PROCESAMIENTO DE LENGUAJE NATURAL
TRABAJO DE FIN DE UNIDAD

1. Busque en la colección Gutenberg (https://www.gutenberg.org/) los 3 tomos en español de "Las mil y una noches" (autor anónimo)
2. Recupere la información desde las páginas web (web scraping).
3. Elimine de los textos todo el contenido que no pertenezca a la obra (links, información sobre el proyecto Gutenberg, índices, caracteres de adorno, líneas en blanco, etc.)
4. Junte los tres tomos en un solo texto (textoRAW)
5. Crear los módulos necesarios para:
   a. Recuperar una historia por su título.
   b. Dividir (tokenizar) el texto por historias.
   c. Dividir (tokenizar) el texto por párrafos.
   d. Dividir (tokenizar) el texto por oraciones.
   e. Tokenizar el texto en palabras.

   Realice las pruebas pertinentes de cada módulo
6. Crear lo siguientes módulos independientes para normalizar un texto:
   a. Quitar signos de puntuación
   b. Convertir a minúsculas
   c. Quitar stop-words
   d. Quitar palabras con extensión mínima (valor por defecto del parámetro: 2)
   e. Lematizar
   f. Eliminar duplicados

   Probar **diferentes secuencias** de los módulos, por ejemplo:

   Secuencia 1:
   a. Quitar signos de puntuación
   b. Convertir a minúsculas
   c. Quitar stop-words
   d. Quitar palabras con extensión mínima
   e. Lematizar
   f. Eliminar duplicados

   Secuencia 2:
   a. Quitar signos de puntuación
   b. Convertir a minúsculas
   c. Quitar palabras con extensión mínima
   d. Quitar stop-words
   e. Lematizar
   f. Eliminar duplicados

   Secuencia 3:
   a. Lematizar
   b. Quitar signos de puntuación
   c. Convertir a minúsculas
   d. Quitar stop-words
   e. Quitar palabras con extensión mínima
   f. Eliminar duplicados

   Definir dos secuencias adicionales.
   Realizar un cuadro comparativo de las 5 secuencias (número de tokens) y analizar las diferencias de los resultados alcanzados ¿Qué conclusiones se pueden extraer del experimento?
7. Convertir los textos planos (textoRAW) a un corpus propiamente dicho, mediante el uso de librerías como PlaintextCorpusReader. Comparar los resultados obtenidos con lo

implementado en los puntos 5 y 6 (realizar las pruebas necesarias para la comparación, explicando adecuadamente la secuencia).

8. Normalizar textoRAW con alguna de las secuencias probadas en el punto 6 y luego escribir módulos para poder generar:
   a. Bolsa de palabras (BoW)
   b. Vectores de palabras de la BoW (mediante one hot encoding, vectores de frecuencia y tf-idf)
   c. Vocabulario etiquetado (PoS).
   d. Lista de adjetivos, verbos, sustantivos y adverbios más comunes (mostrar los 5 primeros en cada caso)
   e. Lista de nombres de entidades (NER).

   Realice la prueba con una historia específica.
9. Implementar los siguientes módulos (considerando tf-idf)
   a. Palabras con mayor peso en cada documento
   b. Palabras con mayor peso en el corpus (considerar el promedio de tf-idf de cada palabra).

   Haga un estudio de los resultados de dichos módulos considerando todo el corpus dividido en historias (normalizar previamente el corpus).
10. Implemente algún módulo que considere interesante, ponga un ejemplo de su funcionalidad y explique el resultado.

INSTRUCCIONES:

1. El trabajo será realizado en grupos de dos estudiantes (los estudiantes deben pertenecer al mismo grupo de laboratorio).
2. Pueden utilizar cualquiera de las herramientas vistas en laboratorio. De utilizar otras librerías documentar adecuadamente su uso y origen.
3. El trabajo será entregado en la plataforma Classroom en un solo archivo COLAB debidamente organizado y documentado. Los trabajos que no cumplan con la debida organización y documentación, no serán calificados.
   En la primera parte del archivo se debe consignar los apellidos, nombres y códigos de los dos estudiantes (en orden alfabético por apellido paterno). El primer estudiante es el único que debe entregar el archivo en el CLASSROOM.
4. En caso de detectarse copias entre los trabajos, se considerará una calificación de CERO puntos para todos los involucrados.
5. Fecha de entrega: viernes 18/11/2022