# Reporte: Análisis de Vocabulario en

# *Los Miserables*

## 1. Introducción

El presente reporte describe el proceso y los resultados del análisis de vocabulario realizado sobre la obra *"Los Miserables"* de Victor Hugo. El objetivo principal fue extraer, limpiar y analizar el texto del libro para crear un vocabulario, realizar estadísticas y obtener insights sobre la distribución y frecuencia de las palabras.

## 2. Metodología

El proceso se dividió en las siguientes etapas:

### 2.1. Conversión a CSV

Ya que el libro "**Los Miserables"** proporcionado por la Fundación Carlos Slim no está completo, se optó por descargar el libro completo en formato PDF del siguiente enlace: https://mep.janium.net/janium/Documentos/2866851.pdf.

Se utilizó la biblioteca `pdfminer` para extraer el texto del archivo PDF y a continuación, este se dividió en líneas y se guardó en un archivo CSV para facilitar su procesamiento.

### 2.2. Limpieza de Datos

Se realizaron las siguientes acciones sobre el texto:

- Se normalizó el texto convirtiendo todas las letras a minúsculas.
- Se eliminaron acentos, caracteres especiales y puntuación.
- Se eliminaron espacios adicionales y se estandarizó el formato del texto.
- Se optó además por omitir los valores numéricos. Esto es ideal para NLP cuando solo interesan palabras sin ruido de números o símbolos.

### 2.3. Creación del Vocabulario

- Se tokenizó el texto en palabras individuales.
- Se identificaron las **stopwords** (palabras funcionales como "de", "la", "que") utilizando la lista de **stopwords** en español de la biblioteca nltk.
- Se contó la frecuencia de cada palabra para crear el vocabulario.

### 2.4 Almacenamiento y Análisis

- El vocabulario se guardó en formato Parquet para facilitar su almacenamiento y consulta.

# 3. Resultados

## 3.1. Estadísticas Generales

- **Total de palabras en el libro:** 548,335
- **Palabra únicas en el vocabulario:** 33,787
- **Stopwords encontradas**: 266,572 (48.6% del texto).
- **Palabras únicas que son stopwords:** 202 (0.59% de las palabras únicas)

## 3.2. Frecuencia de palabras

🔝 10 palabras más frecuentes:

| | palabra | frecuencia |
|---|---|---|
| 6 | de | 28867 |
| 17 | la | 22498 |
| 26 | que | 20013 |
| 39 | y | 16458 |
| 33 | el | 16176 |
| 51 | a | 14683 |
| 83 | en | 13779 |
| 160 | se | 9599 |
| 29 | un | 7681 |
| 52 | los | 7484 |

🔻 10 palabras menos frecuentes:

| | palabra | frecuencia |
|---|---|---|
| 19 | marginado | 1 |
| 40 | redimirse | 1 |
| 45 | logra | 1 |
| 59 | obligada | 1 |
| 60 | prostituirse | 1 |
| 61 | subsistir | 1 |
| 67 | reincidido | 1 |
| 99 | negada | 1 |
| 102 | clasicas | 1 |
| 117 | fundamental | 1 |

🔝 10 palabras más frecuentes (sin stopwords):

| | palabra | frecuencia |
|---|---|---|
| 532 | habia | 3659 |
| 14 | mas | 2387 |
| 75 | si | 2005 |
| 550 | marius | 1403 |
| 55 | hombre | 1398 |
| 260 | senor | 1362 |
| 242 | dos | 1300 |
| 1516 | dijo | 1269 |
| 0 | jean | 1211 |
| 1 | valjean | 1139 |

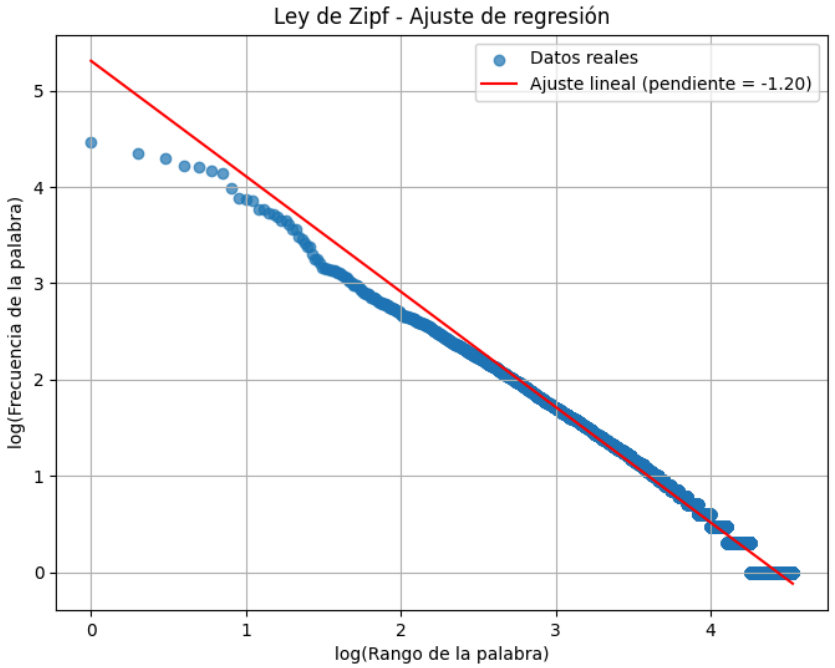
## 3.3 Nube de Palabras

Se generó una nube de palabras para visualizar las 100 palabras más frecuentes sin stopwords. Las palabras más destacadas incluyen "habia", "marius", "hombre", "jean", y "valjean", lo que refleja la importancia de estos términos en la obra.


### 3.4. Análisis de la Ley de Zipf

Esta información permitió verificar si el vocabulario de *Los Miserables* sigue la **Ley de Zipf** según la cual en una determinada lengua la frecuencia de aparición de distintas palabras sigue una distribución tal que la frecuencia de una palabra es inversamente proporcional a su rango en la lista global de palabras, se realizó un análisis en escala log-log entre el rango de las palabras y su frecuencia.

- Se obtuvo una pendiente de **-1.20**, lo que sugiere que la novela sigue una distribución Zipfiana con una ligera desviación. Esto indica que la frecuencia de las palabras más comunes disminuye un poco más rápidamente de lo esperado.




## 4. Conclusiones

- Se logró extraer y analizar el vocabulario de *Los Miserables* exitosamente.
- Se verificó que el texto sigue la **Ley de Zipf**, con una pendiente de **-1.20**.
- La estructura del español y el estilo de Victor Hugo pueden influir en la distribución de palabras.