Laboratorio 1: 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 y convertir a minúsculas.

Materiales:

Instalar Python y NLTK

Descripción:

1. Descarga el e-book en texto plano Around the World in Eighty Days, by Jules Verne https://www.gutenberg.org/files/103/103-0.txt

2. Lee el archivo que descargaste

```
load text
filename = 'xxxxxx'
file = open(filename, 'rt')
text = file.read()
file.close()
```

3. Separa el texto por palabras utilizando el espacio como delimitador y escribe las primeras 100 palabras

```
words = text.split()
print(words[:100])
```

   ¿Qué observas en estas palabras?

4. Utiliza re de python para eliminar signos de puntuación. Lee en la web como funciona re en python, escribe un breve resumen de lo que entendiste. Ahora prueba las siguientes líneas, asegúrate de entenderlas y hacer cambios si es necesario.

```
import re
words = re.split(r'\W+', text)
print(words[:100])
```

   Busca una mejor manera de eliminar los signos de puntuación. Prueba:

```
print(string.punctuation)
Utiliza este método junto con re
re_punc = re.compile('[%s]' % re.escape(string.punctuation))
stripped = [re_punc.sub('', w) for w in words]
print(stripped[:100])
```

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

Escribe los cambios que ves en la separación de las palabras, asegúrate de entender el código. Ejecuta un ejemplo más utilizando re (puedes elegir la situación para ilustrar su utilidad), de acuerdo a lo que leíste. Coloca el código de tu prueba y el resultado.

5. Convierte a minúsculas el texto, utiliza lower(), lee como funciona este método y escribe para que sirve. Utilízalo y escribe nuevamente las 100 primeras palabras. Escribe tus observaciones.
6. Ahora elimina palabras vacías
    a) Importa las palabras vacías en idioma inglés
    b) “from nltk.corpus import stopwords
    c) stop_words = stopwords.words('english')”
    d) Imprime todas las palabras vacías, revísalas, y escribe en tu reporte las primeras 5.
    e) Elimina palabras vacías y escribe nuevamente las 100 primeras palabras.
7. Escribe tu reporte en el formato establecido y colócalo en Teams.