PROCESAMIENTO DEL LENGUAJE NATURAL
TRABAJOS PRÁCTICOS

- 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.

ESTO SE ENCUENTRA RESUELTO EN **TP_NLP.ipynb**

EJERCICIO:
Mediante un proceso de web scraping genere un dataset de datos sobre libros de la siguiente web: Lectulandia

El dataset debe tener de un mínimo de 100 libros con los siguientes datos:

- género literario (mínimo 10)[1]

[1] Nota: tener en cuenta que un libro puede pertenecer a más de un género.

- autor
- título
- síntesis de cada libro

Propósito
El objetivo de este proyecto es desarrollar un programa que interactúe con el usuario para recomendar lecturas. El programa ofrecerá tres opciones principales:

Recomendación Directa: El programa preguntará al usuario **"¿Qué tienes ganas de leer hoy?"** y, mediante la clasificación de la respuesta, propondrá una lista de tres libros acordes a las temáticas mencionadas. Además, se detallará el autor, género y una breve reseña de cada libro que se recomienda descargar.

Elección por Autor: Si el usuario prefiere buscar por autor, el programa ofrecerá una lista de libros del autor especificado. En caso de que haya múltiples resultados, se basará en la similitud de los dos primeros resultados más relevantes, retornando dos títulos (con sus respectivas reseñas) del resultado más cercano y uno del segundo que se recomiendan para descargar. Si existen varios libros del mismo autor aleatorizar.

Elección por Género Literario: Similar a la búsqueda por autor, si el usuario elige buscar por género, el programa ofrecerá una lista de libros del género especificado. Aplicará la misma lógica de similitud para seleccionar y presentar los resultados.

ESTO SE ENCUENTRA RESUELTO EN **NLP_tp.ipynb**