Actividad 1. Análisis de Texto Introducción

Walter Raul Perez Machinena
Facultad de Ciencias de Datos
Universidad Autónoma de Nuevo León
waltermachinena@gmail.com

I. Introducción

Esta actividad presenta un analisis de datos textuales provenientes de libros clásicos del mismo autor Jane Austen. Se realiza un estudio estadístico descriptivo sobre frecuencias, distribuciones de palabras y biagramas, utilizando herramientas como NLTK, pandas y gutenbergpy. Los resultados permiten comparar fuentes textuales y extraer conclusiones significativas sobre el uso del lenguaje. El análisis de texto es una herramienta fundamental para comprender patrones en el uso del lenguaje. Este estudio compara dos libros clásicos utilizando métodos estadísticos y herramientas computacionales.

II. Metodología

Se utilizaron dos libros del mismo autor, utilizando la libreria de gutenbergpy los libros que se utilizaron fueron:

- **Orgullo y Perjuicio**: Autor Jane Austen
- **Emma**: Autor Jane Austen

Para el análisis, se utilizaron las siguientes herramientas y librerías:

- **NLTK**: Tokenización, eliminación de stopwords, y lematización.
- **pandas**: Manipulación y análisis de datos.
- **Matplotlib**: Visualización gráfica.

El flujo de trabajo incluyó:

1) Descarga y limpieza de los textos utilizando *GutenbergPy*.
2) Generacion de metodos para obtener libros, obtener ngramas
3) Utilizacion de librerias para utilizar algoritmos como lenmatizar utilizando WordNetLemmatizer y PorterStemmer para los stenmer.
4) Tokenización y remoción de stopwords.
5) Cálculo de frecuencias y biagramas.
6) Visualizacion de las frecuencias, bigramas, nube de palabras

A. Cantidad de Palabras por Libro

El análisis comenzó con una revisión básica de la extensión de los textos, medida por el número de palabras en cada libro. El Libro 1 contiene un total de 55,253 palabras, mientras que el Libro 2 tiene 67,209 palabras. Esta diferencia de longitud representa un incremento del 21.6% en el Libro 2, lo cual podría influir en la mayor diversidad de biagramas observada en este texto.

B. Signos y simbolos mas utilizados por el autor

Se realizo una revision para identificar que tipo de simbolos utilizaba mas el autor en ambos libros esto para identificar si repite patrones aun cuando son titulos diferentes.

III. Resultados Visuales

A continuación, se presentan los resultados visuales obtenidos a partir del análisis.

