RETO 4

Procesamiento del Lenguaje Natural

Análisis de redes sociales en torno a #SOSMarMenor

Contexto del caso práctico

En esta última práctica imaginaremos que tras acabar de forma brillante el curso de especialización en IA y Big Data consigues tu primer contrato como analista de datos en el grupo de comunicación "Valencia Plazoleta".

El jefe de sección de informativos se ha reunido contigo para encargarte tu primer trabajo, que consistirá en que aportes ciertos datos que permita al grupo de periodistas que tiene a su cargo realizar un reportaje sobre la crisis medioambiental del Mar Menor. En particular, necesitan hacerse una idea de qué ha pasado meses atrás, quiénes son los protagonistas y cuáles son los episodios más relevantes en torno a los cuales se pueda elaborar un relato periodístico de la crisis.

Según te ha trasladado el jefe de sección, les gustaría abordar informativamente qué tipo de respuesta ha dado la ciudadanía a la crisis medioambiental que sufre desde hace años la región de Murcia. Es posible que hayan manifestado su descontento con la situación a través de manifestaciones, protestas, apariciones en medios de comunicación, etc. en diferentes momentos. Como es lógico, cuanto más clara, exhaustiva y útil sea la información que puedas proporcionarles en torno a todo ello, mejor será el reportaje que podrán elaborar después.

Tu objetivo será presentar un informe (cuaderno Jupyter) con las conclusiones que puedas sacar en torno al caso planteado, que deberás argumentar con datos y siempre que se pueda de la forma más atractiva posible.

Tareas a realizar

Como sabemos, cualquier proyecto de *Machine Learning* implica resolver un problema que podemos formular como pregunta, y las aproximaciones desde el PLN no lo son menos. Dado el caso en cuestión, la pregunta a la que deberíamos dar respuesta podría ser: **¿qué podemos aprender de lo que ha sucedido en torno a la crisis del Mar Menor estos últimos años partiendo de la actividad existente en la redes sociales?**

Responder a este pregunta supone dos tareas básicas:

a) contar con un conocimiento básico de cuál es el contexto de la problemática en cuestión. A tal fin, te recomiendo que empieces revisando este vídeo de no más de 2 minutos. Si quieres información más detallada puedes revisar el artículo de la versión online del *National Geographic* titulado "El desastre del Mar Menor, historia de un colapso ambiental que pudo haberse evitado".

b) cabe decidir cuál será la fuente (o fuentes) de la que se extraerá la información. En este caso cuentas con un dataset con datos extraídos de Twitter que recoge la actividad desarrollada en torno al *hashtag* **#SOSMarMenor**.

A partir de este momento, cabe desarrollar un *pipeline* que permita contestar a la pregunta.

Las tareas a realizar se realizan con 2 objetivos: preparar los datos para su procesamiento y análisis posterior, y comprender los datos que vienen en el dataset como parte de ese análisis:

a) **exploración de datos**

En primer lugar, resulta recomendable realizar una revisión superficial de los datos como primera aproximación a su comprensión. Aplica las técnicas que te permita obtener algo de información sobre el dataset.

b) **preprocesamiento de texto**

Como sabemos, preprocesar un texto supone realizar diferentes tareas que lo dejen preparado para tareas posteriores. En este caso, como quizá puede ser útil no eliminar tokens referidos a otros *hashtags* o nombres de usuario (que pueden ser colectivos u organizaciones ciudadanas) que aparezcan en los mismos como parte de su contenido, ya es contenido que nos interesa preservar, puede ser aconsejable guardar los tuits preprocesados de 2 formas: una en cuyo limpiado se incluyan *hashtags* y nombres de usuarios, y otra en la que se preserven como parte del tuit preprocesado.

c) **análisis de datos** ("Exploratory Data Analysis"):

En un tercer momento, puede ser interesante explorar en términos cuantitativos como parte de un análisis más detallado de los datos. Como guía a los aspectos en los que puedes indagar, trata de responder en el cuaderno las siguientes preguntas, argumentando tu respuesta:

- ¿En qué períodos ha existido más actividad ciudadana en relación al *hashtag* #SOSMarMenor?, ¿qué ocurrió en esos momentos?
- ¿Cuántos usuarios han generado tuits en relación al hashtag #SOSMarMenor?
- ¿Cuáles son los usuarios más activos?, ¿son partidos políticos, ONGs o plataformas ciudadanas, o personas particulares?
- ¿Cuáles han sido los tuits más celebrados (i.e. retuiteados)?, ¿qué tipo de mensajes transmiten (de protesta, de indignación, de ánimo, de solidaridad)?

