Laboratorio 2: Recuperación de Información

Objetivo: Aprender a preparar los textos para que sean de utilidad en el proceso de recuperación de información. Para ello deberás separar el texto en tokens, eliminarse los tokens inútiles (signos de puntuación, números), palabras vacías, convertir a minúsculas y truncar las palabras.

Materiales:

Instalar Python y NLTK

Descripción:

1. Utiliza el mismo e-book en texto plano del Lab1 (Around the World in Eighty Days, by Jules Verne)

   https://www.gutenberg.org/files/103/103-0.txt

2. Realiza el mismo pre-procesamiento de la práctica 1 y como último paso trunca las palabras utilizando el algoritmo de Porter Stemming incluido en NLTK.

   ```
   from nltk.stem.porter import PorterStemmer
   ```

   Escribe las 100 primeras palabras. ¿Qué observas? ¿Hay cambios?

3. Escribe tus observaciones.

4. Investiga cómo utilizar:

a) nltk.stem.SnowballStemmer()
b) nltk.wordnet.WordNetLemmatizer()

5. Ejecútalos sobre el mismo e-book escibe las 100 primeras palabras, compara las tres herramientas y obtengan sus conclusiones. Si alguna de ellas tarda demasiado considera elegir una porción del texto.

6. Escribe tu reporte en el formato establecido y colócalo en Teams.