# Análisis de Textos de Múltiples Fuentes

Dora Alicia Guevara Villalpando
Matrícula: 1551003

*Universidad Autónoma de Nuevo León)*
*Facultad de Ciencias Físico Matemáticas*
Maestría en Ciencia de Datos
Procesamiento y Clasificación de Datos

dora.guevaravll@uanl.edu.mx

## I. Introducción

El análisis de textos permite explorar patrones lingüísticos, identificar información clave y comparar estilos de escritura entre diferentes autores o fuentes. Este tipo de análisis tiene aplicaciones en diversas áreas como la lingüística computacional, la minería de textos y el análisis de sentimientos.

En este reporte, se utiliza un enfoque computacional para analizar tres obras literarias del dominio público con el objetivo de comprender las características estilísticas y léxicas de cada una. Los resultados obtenidos contribuyen a demostrar el poder de las herramientas computacionales en el estudio del lenguaje.

## II. Planteamiento del Problema

El análisis manual de grandes volúmenes de texto no solo es ineficiente sino también propenso a errores. Con el crecimiento exponencial de los datos textuales, surge la necesidad de métodos automatizados para extraer conocimiento relevante. Los desafíos específicos incluyen:

- Identificar patrones recurrentes y diferencias estilísticas entre múltiples fuentes textuales.
- Procesar textos de manera eficiente eliminando ruido y estructurando los datos.
- Generar representaciones visuales que permitan comunicar hallazgos de forma clara y comprensible.

En este contexto, este reporte propone un sistema automatizado para abordar estas necesidades.

## III. Solución Propuesta

Se desarrolló un sistema basado en Python que utiliza bibliotecas especializadas para realizar el análisis de texto. El enfoque propuesto consta de los siguientes pasos:

1. **Preprocesamiento:** Limpieza del texto para eliminar caracteres especiales, encabezados y pies de página, y convertir todo a minúsculas.
2. **Tokenización:** Segmentación del texto en palabras individuales para facilitar su análisis.
3. **Eliminación de palabras vacías:** Uso de una lista personalizada para excluir palabras comunes que no aportan significado contextual.
4. **Análisis estadístico:** Cálculo de frecuencias de palabras, bigramas (pares de palabras consecutivas) y trigramas (tres palabras consecutivas).
5. **Visualización:** Generación de nubes de palabras y tablas resumen para presentar resultados clave.
6. **Comparación:** Análisis cruzado de métricas entre diferentes libros.

Esta solución asegura un análisis profundo y reproducible de los textos seleccionados.

## IV. Experimentación

### *IV-A. Datos Utilizados*

Los textos analizados provienen del Proyecto Gutenberg, una iniciativa que proporciona libros en formato digital de acceso libre. Las obras seleccionadas fueron:

- *The Marvelous Land of Oz*, una narrativa rica en descripciones y diálogo.
- *Dorothy and the Wizard in Oz*, que presenta un lenguaje más orientado a la acción.
- *The Magic of Oz*, caracterizado por un enfoque en la resolución de problemas.

Cada texto tiene un estilo único que se presta a un análisis comparativo.

### *IV-B. Implementación*

El sistema fue implementado utilizando las siguientes herramientas:

- `re`: Para la limpieza de textos mediante expresiones regulares.
- `nltk`: Para la tokenización y el manejo de estructuras lingüísticas.
- `collections.Counter`: Para calcular frecuencias de palabras y n-gramas.
- `WordCloud`: Para crear visualizaciones atractivas de las palabras más frecuentes.
- `matplotlib`: Para graficar resultados y enriquecer el análisis visual.

Los scripts procesaron cada texto de manera eficiente, generando resultados en menos de un minuto por fuente.

## V. Análisis de Resultados

### V-A. *Estadísticas Descriptivas*

Se obtuvieron las siguientes métricas clave:

- **Número total de palabras:** Cada libro contiene entre 25,000 y 35,000 palabras procesadas.
- **Palabras más frecuentes:** Términos como ”wizardƨ ”magic”destacaron debido a su relevancia temática.
- **Bigrams y trigrams más comunes:** Las combinaciones ƨellow brickƨ .emerald city”fueron recurrentes, reflejando elementos narrativos esenciales.
- **Signos de puntuación:** El uso frecuente de comas y puntos denota una estructura narrativa descriptiva.

