Práctica 1 (25% nota final)

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.

   Debido al constante auge en el que se encuentra la venta de productos online y, la continua aparición de nuevas empresas de comercio electrónico, la página web donde realicé el webscraping es Amazon.

   Principalmente, el objetivo no es solo encontrar los ordenadores más baratos en esta página (o cualquier otro producto), si no los más baratos en cualquier página de empresas de distribución online.

   Es por ello que, a pesar de que se ha creado en base a Amazon, podría aplicarse a otras páginas web similares y realizar los análisis que deseemos.

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

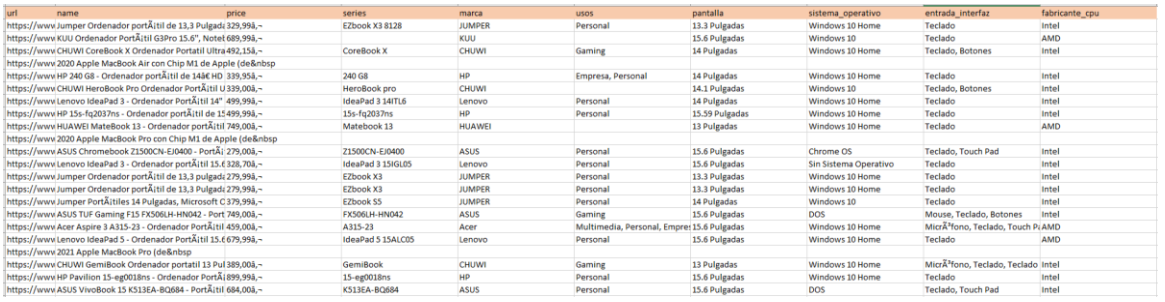
   Al estar trabajando sobre un dataset referente a ordenadores portátiles de Amazon y su descripción, llamaremos al mismo Amazon Laptops Dataset.

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.

   El dataset está compuesto por 989 registros de portátiles que se encuentran a la venta (o incluso que están temporalmente sin stock) en la página web de Amazon. Cada registro está formado por descripciones tales como el nombre, el precio, la marca, el uso (empresa o gaming, entre otros) o el tamaño de la pantalla.

   En el punto 5 se explica detalladamente cada una de las definiciones.

