Práctica 1 (25% nota final)

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 el sitio web elegido, el código y el dataset que se indican más adelante. Se deberá presentar un documento PDF **(máximo 20 páginas)** en el que se resuelvan los siguientes apartados:

1. **Contexto.** Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información. Indicar la dirección del sitio web.

   En este proyecto, se ha decidido investigar sobre la red social de microblogging Twitter. La red permite enviar mensajes de texto plano de corta longitud (280 caracteres) que se muestran en la página principal del usuario.

   Los usuarios pueden suscribirse a las cuentas de otros usuarios –a esto se le llama **seguir** y a los usuarios abonados se les llama **seguidores**– para poder ver sus tweets en su página principal. Además, en la página principal, se muestran una serie de *Trending Topics*, que son los temas del momento o tendencias.

   Recientemente, Twitter ha sido comprado por el multimillonario Elon Musk (dueño de Tesla), lo que ha generado gran controversia, debido a sus declaraciones y las primeras medidas que ha tomado en relación a esta red social. Una de las primeras medidas tomadas por Musk fue la de pagar por el verificado de cuenta de Twitter, lo que ha provocado que una cuenta falsa obtenga el verificado y una farmaceútica caiga en la bolsa. En este escenario, nos parece interesante recabar una serie de datos con el objetivo de realizar un análisis de sentimientos para conocer la opinión de la gente sobre este nuevo propietario de Twitter. Esta información podría ser muy valiosa para el equipo de análisis de redes sociales, community managers o equipo de prensa de Elon Musk, cuyo objetivo es velar por su imagen.

   En este trabajo, se ha elegido la web del propio Twitter https://twitter.com para recoger información de los tweets que contienen la palabra “musk”.

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

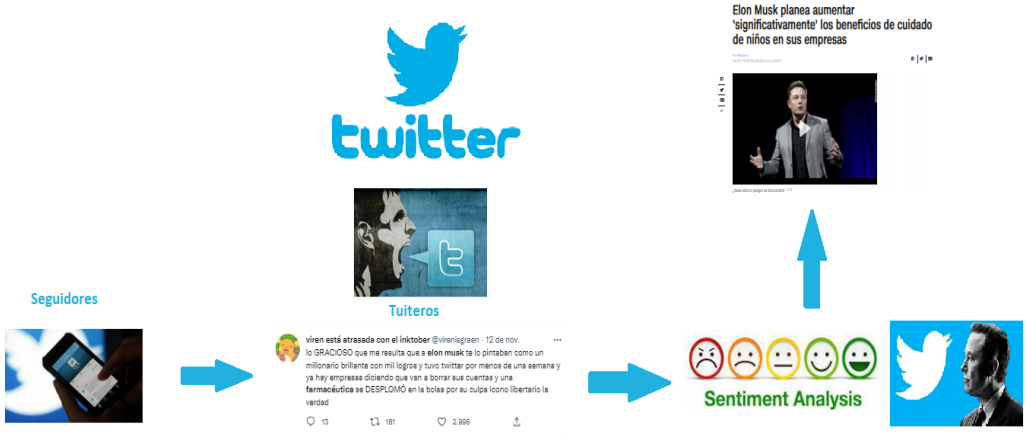
   Opinión pública y reputación online de Elon Musk: datos más relevantes para un análisis basado en sentimientos.

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 formado por datos basados en tweets sobre Musk y las interacciones en los mismos. En este dataset se presentan un total de 192 tweets publicados entre el 13 y 14 de noviembre de 2022. Las unidades o magnitudes de las características extraídas son en formato fecha y hora, texto, cantidad numérica y weblink. Además, se ha llevado a cabo un proceso de limpieza y preprocesado con el fin de que la información sea más comprensible.

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



5. **Contenido.** Explicar los campos que incluye el dataset y el periodo de tiempo de los datos.

   En este dataset, se presentan dichos datos para cada uno de los 192 tweets recogidos, que han sido publicados durante el 13 y 14 de noviembre de 2022. Las características extraídas son las siguientes:

   - **DateTime of Tweet**: Fecha y hora de publicación del tweet. Se trata de datos en formato fecha y hora.
   - **Replying to**: Número de respuestas a este tweet. Se trata de datos en formato numérico (enteros).
   - **Tweet**: Texto plano del tweet (280 caracteres máximo). Se trata de datos en formato texto.
   - **Likes**: Número de *"Me gusta"* que tiene el tweet. Cuenta el número de usuarios que pulsan el botón del icono de corazón. Se trata de datos

en formato numérico (enteros).

- **Retweet**: Número de veces que se hace *"Retweet"* al tweet en cuestión. El usuario que pulsa *"Retweet"* comparte el tweet en su página principal. Se trata de datos en formato numérico (enteros).

- **Image**: URL de la imagen insertada en el tweet. Se trata de datos que indican el enlace en el que se encuentra la imagen adjunta al tweet. Se trata de datos de texto.

Los datos se han recogido mediante técnicas de web scraping en lenguaje Python sobre cada uno de los tweets analizados en la página del sitio web.

El procedimiento llevado a cabo para este análisis es:

1. Se recorre la página de Twitter y se eligen varios tweets que incluyen la palabra "Musk", comprobando que no se produzcan duplicados.
2. Se extrae información relevante de cada tweet elegido. Sobre cada tweet, se aplica el scraping para recolectar la información deseada.
3. Se guardan los datos extraídos en un fichero CSV.

6. **Propietario.** 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. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

El propietario del sitio web actualmente es Elon Musk, fundador de SpaceX; CEO de Tesla, fundador de The Boring Company; cofundador de Neuralink y OpenAI, además de ser el actual dueño de Twitter.

