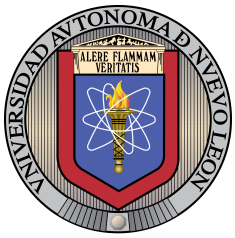

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Procesamiento y Clasificación de Datos

Gloria Samanta Servín García
1731703

PROFESORA

Mayra Cristina Berrones Reyes

MATERIA

Tarea 1

19 de mayo de 2022

Introducción

El propósito de esta actividad consiste en utilizar de forma apropiada la extracción de palabras clave, stopwords, lematización, y stemming para hacer un preprocesamiento de la información.Además, se deben utilizar gráficos adecuados para representar la información, y describir de manera general cual puede ser el propósito de la base de datos.

Planteamiento

Se obtiene del Proyecto Gutenberg: https://www.gutenberg.org/ un archivo .html que contiene el libro "Mujercitas", escrito por Louisa May Alcott y publicado el 30 de septiembre de 1868. Tomando experiencias propias, la escritora nos presenta las aventuras dentro de la vida de cuatro niñas que, tras pasar la adolescencia con la Guerra Civil en los Estados Unidos como fondo, se convierten en mujeres. Mediante el uso de herramientas de preprocesamiento de datos se pretende encontrar las palabras más significativas dentro de la novela.

Descripción de las técnicas y métodos empleados

Utilizando Google Colaboratory, tomamos de la cuenta personal de drive el archivo .html. En una primera instancia revisamos un extracto del texto para asegurarnos que el documento es el adecuado y es posible trabajarlo, una vez que hayamos confirmado esto es posible continuar con el proceso de limpieza de datos. La limpieza de nuestros datos consiste en separar el escrito por palabras y transformarlas todas a minúsculas para poder identificar las "stop words"que son palabras utilizadas en el cuerpo del documento pero no tienen significancia alguna vistas de manera individual. Al contar con las palabras relevantes se continua a seleccionar la representante de todas las formas flexionadas de una misma palabra para así finalizar el preprocesamiento de los datos.

Procedemos entonces a buscar palabras significativas dentro de Mujercitas, comenzando con las diez palabras con mayor frecuencia en el libro

```
[('jo', 1362),
 ('said', 827),
 ('little', 730),
 ('one', 725),
 ('meg', 686),
 ('amy', 652),
 ('laurie', 598),
 ('like', 591),
 ('beth', 467),
 ('good', 462)]
```

Figura 1: Palabras más mencionadas en el libro

Es posible ver que dentro de este top 10 se encuentran los nombres de los personajes principales de la historia, así que se creó una gráfica de pastel para ver el porcentaje de aparición que tienen entre sí mismos

Figura 2: Gráfica de pastel de los personajes con mayores apariciones

Continuamos con un gráfico que muestra la frecuencia que tienen las palabras con mayor presencia en el libro

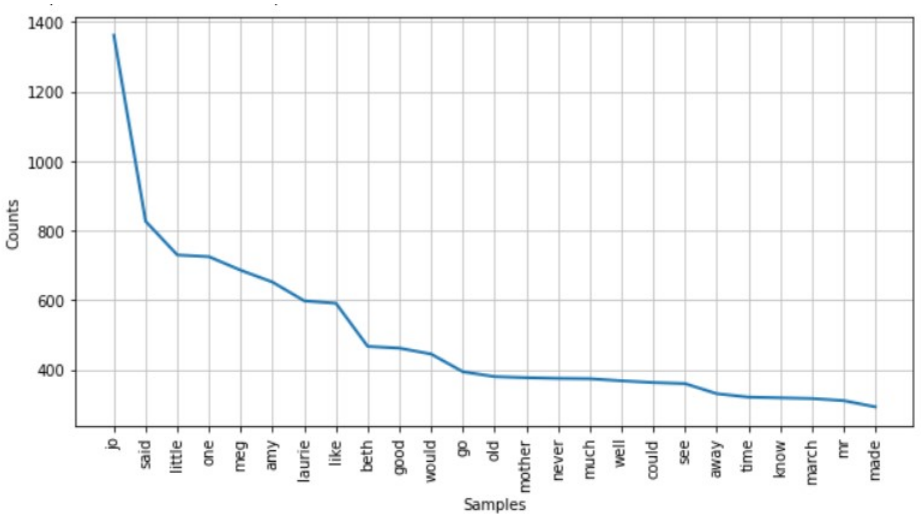

Figura 3: Gráfica de frecuencia por palabra

Para finalizar, se generó una imagen donde se pueden apreciar las palabras que cuentan con más frecuencia dentro de la historia, mientras más grande sea su tamaño, más veces se presentó en la lectura.

Figura 4: Representación visual de palabras con mayor frecuencia en el libro

Conclusiones

Gracias el apoyo de los gráficos generados en esta actividad así como del top 10 y la vista de palabras, podemos determinar fácilmente que el personaje principal de la historia es Jo, también se puede inferir quiénes son sus acompañantes dentro del relato. En general, el desarrollo de este análisis arrojó la relevancia de los personajes más que de otras palabras que nos puedan dar indicios sobre de qué trata la historia. Se espera continuar este ejercicio en sesiones futuras para poder determinar el tono de la novela así como otras características que nos permitan pintar una imagen clara de este clásico literario.

Github

ClarasDinner
https://github.com/ClarasDinner/MCD_Procesamiento