**Laboratorio 2: Recuperación de la Información**

Agis, Jaime., Reyes, Alfonso., y Torres, Nancy.

Benemérita Universidad Autónoma de Puebla

Recuperación de la información

Fecha: 30 de agosto de 2022

**Resumen:** En esta práctica se busca implementar los conocimientos adquiridos en la practica **Laboratorio 1: Recuperación de la Información** dando continuidad a la misma ahora trabajando con nuevas herramientas de la distribución implementando el algoritmo de Porter Stemming incluido en NLTK entre otras funciones presentadas en el punto **III. DESARROLLO EXPERIMENTAL.**

**Palabras clave:** Tokens, Python, NLTK, texto, Porter Stemming.

**I. INTRODUCCIÓN**

Se utiliza el texto “Around the World in Eighty Days, by Jules Verne” para dar continuidad a la práctica de laboratorio anteriormente realizada se deben de utilizar las herramientas implementadas en la práctica 1, así como implementar ahora el algoritmo de Porter Stemming incluido en NLTK.

Comprender y utilizar los siguientes recursos ejecutando las mismas realizando diferentes pruebas para comprender el funcionamiento:

* nltk.stem.SnowballStemmer()
* nltk.wordnet.WordNetLemmatizer()

**II. OBJETIVO Y PLANTEAMIENTO DEL PROBLEMA**

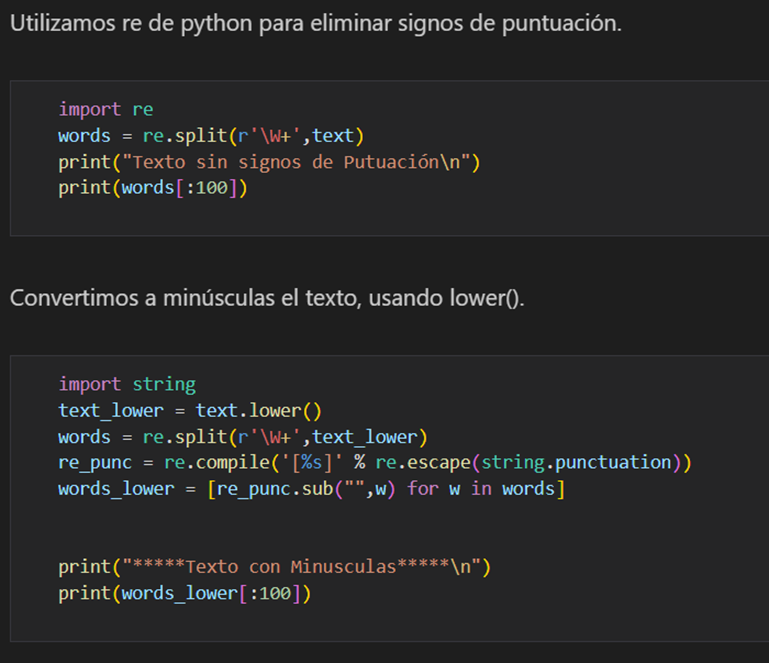
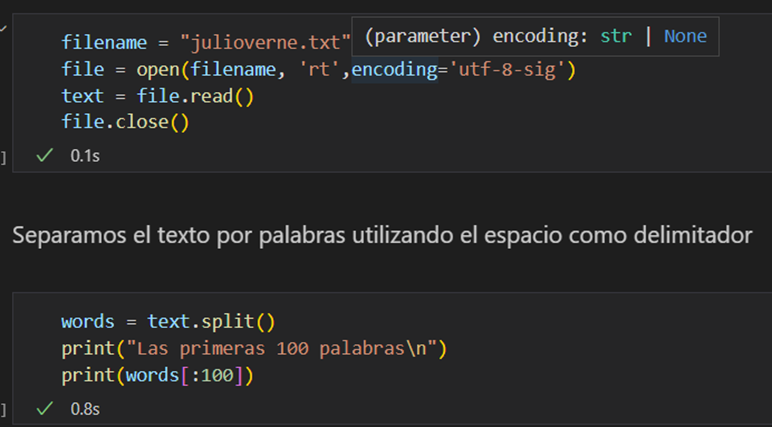
Aprender a preparar los textos para que sean de utilidad en el proceso de recuperación de información. Para ello deberás separar el texto en tokens, eliminar los tokens inútiles (signos de puntuación, números, palabras vacías y convertir a minúsculas.

