REPORTE 14 de Marzo:

Vocabulary Challenge - LINK COLAB

Explicacion del código

En la estandarización se realiza lo siguiente:

- Convierte el texto a minúsculas
- Elimina los acentos
- Elimina la puntuación y los números
- Elimina los espacios extra

Creación del vocabulario:

1. **WordEntry**: Se define una clase para representar las palabras del vocabulario y su frecuencia.
2. **Vocabulary**: Se efine una clase para el diccionario de palabras y sus frecuencias, con un método para agregar palabras o incrementar sus frecuencias.
3. **Procesamiento de textos**: Se recorre cada texto en el dataframe, divide los textos en palabras y las agrega al vocabulario usando las clases previamente definidas.
4. **Generación del dataframe**: Después de procesar todos los textos, se genera un dataframe con las palabras y sus frecuencias.

Estadisticas del vocabulario:

1. Se calcula el total de palabras en todos los textos de la columna del dataframe original. Se utiliza split para dividir el texto en palabras y se mide la longitud para contar cuántas palabras tiene cada texto.
2. Se obtiene el número de palabras únicas en el vocabulario ya que corresponde con el número de filas en el vocabulario.
3. Se ordenan las palabras por frecuencia de manera descendente (ascending=False) y se obtienen las 100 palabras más frecuentes.
4. Se ordenan las palabras por frecuencia de manera ascendente (ascending=True) y se obtienen las 100 palabras menos frecuentes.

Conclusiones

Mediante la comparación de resultados con los compañeros se observó que todos obtuvimos respuestas diferentes en cuanto a cantidad total de palabras y conteo de palabras diferentes, lo que esta relacionado con el preprocesamiento del texto, es decir, si se retiraron los acentos, los signos de puntuación, etc.

Respecto a las palabras mas y menos frecuentes los resultados entre los compañeros fueron mas homogéneos al menos en cuanto al ranking en el cual se colocaron las palabras.

Un vocabulario permite:

- Contar las palabras y frecuencias de manera eficiente.
- Preparar datos de manera adecuada para modelos de aprendizaje automático.
- Controlar el preprocesamiento de texto.
- Ser escalable para textos grandes y complejos.

Gracias a este ejercicio comprendí estas ventajas, ya que por ejemplo si se hace simplemente una busqueda directa, esto implicaría revisar todo el texto cada vez que se necesite contar una palabra, por lo que tener un vocabulario representa una mejor opcion cuando se desea realizar un analisis de texto.