TUIA NLP 2023
TRABAJO PRÁCTICO 1

Pautas generales:

- El trabajo deberá ser realizado en grupos de tres o excepcionalmente dos integrantes.
- Deberá informar cómo está conformado su equipo de trabajo y cuál es la url del repositorio con el que van a trabajar en el siguiente formulario: https://forms.gle/pCbd6bWRWmQFcF5y7
- Considerar que todos los integrantes deben conocer todos los aspectos del trabajo entregado, ya que será evaluado luego de la entrega de forma oral.
- Se debe entregar un informe en el cual se incluya las justificaciones y un vínculo a los archivos que permitan reproducir el proyecto. Recomendamos **gitlab o github** para tal fin, si se realiza en colab entregar en el formato de Jupyter Notebook **.ipynb**, dentro de un repositorio. Además recomendamos se utilice **virtualenv, venv** o similar para registrar las dependencias de las distintas partes del proyecto.

Ejercicio 1:

Construir un dataset haciendo web scraping de páginas web de su elección.

- Definir 4 categorías de noticias/artículos.
- Para cada categoría, extraer los siguientes datos de 10 noticias diferentes:
    - url (sitio web donde se publicó el artículo)
    - título (título del artículo)
    - texto (contenido del artículo)

Recomendaciones: elegir blogs para evitar los límites de lectura para los medios que exigen suscripción. Investigue sobre el archivo *robots.txt* y téngalo en cuenta. Considere también espaciar las consultas para evitar saturar el sitio.

Utilizando los datos obtenidos construya el dataset en formato csv.

Ejercicio 2:

Utilizando los datos de título y categoría del dataset del ejercicio anterior, entrenar un modelo de clasificación de noticias en categorías específicas.

Ejercicio 3:

Para cada categoría, realizar las siguientes tareas:

- Procesar el texto mediante recursos de normalización y limpieza.
- Con el resultado anterior, realizar conteo de palabras y mostrar la importancia de las mismas mediante una nube de palabras.

Escribir un análisis general del resultado obtenido.

Ejercicio 4:

Use los modelos de embedding propuestos sobre el final de la Unidad 2 para evaluar la similitud entre los títulos de las noticias de una de las categorías.

Reflexione sobre las limitaciones del modelo en base a los resultados obtenidos, en contraposición a los resultados que hubiera esperado obtener.

Ejercicio 5:

Escriba un programa interactivo que, según la categoría seleccionada por el usuario, devuelva un resumen de las noticias incluidas en ella.

Justifique la elección del modelo usado para tal fin.

Opcional: Investigar y programar un bot de Telegram que entregue un resumen de noticias del blog de su elección. Recomendamos el uso de pyTelegramBotAPI.