Reporte Brandon Villada Vocabulary

Para este trabajo, me enfoqué en extraer y analizar el vocabulario de un archivo en formato PDF. El proceso consistió en varias etapas, desde la conversión del PDF a texto, la limpieza de los datos y la generación de un vocabulario con sus respectivas frecuencias.

Extracción del Texto

Lo primero que hice fue convertir el contenido del PDF en texto utilizando la biblioteca `pdfplumber`. Esta herramienta me permitió leer cada página del documento y extraer su contenido en un solo bloque de texto. Posteriormente, guardé este contenido en un archivo CSV para facilitar su manipulación y almacenamiento.

Limpieza del Texto

Después de extraer el texto, procedí a limpiarlo para eliminar caracteres no deseados y normalizar las palabras. Implementé una función en Python que convirtió todo el texto a minúsculas y eliminó signos de puntuación y caracteres especiales. La limpieza de datos es fundamental para garantizar que el análisis del vocabulario sea preciso y no se vea afectado por variaciones en la escritura de las palabras.

Construcción del Vocabulario

Una vez que tenía el texto limpio, separé las palabras y creé un diccionario donde cada palabra era una clave y su frecuencia de aparición era el valor asociado. Esto me permitió generar un listado ordenado de palabras según su frecuencia de uso. Posteriormente, guardé esta información en un archivo Parquet para facilitar su acceso y análisis en futuras ocasiones.

Dificultades Encontradas

El proceso en general fue fluido, pero encontré algunos desafíos. Uno de los principales fue la presencia de caracteres especiales y formatos de

texto dentro del PDF que podían afectar la extracción. También tuve que asegurarme de que la limpieza de texto no eliminara información relevante. Otro reto fue manejar palabras con acentos o diferentes formas gramaticales que podían afectar el conteo del vocabulario. A pesar de estos detalles, el resultado final me permitió obtener un análisis confiable del contenido del PDF.

En conclusión, este trabajo me ayudó a mejorar mis habilidades en procesamiento de texto y análisis de datos en Python. La combinación de `pdfplumber`, `pandas` y `re` fue clave para lograrlo de manera eficiente.