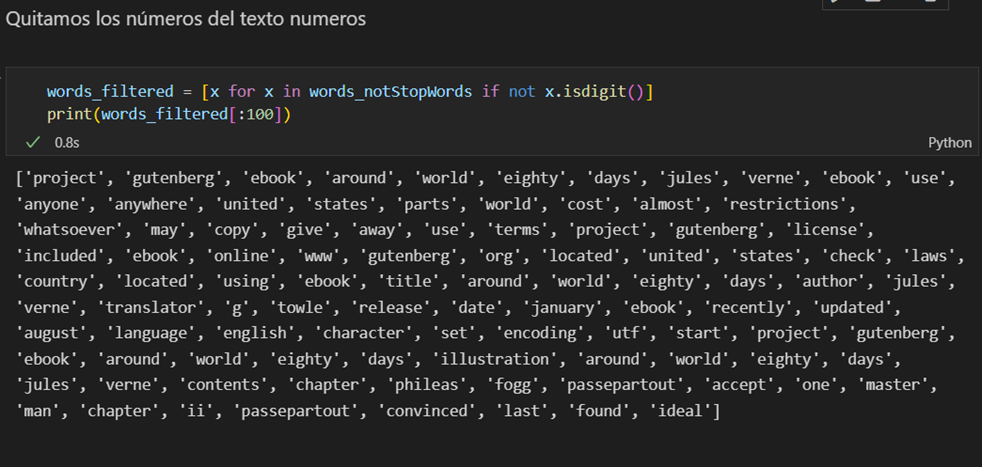
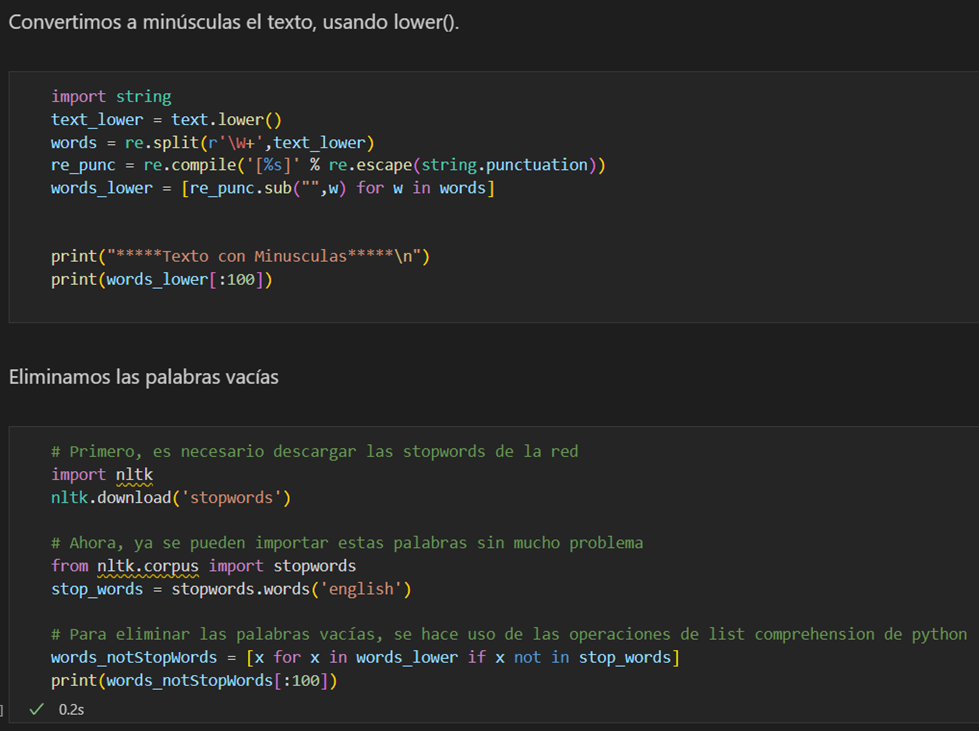
**III. DESARROLLO EXPERIMENTAL**

1. Utiliza el mismo ebook en texto plano del Lab1 (Around the World in Eight Days by Jules Verne) https://www.gutenberg.org/files/103/103-0.txt

2. Realizar el mismo preprocesamiento de la práctica 1 y como último paso trunca las palabras utilizando el algoritmo de Porter Stemming incluido en NLTK.

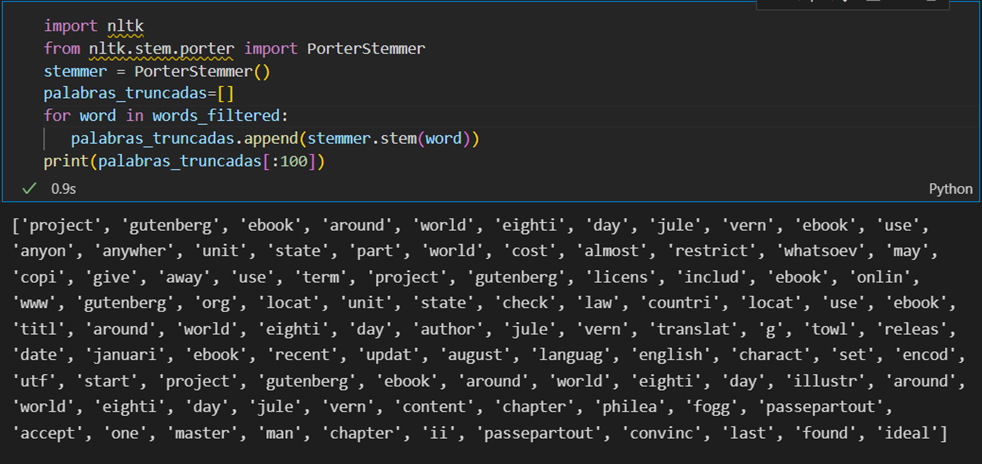
Hacemos el preprocesamiento de la Práctica 1





Importamos el algoritmo Porter Stemming incluido en NLTK

Escribimos las primeras 100 palabras ¿Que observas? ¿Hay cambios?