  En relación a esta última pregunta, puedes echar mano a efectos ilustrativos de la siguiente función, que recibe un nombre de usuario y el identificador de un tuit, y muestra el tuit en cuestión por pantalla.

```
from IPython.display import HTML
import requests

# Muestra un tuit a partir de su id y del nombre del usuario
def show_tweet(user, tweet_id):
   url = 'https://twitter.com/' + user + '/status/' + str(tweet_id)
   url_to_json = 'https://publish.twitter.com/oembed?url=%s' % url
   response = requests.get(url_to_json)
   html = response.json()["html"]
   display(HTML(html))
```

Responde a las preguntas anteriores en tu cuaderno, aportando los datos necesarios y eligiendo la forma más adecuada de mostrarlos.

d) **comparación de modelos para la identificación de temas**

Más allá de la información y conclusiones obtenidas en el apartado previo, una forma de identificar las temáticas en torno a la crisis del Mar Menor sería realizar un análisis de tipo léxico y temático del texto que contiene los tuits.

En particular, se propone en relación a ello la realización de 3 tareas:

1. elabora una nube de términos (**wordcloud**) para cada uno de los períodos identificados en el apartado anterior (la nube de términos debería tener en cuenta únicamente sustantivos).

2. aplica la técnica conocida como ***topic modeling*** al contenido de los tuits. La aplicación de esta técnica debería ayudarte a ser capaz de identificar los temas en torno a los cuales gira la conversación ciudadana en Twitter y los términos asociados a los mismos. Valora el uso de la librería pyLDavis para la visualización de datos.

3. aplica en segundo momento la técnica del **K-Means** para evaluar si se obtienen unos resultados parejos a lo del topic modeling, o bien se apartan de los mismos.

   Compara los resultados obtenidos aplicando cada modelo en la Identificación de posibles temas en el *corpus* de tuits, aportando los datos necesarios y eligiendo la forma más adecuada de mostrarlos.

Recursos para consulta

- **Introducción al topic modeling con Gensim (I): fundamentos y preprocesamiento de textos**
  - https://elmundodelosdatos.com/topic-modeling-gensim-fundamentos-preprocesamiento-textos/
- **Introducción al topic modeling con Gensim (II): asignación de tópicos**
  - https://elmundodelosdatos.com/topic-modeling-gensim-asignacion-topicos/
- **Improving the Interpretation of Topic Models**
  - https://towardsdatascience.com/improving-the-interpretation-of-topic-models-87fd2ee3847d
- **pyLDAvis: Topic Modelling Exploration Tool That Every NLP Data Scientist Should Know**
  - https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know

Entrega y evaluación

- La entrega de la práctica deberá incluir un fichero tipo **cuaderno Jupyter** (con extensión "ipynb"). El cuaderno será el informe que presentarías al jefe de sección de la empresa, y debe dar respuesta de una forma argumentada a las preguntas y apartados previos (es decir, debe combinar código con la interpretación que haces de resultados). Asegúrate antes de hacer la entrega que el cuaderno carga y visualiza las gráficas que incluya una vez abierto.

- La fecha límite de entrega será el **25 de Febrero de 2024**.

- La evaluación del ejercicio se hará conforme a los criterios recogidos en la siguiente rúbrica, pudiendo alcanzar un total de 10 puntos como máximo.

Rúbrica de evaluación

RETO 4 :: Análisis de redes sociales en torno a #SOSMarMenor				
Criterios	**Puntuación**			
	0	**1**	**2**	**3**
Análisis de datos	Se dejan varias preguntas del enunciado sin contestar o son contestadas gran parte de ellas incorrectamente / No se aportan los datos solicitados	Se contesta correctamente sólo a algunas de preguntas planteadas en el enunciado	Se contesta correctamente a la mayoría de preguntas planteadas en el enunciado	Se contesta correctamente a la totalidad de preguntas planteadas en el enunciado
Aplicación del *Topic modeling*	No se aplica el modelo, o si se hace, se obtiene un conjunto de temas que no responden a lo esperado, ni tampoco se usan muestran los datos visualmente de la forma apropiada	O bien no se aplica el modelo, o si se hace, se obtiene un conjunto de temas que no responden en gran medida a lo esperado; o bien no se usan de recursos de visualización	Se aplica el modelo, visualizándolo los datos adecuadamente, pero o bien identificando parcialmente los temas, o bien no haciendo una interpretación del todo correcta de los mismos,	Se aplica el modelo identificando acertadamente los temas y haciendo una interpretación correcta de los mismos
Aplicación de K-Means	No se aplica el modelo, o si se hace, se obtiene un conjunto de temas que no responden a lo esperado	Se aplica el modelo, pero sin llegar a interpretar los datos, o no haciéndolo correctamente	Se aplica el modelo, interpretando los datos correctamente, comparándolos con los del Topic Modeling	—
Presentación	No se presentan por lo general los resultados o conclusiones (textual, gráfica o ambas) de una forma adecuada	Se elige la forma más adecuada de presentar los resultados o conclusiones (textual, gráfica o ambas) sólo en algunos casos	Se elige la forma más adecuada de presentar los resultados o conclusiones (textual, gráfica o ambas) en la mayoría o totalidad de casos	---