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.
2. Definir un título para el dataset. Elegir un título que sea descriptivo.
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).
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.
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.
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.
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
9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.
10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

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.

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.
   La información se ha obtenido del sitio web de documentación de Microsoft.
2. Definir un título para el dataset. Elegir un título que sea descriptivo.
   Información sobre la documentación de PowerBI
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 contiene información sobre los artículos de Microsoft relacionados con la aplicación PowerBI. La idea es poder determinar qué artículos se han actualizado recientemente de cara a estar al día en cuanto a conocimiento de la solución.
4. Representación gráfica. Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido

Artículo	Fecha	Tiempo	URL
What is Power BI? - Power BI | Microsoft Docs	03/29/2021	4 minutes to read	https://docs.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview
Sign up for the Power BI service as an individual - Power BI | Microsoft Docs	06/24/2020	7 minutes to read	https://docs.microsoft.com/en-us/power-bi/fundamentals/service-self-service-signup-for-power-bi
What is a Power BI "business user"? - Power BI | Microsoft Docs	04/02/2021	3 minutes to read	https://docs.microsoft.com/en-us/power-bi/consumer/end-user-consumer
Getting around in Power BI service - Power BI | Microsoft Docs	04/02/2021	7 minutes to read	https://docs.microsoft.com/en-us/power-bi/consumer/end-user-experience
What is a dashboard and how do I open it? - Power BI | Microsoft Docs	12/03/2020	2 minutes to read	https://docs.microsoft.com/en-us/power-bi/consumer/end-user-dashboards

5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.
   - Artículo – Título del artículo
   - Fecha – Fecha en la que se ha realizado la última modificación.
   - Tiempo – Tiempo de lectura estimado
   - URL – URL del site
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.
   Microsoft
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.
   La idea para realizar este web scraping surgió de la necesidad profesional de mantener actualizados los conocimientos en PowerBI. Aunque en el blog de la herramienta se avisa de las nuevas funcionalidades, muchas veces se realizan actualizaciones en la documentación sin tener un reflejo en el blog.
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

- La CC0 ya que debería ser libre de usarlo cualquiera.

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

```
10.     import requests
import csv
from bs4 import BeautifulSoup

url_principal = "https://docs.microsoft.com/en-us/power-bi/"

#Función que obtiene las url hijas de la página principal de documentación
def obtener_links(p_url):
    #Inicializamos la lista de salida
    lista_urls = []
    # Realizamos la llamada a requests
    response = requests.get(p_url)
    # Almacenamos el contenido de las web
    webpage = response.content
    # Creamos un objeto BeautifulSoup con el contenido
    soup = BeautifulSoup(webpage, "html.parser")
    # Buscamos los links y los almacenamos en una lista
    for i in soup.find_all('li'):
        for url in i.find_all('a'):
            hijo = url['href']
            #Descartamos las urls que no se basen en la url relativa de
nuestro interes
            if not hijo.startswith("https"):
                lista_urls.append(p_url+hijo)
    #Quitamos duplicamos de la lista
    lista_urls = list(dict.fromkeys(lista_urls))
    return lista_urls

#Esta función obtiene el título, tiempo estimado de lectura y fecha de
actualización de los artículos de documentación
def obtener_info(url):
    # Realizamos la llamada a requests
    response = requests.get(url)
    # Almacenamos el contenido de las web
    webpage = response.content
    # Creamos un objeto BeautifulSoup con el contenido
    soup = BeautifulSoup(webpage, "html.parser")
    # Buscamos los datos que queremos extraer y los almacenamos en una tupla
    try:
        # Tiempo de lectura
        tiempo = soup.find("li", {"class": "readingTime"}).string
        # Fecha de modificación
        fecha = soup.find("time").string
        # Título del artículo
        titulo = soup.title.string
        # Almacenamos los datos en una tupla
        resultado = (titulo, fecha, tiempo, url)
        return resultado
    except(ConnectionError, Exception):
        resultado_error=(None,None,None,url)
        return resultado_error

#Llamamos a la función que obtiene las urls y metemos los resultados en una
lista
```

```
salida=obtener_links(url_principal)

#Para cada url de la lista llamamos a la función que obtiene la info que
buscamos.
#Guardamos la información en una lista de tuplas
lista=[]
for web in salida:
    tupla=obtener_info(web)
    if tupla[0]!=None:
        lista.append(tupla)

#Guardamos la salida en un fichero
with open('D:/powerbidoc.csv','w', newline='') as out:
    csv_out=csv.writer(out,delimiter=';')
    csv_out.writerow(['Artículo','Fecha','Tiempo','URL'])
    for row in lista:
        csv_out.writerow(row)
```

11. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

    10.5281/zenodo.4699732