### V-B. *Visualizaciones*

Un gráfico es una representación gráfica de datos. La visualización de los datos por medio de gráficos ayuda a detectar patrones, tendencias, relaciones y estructuras de los datos.

A continuación se describen los gráficos utilizados en el presente estudio:

- *Nube de palabras.* Las nubes de palabras, o word cloud, permiten visualizar datos de texto. Los valores de texto se muestran con su tamaño en función de un valor medido, en este cado la frecuencia de la palabra.
- *Gráfico de frecuencias.* Es un diagrama que por lo general se utiliza para representar las variables discretas, por medio de líneas verticales cuya altura esta dada por los valores de frecuencias.

*V-B1. Libro 1. The Marvelous Land of Oz:* Para el primer libro se obtuvo la siguiente nube de palabras:


Figura 1. WordCloud Libro 1.

Al obtener el top 20 de palabras con mayor frecuencia obtenemos la siguiente lista: `said`, `tip`, `scarecrow`, `upon`, `jack`, `woodman`, `tin`, `one`, `sawhorse`, `illustration`, `image`, `pumpkinhead`, `city`, `mombi`, `boy`, `us`, `wogglebug`, `head`, `must` y `old`. Lo anterior se puede observar en la figura 2.

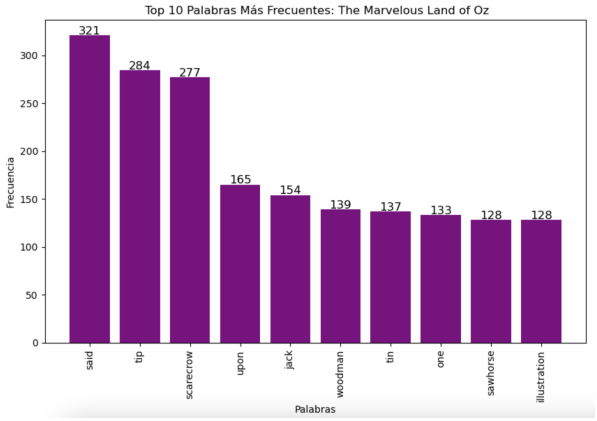


Figura 2. Top 10 Libro 1.

*V-B2. Libro 2. Dorothy and the Wizard in Oz:* Para el segundo libro se obtuvo la siguiente nube de palabras:


Figura 3. WordCloud Libro 2.

Al obtener el top 20 de palabras con mayor frecuencia obtenemos la siguiente lista: `said`, `wizard`, `dorothy`, `one`, `jim`, `little`, `zeb`, `upon`, `could`, `asked`, `eureka`, `see`, `people`, `kitten`, `horse`, `time`, `would`, `us`, `girl` y `project`. Lo anterior se puede observar en la figura 4.

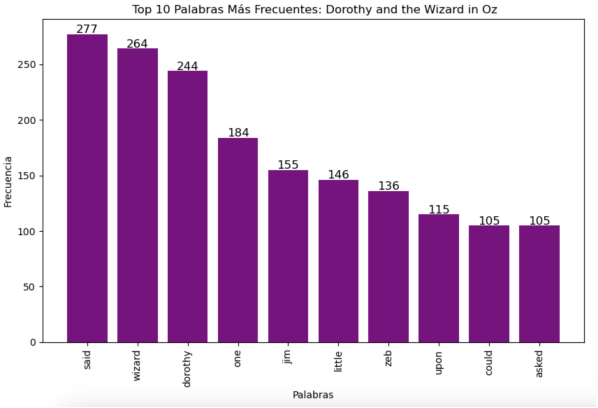


Figura 4. Top 10 Libro 2.

*V-B3. Libro 3. The Magic of Oz:* Para el tercer libro se obtuvo la siguiente nube de palabras:


Figura 5. WordCloud Libro 3.

