Análisis de Vocabulario de "Los Miserables"

Joseph Tuffit Hadad Piña

14-03-25

1. Introducción

Este el siguiente paper presentamos los resultados de crear un voabulario de las palabras en la obra "Los Miserables"de Victor Hugo y realizar estadisticas para conocer mas sobre el vocabulario usado.

2. Metodología

El análisis se realizó siguiendo estos pasos:

2.1. Extracción del texto

Se procesó un archivo PDF de 305 páginas. Usando python pudimos extraer el texto completo.

2.2. Limpieza y normalización

Se la siguiente limpieza al texto:

- Conversión a minúsculas.
- Eliminación de acentos.
- Sustitución de signos de puntuación por espacios.
- Eliminación de números.

2.3. Creación del vocabulario

El texto limpio se dividió en palabras individuales, obteniendo un total de 109,502 palabras. Para cada palabra en el texto:

1. Se verificó si ya existía en el vocabulario
2. Si existía, se incrementó su contador de frecuencia
3. Si no existía, se añadió al vocabulario con frecuencia inicial de 1

Obtuvimos un vocabulario de 13,121 palabras únicas con sus respectivas frecuencias de aparición.

3. Resultados del Análisis

3.1. Estadísticas generales

- Total de palabras en el texto: 109,502
- Palabras únicas: 13,121

3.2. Palabras más frecuentes

Las cinco palabras más frecuentes son elementos gramaticales:

Palabra	Frecuencia	% del texto
de	5,328	4.87 %
la	3,918	3.58 %
que	3,818	3.49 %
el	3,394	3.10 %
y	3,123	2.85 %

Entre las palabras de contenido más frecuentes encontramos: "hombre"(367), .ºbispo"(289), "madeleine"(214), "jean"(203), "fantine"(194) y "valjean"(192)

3.3. Palabras menos frecuentes

Las palabras que aparecen una sola vez son numerosas e incluyen términos como "promiscuidad", "holocausto", "florecer", .emocionantesz "fulgor".