En internet nos encontramos varios análisis y extracciones de datos de Twitter similares al nuestro:

- https://www.kaggle.com/code/marta99/elon-musk-s-tweets-sentiment-analysis
- https://pub.towardsai.net/how-to-scrape-tweets-without-twitters-api-using-twint-797b196b951c
- https://link.springer.com/chapter/10.1007/978-3-031-09176-6_70

En el primero se realiza un análisis de sentimiento de los tweets realizados por Elon Musk en 2022. El dataset utilizado es muy similar al nuestro. En el segundo enlace nos encontramos un código para hacer web scraping en Twitter utilizando la herramienta Twint. Por último, tenemos una publicación que analiza mediante modelos LTSM y SVM el impacto de los tweets publicados por Elon Musk en un horizonte temporal de 5 años en el mercados de monedas y la bolsa de valores.

Aunque Twitter dispone de una API propia para realizar web scraping de la información, hemos procedido a leer el archivo *robots.txt* para confirmar que la información extraída no se encuentra entre las restricciones de rastreado. Además,

el scraping de los datos se ha realizado sin iniciar sesión en la web, con lo cual, no se han tenido que aceptar términos ni condiciones; por lo que toda la información extraída es enteramente pública.
Por otra parte, se trata de datos utilizados con fines académicos, por lo que su uso es legítimo. Además, no se ha sobrecargado el servidor con muchas peticiones, ya que el dataset es de tamaño moderado y no se utiliza la información de forma injusta.

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.

   Es interesante analizar este conjunto de datos para conocer la opinión en redes sociales sobre la compra y gestión de Twitter por parte de Elon Musk. El análisis de sentimientos en los tweets que publica la gente sobre Musk, nos permiten saber si las menciones a él en Twitter son positivas, negativas o neutras. Los resultados de este análisis, podrían ser decisivos en la toma de decisiones de su equipo de cara a la futura gestión de Twitter, así como para destacar tendencias y poder crear campañas publicitarias o difundir noticias que mejoren su imagen.

   Este dataset nos proporciona tweets de la opinión pública, al contrario que los análisis presentados anteriormente, que proporcionan tweets del propio Elon, y no pueden responder a las mismas preguntas.

   Las preguntas que pretende responder este dataset son las siguientes:

   - ¿Qué se tuitea sobre Elon Musk en los últimos días?
   - ¿Cuáles son los tweets con más favoritos?
   - ¿Cuáles y cuántos son los tweets más compartidos?
   - ¿Cuáles son los hilos más largos hablando sobre Elon Musk?

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).

   La licencia escogida para la publicación de este conjunto de datos es CC BY-SA 4.0 License. Los motivos que han llevado a la elección de esta licencia tiene que ver con la idoneidad de las cláusulas que esta presenta en relación con el trabajo realizado:

- Se debe proveer el nombre del creador del conjunto de datos generado, indicando los cambios que se han realizado. De esta manera, se reconoce el trabajo ajeno y en qué medida se han realizado aportaciones en relación con el trabajo original.

- Se permite un uso comercial. Esto haría que incrementen las probabilidades de que una empresa utilice los datos generados y realicen trabajos de calidad que reporten cierto conocimiento al autor original.
- Las contribuciones realizadas a posteriori sobre el trabajo publicado bajo esta licencia deberán distribuirse bajo la misma. Esto hace que el trabajo del autor original continúe distribuyéndose bajo los términos que él mismo planteó.

9. **Código.** Código con el que se ha obtenido el 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 el documento 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 las habéis resuelto.

El código fue hecho en lenguaje Python con la implementación de la librería Selenium. El código fuente se encuentra dentro de la carpeta **source**.

Las dificultades del sitio web con las que nos hemos encontrado durante el web scraping y las soluciones a las mismas son las siguientes:

★ Se ha modificado la cabecera del user-agent con el objetivo de no ser bloqueados en la web.
★ Puesto que en Twitter existe la opción de iniciar sesión, se utiliza el objeto *session* para agregar las *cookies* de forma automática a las posteriores peticiones realizadas en el mismo sitio.
★ Se verifica que se cumplen las restricciones del archivo *robots.txt*
★ Se introduce un retardo entre peticiones con el objetivo de no colapsar el servidor web.
★ Se comprueba que no se capturen tweets duplicados (información duplicada).

10. **Dataset.** Publicar el dataset obtenido en formato CSV en Zenodo, incluyendo una breve descripción. Obtener y adjuntar el enlace del DOI del dataset (https://doi.org/…). El dataset también deberá incluirse en la carpeta **/dataset** del repositorio.

ZENODO y atributos del DOI	
DOI	10.5281/zenodo.7334461
URL ZENODO	https://zenodo.org/badge/DOI/10.5281/zenodo.7334461.svg
URL del DOI	https://doi.org/10.5281/zenodo.7334461
HTML	<img src="https://zenodo.org/badge/DOI/10.5281/zenodo.7334461.svg" alt="DOI">
TARGET URL	https://doi.org/10.5281/zenodo.7334461

Si existe alguna circunstancia que impida publicar abiertamente el dataset real en Zenodo, se deberá: (1) comentar esta circunstancia y justificar el motivo en este apartado; (2) generar un dataset simulado y publicarlo en Zenodo, obteniendo el enlace del DOI; y (3) comunicar al profesor el dataset real de forma privada (p. ej., utilizando un repositorio privado).

El dataset en formato CSV se encuentra dentro de la carpeta **dataset** del repositorio.

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.

Contribuciones	Firma
Investigación previa	ACF, DGG
Redacción de las respuestas	ACF, DGG
Desarrollo del código	ACF, DGG
Participación en el vídeo	ACF, DGG