Reporte del Proceso de Análisis del Libro “Los Miserables”

---

1. Introducción

Se analizó el libro *Los Miserables* de Victor Hugo, descargado desde el sitio de la Fundación Carlos Slim. El objetivo fue extraer, limpiar y procesar el texto para construir un vocabulario en formato Parquet y generar estadísticas sobre las palabras más y menos frecuentes.

---

2. Metodología

- **Extracción de texto:** Usé `pdfplumber` para extraer el contenido del PDF, filtrando números de página.
- **Limpieza:** Normalicé el texto eliminando mayúsculas, acentos y signos de puntuación con `re` y `unicodedata`.
- **Construcción del vocabulario:** Tokenicé el texto en palabras y conté su frecuencia con `collections.Counter`.
- **Almacenamiento:** Guardé el vocabulario en formato Parquet con `pyarrow`.
- **Estadísticas:** Analicé la cantidad total de palabras, palabras únicas y sus frecuencias.

---

3. Resultados

Métrica	Valor
Total de palabras extraídas	109,627
Palabras únicas en el vocabulario	13,185

Top 3 palabras más frecuentes:

- `de`: 5,328 veces.
- `la`: 3,918 veces.
- `que`: 3,818 veces.

Top 3 palabras menos frecuentes:

- `promiscuidad`: 1 vez.
- `sufrio`: 1 vez.
- `encontrados`: 1 vez.

---

4. Experiencia

El mayor reto fue extraer el texto correctamente, ya que el PDF contenía elementos innecesarios, como números de página. También tuve que normalizar bien las palabras para evitar variaciones ortográficas innecesarias para garantizar datos consistentes y precisos.

El uso de paquetes como `pandas` y `pyarrow` facilitó la manipulación y almacenamiento del vocabulario. El usar el formato Parquet resultó ser una decisión eficiente, tanto en términos de espacio como de velocidad.

---

5. Conclusión

Este análisis permitió obtener estadísticas sobre el vocabulario de *Los Miserables* y experimentar con técnicas de procesamiento de texto. Fue un ejercicio útil para aplicar herramientas de extracción y análisis de datos.