Vocabulary challenge

Felipe Andres Castillo
email: 3.1416.p@ciencias.unam.mx

Diplomado Inteligencia Artificial y Ciencia de Datos

14 de marzo de 2025

Resumen

Se implementó un script en Python con el objetivo de construir un vocabulario que registre las diferentes palabras y su frecuencia de aparición en el libro *Los miserables*, de Victor Hugo. En el volumen 1, se identificaron un total de 109,224 palabras, de las cuales 13,240 son únicas.

1. Introducción

1.1. *Les misérables*

Los miserables es una novela del poeta y escritor francés Victor Hugo publicada en 1862, considerada una de las obras más importantes del siglo XIX. El texto, de estilo romántico, plantea por medio de su argumento una discusión sobre el bien y el mal, sobre la ley, la política, la ética, la justicia y la religión. En su núcleo, la novela sirve como una defensa de los oprimidos, sea cual sea el lugar o la situación sociohistórica.

Los miserables ha sido objeto de numerosas adaptaciones teatrales, cinematográficas y televisivas, entre ellas el famoso musical del mismo nombre que ya ha recorrido con gran éxito medio mundo.

1.2. El procesamiento del lenguaje natural

El procesamiento del lenguaje natural (PLN), como rama de la inteligencia artificial, emplea el aprendizaje automático para analizar e interpretar textos y datos, permitiendo extraer información valiosa a partir de datos no estructurados basados en texto.[1]

Este enfoque es fundamental para procesar de manera eficiente y profunda datos provenientes de texto y voz, pudiendo abordar desafíos como diferencias en dialectos, jergas e irregularidades gramaticales comunes en el lenguaje cotidiano. Algunas de sus principales aplicaciones incluyen el análisis de documentos, la evaluación de comentarios de clientes, la implementación de chatbots para atención automatizada, y la clasificación y extracción de contenido, entre otras.

Al igual que en otros modelos de deep learning, la aplicación de técnicas de PLN requiere un preprocesamiento adecuado del texto. Esto implica tareas como la estandarización, tokenización y vectorización de palabras o frases, con el fin de estructurar los datos para su análisis y procesamiento.

2. Metodología

2.1. El texto

El texto fue obtenido de la página web de la Fundación Carlos Slim y corresponde al primer volumen de Los miserables.[2] El libro fue descargado en formato PDF para luego realizar el preprocesamiento del texto en Python. Se omitieron las primeras dos páginas del PDF, las cuales contenían la portada y la información editorial.

El preprocesamiento consistió en segmentar el texto en palabras individuales y eliminar caracteres especiales (como guiones, comas, puntos y signos de interrogación), así como números, manteniendo únicamente las palabras, incluidas aquellas con acentos. Además, todas las palabras fueron convertidas a minúsculas para mantener uniformidad en el análisis. Para extraer el contenido del PDF y segmentar las palabras, se utilizó la biblioteca `PyMuPDF`.

Tras completar el preprocesamiento, se generó un archivo CSV con todas las palabras extraídas del libro.

2.2. El vocabulario

Mediante una iteración sobre todas las palabras obtenidas y utilizando un diccionario, donde cada palabra se almacenó como clave y su frecuencia de aparición como valor, se construyó una versión preliminar del vocabulario.

Para facilitar el análisis estadístico y almacenar el vocabulario en formato Parquet, los datos fue-

ron transformados en un DataFrame. Esta estructura permitió ordenar fácilmente las palabras según su frecuencia de aparición, así como aplicar filtros y extraer subconjuntos de interés.

3. Resultados y análisis

El código y el vocabulario obtenido pueden consultarse a través del notebook. Los resultados principales son los siguientes:

- El libro contiene un total de 109,224 palabras.
- Se identificaron 13,240 palabras únicas.
- La Tabla 1 muestra las diez palabras más frecuentes, las cuales corresponden mayormente a nexos, lo que explica su alta aparición en el texto.
- Para reducir la influencia de este tipo de palabras en el análisis, la Tabla 2 presenta las diez palabras más frecuentes con al menos cuatro letras.
- La Tabla 3 incluye ejemplos de las palabras menos frecuentes. Se encontró que 7,297 palabras aparecen únicamente una vez en todo el texto.

Palabra	Frecuencia
de	5321
la	3917
que	3505
y	3122
el	3081
en	2835
a	2488
se	1632
un	1601
no	1498

Tabla 1: Las 10 palabras más frecuentes.

Palabra	Frecuencia
había	858
señor	447
hombre	363
obispo	286
estaba	273
sobre	269
cuando	265
aquel	247
madeleine	214
tenía	211

Tabla 2: Las 10 palabras (mayores a tres letras) más frecuentes.

Palabra	Frecuencia
bujía	1
singularidad	1
reparó	1
valga	1
holocausto	1
seguidas	1
mintió	1
evadido	1
insista	1
desfallecer	1

Tabla 3: Ejemplo de palabras menos frecuentes.

4. Conclusiones

A partir de la extracción y estandarización de las palabras del libro, se obtuvo la distribución de los términos que lo conforman, lo que permite su uso en análisis y tareas de procesamiento del lenguaje natural.

Los resultados destacan la importancia de excluir las palabras que funcionan como nexos, ya que no aportan información relevante o distintiva. En este caso, se filtraron las palabras con menos de cuatro letras, dado que es más común que los nexos sean términos cortos. No obstante, el vocabulario podría mejorarse aún más si se contara con una lista más detallada de estas palabras para realizar una eliminación más precisa.

Referencias

[1] Google Cloud. *¿Qué es el procesamiento del lenguaje natural?* URL: https://cloud.google.com/learn/what-is-natural-language-processing?hl=es.

[2] Victor Hugo. *Los miserables I*. Fundación Carlos Slim. URL: https://cdn.pruebat.org/recursos/recursos/Los-miserables.pdf.