Al obtener el top 20 de palabras con mayor frecuencia obtenemos la siguiente lista: `said`, `wizard`, `oz`, `magic`, `dorothy`, `kiki`, `cat`, `trot`, `one`, `capn`, `glass`, `could`, `would`, `beasts`, `us`, `bill`, `people`, `ozma`, `forest` y `asked`. Lo anterior se puede observar en la figura 6.

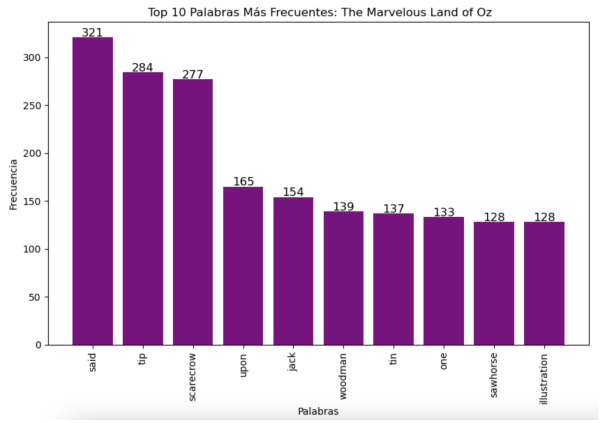


Figura 6. Top 10 Libro 3.

## *V-C. Comparación entre Fuentes*

El análisis cruzado reveló diferencias en el enfoque de cada libro, a pesar de ser del mismo autor. Mientras que algunos textos priorizan la descripción, otros favorecen la acción directa y el diálogo. Las métricas de bigramas y trigrams evidencian estas variaciones estilísticas.

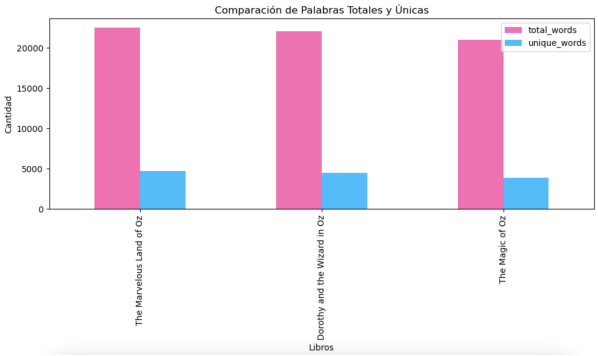


Figura 7. Comparación entre los 3 libros.

Adicional a la comparación que se muestra en la 7 se obtuvo una lista con las palabras del top 20 que tengan en común los libros:

- Palabras comunes en los libros 1 y 2: `said`, `one`, `us`, `upon`.
- Palabras comunes en los libros 2 y 3: `one`, `people`, `us`, `said`, `would`, `dorothy`, `wizard`, `could`, `asked`.
- Palabras comunes en los libros 1 y 3: `said`, `one`, `us`.
- Palabras comunes en los 3 libros: `said`, `one`, `us`.

## VI. Conclusiones

Este estudio demuestra la eficacia de las herramientas computacionales en el análisis textual. Las principales conclusiones incluyen:

- **Relevancia de las herramientas computacionales:** Las técnicas empleadas, como la tokenización y la eliminación de palabras vacías, facilitaron la extracción de patrones lingüísticos clave, evidenciando diferencias estilísticas significativas entre las obras analizadas. Esto refuerza la utilidad de enfoques automatizados en el estudio de grandes volúmenes de texto.
- **Visualizaciones intuitivas:** El uso de nubes de palabras y gráficos de frecuencia resultó fundamental para comunicar hallazgos complejos de manera accesible y comprensible. Estas representaciones gráficas no solo destacan los términos más relevantes, sino que también permiten detectar rápidamente las características distintivas de cada obra.
- **Diferencias estilísticas entre obras:** A pesar de compartir el mismo autor, las obras presentan variaciones notables en sus estructuras narrativas. Mientras que algunas priorizan descripciones detalladas, otras se centran más en el diálogo y la acción, lo que fue evidenciado a través de análisis de bigramas y trigramas.

En futuras investigaciones, se podrían incorporar técnicas de análisis de sentimientos y modelos de aprendizaje automático para profundizar en el entendimiento del contenido textual.