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



url	name	price	series	marca	usos	pantalla	sistema_operativo	entrada_interfaz	fabricante_cpu
https://www	Jumper Ordenador portÃ¡til de 13,3 Pulgad	329,99â‚¬	EZbook X3 8128	JUMPER	Personal	13.3 Pulgadas	Windows 10 Home	Teclado	Intel
https://www	KUU Ordenador portÃ¡til G3Pro 15.6", Notet	689,99â‚¬		KUU		15.6 Pulgadas	Windows 10	Teclado	AMD
https://www	CHUWI CoreBook X Ordenador Portatil Ultra	492,15â‚¬	CoreBook X	CHUWI	Gaming	14 Pulgadas	Windows 10 Home	Teclado, Botones	Intel
https://www	2020 Apple MacBook Air con Chip M1 de Apple (de								
https://www	HP 240 G8 - Ordenador portÃ¡til de 14â€ HD	339,95â‚¬	240 G8	HP	Empresa, Personal	14 Pulgadas	Windows 10 Home	Teclado	Intel
https://www	CHUWI HeroBook Pro Ordenador PortÃ¡til U	339,00â‚¬	HeroBook pro	CHUWI		14.1 Pulgadas	Windows 10	Teclado, Botones	Intel
https://www	Lenovo IdeaPad 3 - Ordenador PortÃ¡til 14"	499,99â‚¬	IdeaPad 3 14ITL6	Lenovo	Personal	14 Pulgadas	Windows 10 Home	Teclado	Intel
https://www	HP 15s-fq2037ns - Ordenador portÃ¡til de 15	499,99â‚¬	15s-fq2037ns	HP	Personal	15.59 Pulgadas	Windows 10 Home	Teclado	Intel
https://www	HUAWEI MateBook 13 - Ordenador portÃ¡til	749,00â‚¬	Matebook 13	HUAWEI		13 Pulgadas	Windows 10 Home	Teclado	AMD
https://www	2020 Apple MacBook Pro con Chip M1 de Apple (de								
https://www	ASUS Chromebook Z1500CN-EJ0400 - PortÃ¡	279,00â‚¬	Z1500CN-EJ0400	ASUS	Personal	15.6 Pulgadas	Chrome OS	Teclado, Touch Pad	Intel
https://www	Lenovo IdeaPad 3 - Ordenador portÃ¡til 15.€	328,70â‚¬	IdeaPad 3 15IGL05	Lenovo	Personal	15.6 Pulgadas	Sin Sistema Operativo	Teclado	Intel
https://www	Jumper Ordenador portÃ¡til de 13,3 pulgad	279,99â‚¬	EZbook X3	JUMPER	Personal	13.3 Pulgadas	Windows 10 Home	Teclado	Intel
https://www	Jumper Ordenador portÃ¡til de 13,3 Pulgad	279,99â‚¬	EZbook X3	JUMPER	Personal	13.3 Pulgadas	Windows 10 Home	Teclado	Intel
https://www	Jumper PortÃ¡tiles 14 Pulgadas, Microsoft C	379,99â‚¬	EZbook S5	JUMPER	Personal	14 Pulgadas	Windows 10	Teclado	Intel
https://www	ASUS TUF Gaming F15 FX506LH-HN042 - Port	749,00â‚¬	FX506LH-HN042	ASUS	Gaming	15.6 Pulgadas	DOS	Mouse, Teclado, Botones	Intel
https://www	Acer Aspire 3 A315-23 - Ordenador PortÃ¡til	459,00â‚¬	A315-23	Acer	Multimedia, Personal, Empre	15.6 Pulgadas	Windows 10 Home	MicrÃ³fono, Teclado, Touch P	AMD
https://www	Lenovo IdeaPad 5 - Ordenador PortÃ¡til 15.€	679,99â‚¬	IdeaPad 5 15ALC05	Lenovo	Personal	15.6 Pulgadas	Windows 10 Home	Teclado	AMD
https://www	2021 Apple MacBook Pro (de								
https://www	CHUWI GemiBook Ordenador portatil 13 Pul	389,00â‚¬	GemiBook	CHUWI	Gaming	13 Pulgadas	Windows 10 Home	MicrÃ³fono, Teclado, Teclado	Intel
https://www	HP Pavilion 15-eg0018ns - Ordenador PortÃ¡	899,99â‚¬	15-eg0018ns	HP	Personal	15.6 Pulgadas	Windows 10 Home	Teclado	Intel
https://www	ASUS VivoBook 15 K513EA-BQ684 - PortÃ¡til	684,00â‚¬	K513EA-BQ684	ASUS	Personal	15.6 Pulgadas	DOS	Teclado, Touch Pad	Intel

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

    - **url**: La URL de cada producto, en mi caso, de cada ordenador.
    - **name**: El nombre proporcionado por Amazon de cada ordenador.
    - **Price**: el precio del ordenador.
    - **Series**: la Serie del ordenador.
    - **Marca**: la Marca del ordenador.
    - **Usos**: Los usos (etiquetas) a las que se asocia el ordenador. P.e., Gaming, Personal o Empresa. Esto es especialmente útil cuando no se tiene grandes nociones técnicas sobre ordenadores y se desea un resumen breve de su uso.
    - **Pantalla**: Tamaño de la pantalla en pulgadas.
    - **Sistema Operativo**: Sistema Operativo que lleva el ordenador.
    - **entrada_interfaz**: tipos de entradas del ordenador, tales como micrófono o mouse.
    - **fabricante_cpu**: fabricante de la CPU.

6. **Agradecimientos.**

    Teniendo en cuenta que los datos proceden directamente de la web de Amazon, éste es el propietario de los mismos.

    La idea de esta práctica radica principalmente en el análisis del mercado online, debido al auge del mismo durante y a posteriori de la pandemia.

    Si tenemos en cuenta el aumento de demanda online debido al Covid, este set, así como otros que se pueden extraer de otras páginas web similares, el webscraping orientado a este mercado es especialmente útil, ya que permite una comparativa directa de productos (no necesariamente los que he propuesto en esta práctica, sino cualquier otro) entre distintas empresas del mismo sector y, por ende, una mayor posibilidad de comparar precios o atributos de los mismos artículos.

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.
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 el sentido de esta práctica, es decir, debido a que la creación de este dataset es puramente académico, considero que la licencia que mejor se adapta al mismo es CC0: Public Domain License. Esto es debido a que las demás licencias, como CC BY-NC-SA 4.0, que no permite un uso comercial del dataset y, como CC BY-SA 4.0, que mantiene la autoría del mismo, no tiene especial relevancia por lo mencionado anteriormente.

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

   El código ha sido adjuntado en Github: **https://github.com/sholod96/uoc-datascrapping-amazon.**

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

    El Dataset ha sido subido en Zenodo bajo el DOI de **10.5281/zenodo.5655733**.

    (*) 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).

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.