3. Escribe tus observaciones.

Como podemos observar se han quitado las últimas partes de algunas palabras, ya que Stemming o el truncamiento de las palabras es un método de normalización de palabras en el procesamiento del lenguaje natural. En este método, se normalizan las palabras que tienen el mismo significado, pero tienen algunas variaciones según el contexto o la oración.

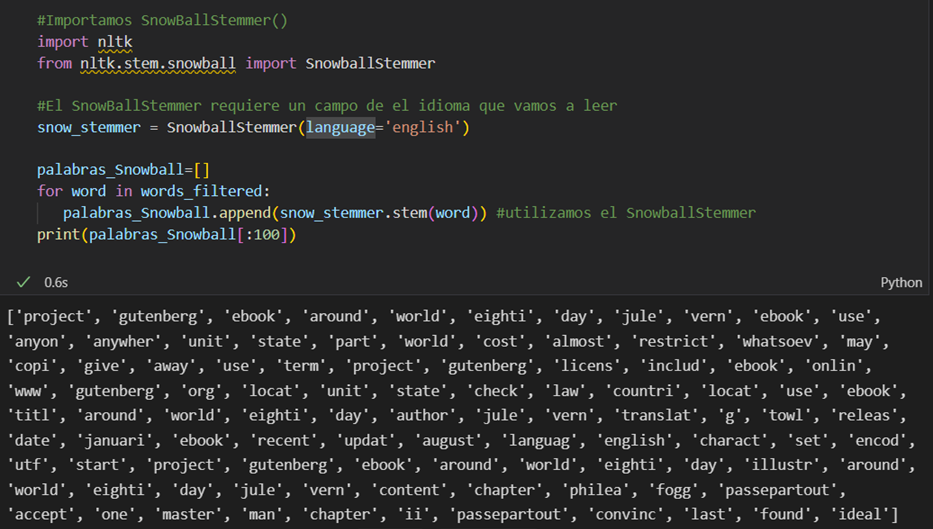
Por ejemplo:

* eighty ----- eighti
* days --------- day

4. Investiga cómo utilizar

a) nltk.stem.SnowBallStemmer()

El Snowball Stemmer: Es un algoritmo de derivación que también se conoce como el algoritmo de derivación de Porter2, ya que es una versión mejorada de Porter Stemmer.



b) nltk.wordnet.WordNetLemmatizer()

La lematización es el proceso de convertir una palabra a su forma base. La diferencia entre el Stemming y la lematización es que la lematización considera el contexto y convierte la palabra a su forma base significativa, mientras que la derivación simplemente elimina los últimos caracteres, lo que a menudo genera significados incorrectos y errores ortográficos.



5. Ejecutalos sobre el mismo ebook escribe las 100 primeras palabras, compara las tres herramientas y obtengan sus conclusiones. Si alguna de ellas tarda demasiado considera elegir una porción del texto.

**IV. DISCUSIÓN Y RESULTADOS**

Al comparar las tres herramientas concluimos que el algoritmo Porter Stemmer y el de SnowballStemmer hacen prácticamente los mismos cambios al texto. En cambio, el WordNetLemmatizer es menos agresivo, ya que muchas palabras no cambiaron a comparación de los otros algoritmos, pero no se generaron errores ortográficos e incorrectos.

Si eliminamos las palabras repetidas (generadas por el proceso de segmentación/lematización) podemos ver las diferencias entre las longitudes de los vocabularios generados.



Si vemos, al utilizar los 3 algoritmos se consigue una reducción en el vocabulario.

Vemos claramente el cambio de “agresividad” entre los algoritmos de porter y snowball, y lo noble que es el algoritmo de lematización, pero también en casos necesario para evitar posibles errores ortográficos generados en el truncamiento de las palabras.

**V. CONCLUSIONES**

En el proceso de recuperación de información, ya que se realizó el proceso de eliminar las palabras vacías, es vital reducir lo más posible el vocabulario a utilizar en una recuperación de la información. Con la utilización de algoritmos de segmentación o lematización, podemos ver una clara disminución del vocabulario final para culminar así el paso del preprocesamiento de un documento.

**VI. BIBLIOGRAFÍA**

NLTK Project. (2022). *NLTK :: Sample usage for stem*. NLTK Documentation. Recuperado 30 de agosto de 2022, de https://www.nltk.org/howto/stem.html