Reporte del Vocabulary Challenge **14 de marzo de 2025**

Explicación del procedimiento

Para realizar este ejercicio primeramente se descargó el archivo PDF del libro “Los miserables” de Víctor Hugo. Este fue bajado de la página de la fundación “Carlos Slim”. Una vez descargado el archivo se procedió a extraer el texto de el haciendo uso de la librería PyPDF2, con ella se extrajo el contenido del libre y se guardó como un string.

Teniendo la cadena de texto, se procedió a definir una función de limpieza, la cual sirve para limpiar el texto extraído, convirtiendo todo a minúsculas, quitando diacríticos exceptuando la ñ, eliminando números y signos de puntuación, de esta manera logramos un texto plano limpio y que nos permitirá trabajarlo de manera más adecuada.

Una vez limpio el texto, se generó un vocabulario, separando las palabras de la cadena de texto y contando las veces que se repiten cada una. Con este vocabulario generado se guardo en un archivo CSV y también en un archivo parquet, para generar el archivo parquet se usó la librería *pyarrow.*

Mediante la lectura del archivo parquet como un dataframe, se realizó el último punto del ejercicio, el cual consiste en obtener datos estadísticos relevantes del libro. Primeramente, se obtuvo que el total de palabras que contiene es de **109271**, siendo un total de **13125** palabras distintas las que se utilizan. Se obtuvieron las 100 palabras más comunes y las 100 menos comunes con su respectiva frecuencia, finalmente también se obtiene el valor medio de uso de una palabra en el libro que es de **8.32**.

Las 100 palabras más comunes son: "de" con 5325 repeticiones, "la" con 3918, "que" con 3818, "el" con 3393, "y" con 3123, "en" con 2836, "a" con 2488, "se" con 1681, "un" con 1601, "no" con 1498, "los" con 1352, "una" con 1316, "su" con 1245, "las" con 935, "por" con 935, "con" con 924, "habia" con 858, "del" con 813, "al" con 756, "es" con 749 y otros 80.

Las 100 palabras menos comunes son: "inarticulados" con 1 repeticion, "persiguiendome" con 1, "turbaria" con 1, "alboroto" con 1, "murmullos" con 1, "protestas" con 1, "ambiente" con 1, "respirable" con 1, "tranquilos" con 1, "alocado" con 1, "salvarse" con 1, "devolverle" con 1, "robados" con 1,

"proyectos" con 1, "abominables" con 1, "desfallecer" con 1, "importunan" con 1, "indistintamente" con 1, "ataque" con 1, "cedia" con 1 y otros 80.

Comentarios y conclusión

Como comentarios finales, el trabajar este ejercicio resulto bastante enriquecedor en varias áreas, algunos puntos a remarcar en los que me ayudó fueron:

1. Conocer librerías para el manejo de archivos PDF y el extraer el texto ellos gracias a herramientas muy útiles y fáciles de implementar, como lo fue el caso de PyPDF2. Algo sorprendente es el tiempo de ejecución de estas herramientas, pues mas de 300 paginas y mas de cien mil palabras son leídas y convertidas a texto plano en menos de 1 segundo, algo que no esperaba.
2. El uso de expresiones regulares para identificar símbolos y letras, si bien hay toda una biblia atrás de las expresiones regulares, este ejercicio sirvió para refrescar un poco su uso, significado e importancia en el manejo de texto.
3. El manejo de dataframes, un clásico que no puede faltar y siempre es bueno estar practicando, en este ejercicio se presto gratamente para su uso, obteniendo un manejo muy simple y eficiente de los datos. Con su uso fue relativamente sencillo obtener los datos mas relevantes que se buscaban, algo sorprendente que mas de 100 mil palabras, teniendo mas de 13 mil distintas, sea posible identificarlas y contarlas en menos de 1 segundo.

Como conclusión este ejercicio fue de gran ayuda para conocer nuevas herramientas y funcionalidades, mostrando el gran campo de posibilidades para el manejo de información, en este caso, de textos.