Emilio Sanchez 202321349
María Alejandra Villa: 202320315

Entendimiento y Preparación de los datos:
Un paso esencial en el proceso de análisis de texto es la comprensión y preparación de los datos. En este caso, se utilizó la biblioteca Pandas para leer el archivo de datos y realizar las operaciones requeridas. La composición del conjunto de datos y las operaciones realizadas determinan la complejidad de esta etapa. Primero, se determinó el idioma de los textos y se eliminaron los textos en inglés. Esto requiere un paso de procesamiento adicional para identificar el idioma de cada texto, lo que podría alargar el tiempo de preparación y la complejidad de los datos. Posteriormente, se confirmó que el conjunto de datos no contenía valores duplicados ni nulos. Esta verificación es necesaria para garantizar la calidad de los datos y evitar problemas durante el análisis posterior. La cantidad de valores nulos o duplicados presentes y el tamaño del conjunto de datos determinan el tiempo de preparación de los datos en esta etapa. Posteriormente, se llevó a cabo una limpieza de texto, que implicó corregir palabras codificadas incorrectamente, eliminar caracteres especiales y puntuación, tokenizar palabras y normalizar el texto, lo que incluyó la obtención de palabras raíces y lemas. Estas tareas pueden ser bastante laboriosas, sobre todo si se agrupan con un gran número de textos. La cantidad de tiempo dedicada a preparar los datos en esta etapa puede ser significativa, pero es esencial para garantizar la calidad y coherencia de los datos. Al final, la columna "palabras" se redujo a un solo campo de texto y el Data Frame actualizado se almacenó en un archivo CSV. Esta operación es realmente rápida y no requiere mucha complejidad. En cuanto al acceso al software, se utilizó la biblioteca Pandas Python para realizar las operaciones de lectura, limpieza y transformación de datos. Esta biblioteca es ampliamente utilizada y cuenta con documentación completa, lo que facilita su acceso y uso. Según lo analítico de textos, se apropian otras bibliotecas de procesamiento de texto, como NLTK, que también están disponibles y ampliamente utilizadas.

Puntos a Mejorar:
Durante la preparación de los datos se pueden realizar algunas mejoras para reducir la complejidad, el tiempo de preparación, el tiempo de ejecución y mejorar la accesibilidad del software. A continuación se presentan algunas recomendaciones:

1. *Paralelización del procesamiento de datos:* La paralelización del procesamiento es una manera de disminuir el tiempo necesario para la preparación y ejecución de los datos. Esto implica dividir la recopilación de datos en partes más pequeñas y procesarlas simultáneamente en varios núcleos o máquinas. Esto podría acortar significativamente el tiempo necesario para preparar y ejecutar los datos.

2. *Utilización de técnicas de procesamiento de datos en tiempo real:* en lugar de procesar todos los datos a la vez, se pueden utilizar técnicas de procesamiento de datos en tiempo real para procesar los datos a medida que llegan. Esto podría acortar el tiempo de preparación y la complejidad de los datos porque elimina la necesidad de esperar a que todos los datos estén disponibles antes de comenzar el procesamiento.

Emilio Sanchez 202321349
María Alejandra Villa: 202320315

3. *Optimización de algoritmos y operaciones:* Para acortar los tiempos de ejecución, es posible optimizar los algoritmos y operaciones utilizadas en la preparación de datos. Esto puede implicar, entre otras cosas, el uso de algoritmos más efectivos, la optimización de canalizaciones y operaciones vectorizadas.

4. *Uso de herramientas y bibliotecas efectivas:* es fundamental utilizar herramientas y bibliotecas efectivas para el procesamiento de datos. Esto puede implicar el uso de bibliotecas de procesamiento de texto optimizadas para la tarea, como spaCy o fastText, que pueden acortar el tiempo necesario para completar las operaciones de limpieza y transformación de texto.

5. *Adoptar técnicas eficientes de almacenamiento y acceso:* por ejemplo, se pueden utilizar técnicas eficientes de almacenamiento y acceso para mejorar el tiempo de preparación de datos y mejorar el acceso al software. Esto incluye el uso de bases de datos optimizadas para procesar grandes volúmenes de datos, así como el uso de técnicas de indexación para acelerar la búsqueda y recuperación de datos.