Si existe alguna circunstancia que impida publicar abiertamente el dataset real en Zenodo, se deberá:
    a. Comentar esta circunstancia y justificar el motivo.
    b. Generar un dataset simulado y publicarlo en Zenodo, obteniendo el enlace del DOI.
    c. Comunicar al profesor el dataset real de forma privada (p. ej., en el repositorio privado o en una carpeta de Google Drive privada).
11. **Vídeo.** Realizar un breve vídeo explicativo de la práctica **(máximo 10 minutos)**, que deberá contar con la participación de los dos integrantes del grupo. En el vídeo se deberá realizar una presentación del proyecto, destacando los puntos más relevantes, tanto de las respuestas a los apartados como del código utilizado para extraer los datos. Indicar el enlace del vídeo (https://drive.google.com/…), que deberá ubicarse en el Google Drive de la UOC.

Consideraciones sobre el sitio web elegido, el código y el dataset

A la hora de elegir un sitio web para realizar esta práctica, es importante tener en cuenta que el objetivo primordial consiste en extraer un dataset potencialmente “interesante” para un hipotético análisis posterior, y que el proceso de extracción de los datos no sea completamente “trivial”:

- El idioma del sitio web escogido debe ser **español**, **inglés** o **catalán**.
- El sitio web elegido no puede ser un sitio de "prácticas" (p.ej. https://books.toscrape.com/), sino **un sitio real**.
- El código generado para obtener el dataset debe incluir **descubrimiento de enlaces y navegación autónoma**. P. ej., no basta con procesar el contenido de una única página web donde aparezca todo el dataset en una única tabla.

- El código debe implementar mecanismos que permitan ejecutar un **uso responsable del *web scraping*** (p. ej., evitar saturar al servidor).
- Debe comprobarse **qué User-Agent está utilizando el código**, aunque se utilice un WebDriver.
- **No se permite el uso de APIs como parte principal de la práctica.** En el caso de que el sitio web elegido ofrezca alguna API para acceder a los datos, se deberá prescindir de la misma. Se permite el uso de APIs sólo como parte complementaria a la práctica, p. ej., para realizar consultas a algún servicio adicional para tratar o completar los datos recogidos.
- El código debe tener un **nivel adecuado de modularidad y estar debidamente comentado**. No se trata de insertar un comentario por cada línea de código, sino de comentar puntos claves que ayuden a entender qué se está realizando.
- **No es necesario realizar la limpieza del dataset resultante**, ya que este proceso será uno de los objetivos de la Práctica 2. Si se desea utilizar el dataset generado para la Práctica 2 (no es obligatorio utilizar este dataset), sería conveniente que incluyera tanto datos numéricos como categóricos.

La nota final tendrá en cuenta las dificultades abordadas en la recolección del dataset. Algunos aspectos que incrementan la dificultad son:

- Uso de tecnologías avanzadas como *Selenium* o *Scrapy*.
- Recolección de datos de sitios web con contenido dinámico (p. ej., *infinite scroll*, *mouseover*).
- Uso de métodos avanzados para saltarse la prevención de *web scraping*.
- Gestión de contenido audiovisual.
- Gestión de usuarios y contraseñas.
- Gestión de código *JavaScript*.

Recursos

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

- 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 apartados es la siguiente:

Apartado	1	2	3	4	5	6	7	8	9	10	11
Puntos	0,25	0,25	0,25	0,5	1	1,5	1,25	0,5	2	2	0,5

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 de datos. Es importante tener en cuenta que la complejidad será un factor que se evaluará y dependerá tanto del sitio elegido como del análisis realizado en la práctica.
- 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

En referencia a la entrega de la práctica, se pide:

a. **La memoria de la práctica**, que deberá ser **un único documento PDF**, cuya extensión **no debe superar las 20 páginas**. En la primera página deberá contener:
   - Los nombres de los integrantes del grupo.
   - El enlace al sitio web elegido.
   - El enlace al repositorio con el código de la práctica.
   - El enlace al dataset publicado en Zenodo.
   - El enlace al vídeo de presentación de la práctica.

   A continuación, la memoria debe contener las **respuestas a los 11 apartados**.

   Además, al final del documento, debe aparecer la siguiente tabla de contribuciones al trabajo realizado, 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
Participación en el vídeo	Integrante 1, Integrante 2

Este documento se entregará, por cada uno de los integrantes del grupo, en el **Espacio de Entrega PR1 del aula Canvas**.

b. **Un repositorio Git** con el código de de la práctica en la rama "**main**". El repositorio se creará en GitHub (https://github.com/). Deberá ser un **repositorio privado**, por lo que se deberá dar acceso al profesor en el momento de la entrega. El repositorio deberá contener:
   1. **Un documento README.md**: Estará ubicado en la carpeta raíz y deberá contener:
      - Los nombres de los integrantes del grupo
      - Un apartado donde se describan los archivos que componen el repositorio.
      - Un apartado donde se describa cómo usar el código del repositorio. Deberá incluir información sobre los posibles parámetros que admita el script y uno o varios ejemplos replicables de su uso.
      - El DOI de Zenodo del dataset generado.
   2. Un archivo `requirements.txt` con las librerías necesarias para ejecutar el código.
   3. **Carpeta /source**: Deberá contener el código Python o R implementado para la obtención de los datos.
   4. **Carpeta /dataset**: Deberá contener el dataset resultante en formato CSV.

c. **Un vídeo explicativo**, cuya duración **no debe superar los 10 minutos**. El enlace del mismo se debe indicar en el apartado 11 del documento PDF.

El documento PDF (memoria) se tiene que subir al Espacio de Entrega PR1 del aula Canvas antes de las **23:59h CET del día 12 de noviembre de 2024**. No se aceptarán entregas fuera de plazo. **No podrá modificarse ningún elemento de la práctica pasada la fecha de entrega** (repositorio, archivos de Google Drive, etc.).

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

Propiedad intelectual

Al presentar una práctica o PEC que haga uso de recursos ajenos, se tiene que presentar junto con ella un documento en que se detallen todos ellos, especificando el nombre de cada recurso, su autor, el lugar donde se obtuvo y su estatus legal: si la obra está protegida por el copyright o se acoge a alguna otra licencia de uso (Creative Commons, licencia

GNU, GPL etc.). El estudiantado tendrá que asegurar que la licencia no impide específicamente su uso en el marco de la práctica o PEC. En caso de no encontrar la información correspondiente tendrá que asumir que la obra está protegida por el copyright.