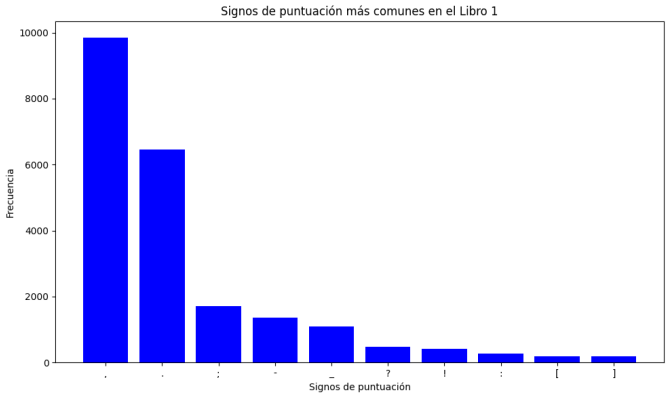

Fig. 1. Signos de puntuacion mas comunes en el Libro 1

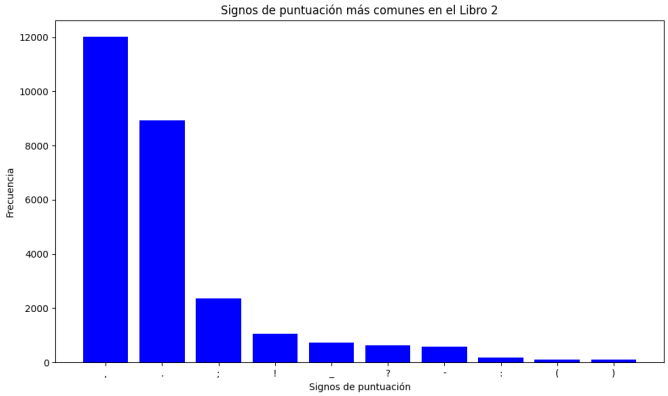

Fig. 2. Signos de puntuacion mas comunes en el Libro 2

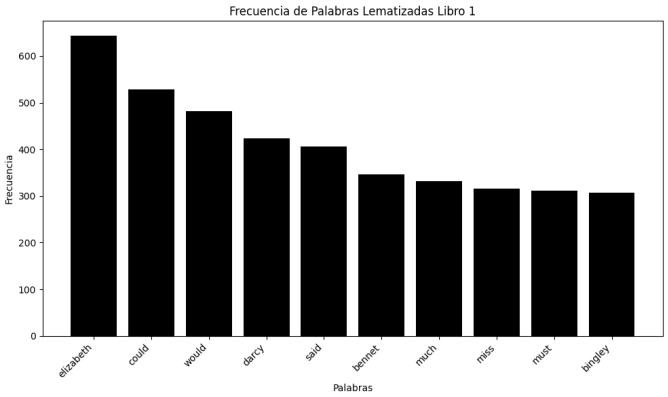

Fig. 3. Frecuencias de palabras Lematizadas en el Libro 1.

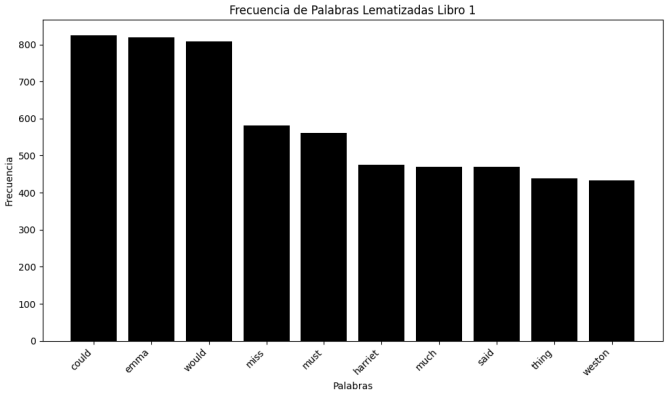

Fig. 4. Frecuencias de palabras Lematizadas en el Libro 2.

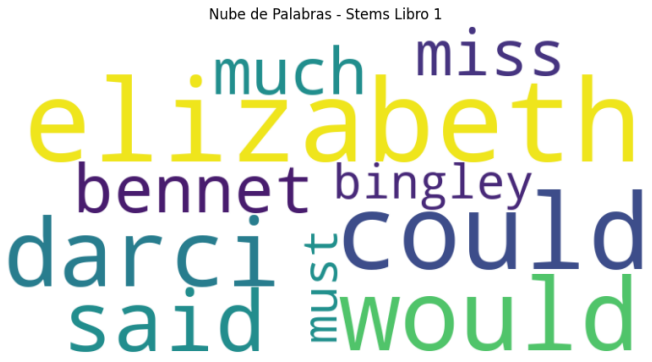

Fig. 5. Nube de palabras para Steams Libro 1

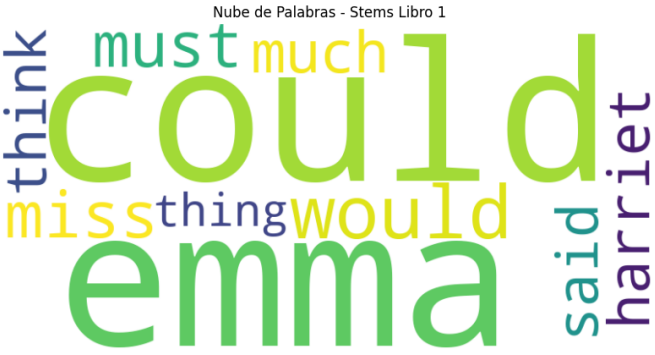

Fig. 6. Nube de palabras para Steams Libro 2

IV. Resultados

Se analizaron las palabras y biagramas más frecuentes en los libros seleccionados. A continuación, se muestran ejemplos:

A. Palabras más frecuentes

En el Libro 1, la palabra más frecuente fue *Elizabeth*, mientras que en el Libro 2 fue *Emma*. Si consideramos la frecuencia entre ambos libros la palabra que mas tuvieron en comun fue *said* y esto lo podemos ver repesentado tanto en las Lematizadas como en las Stenm.

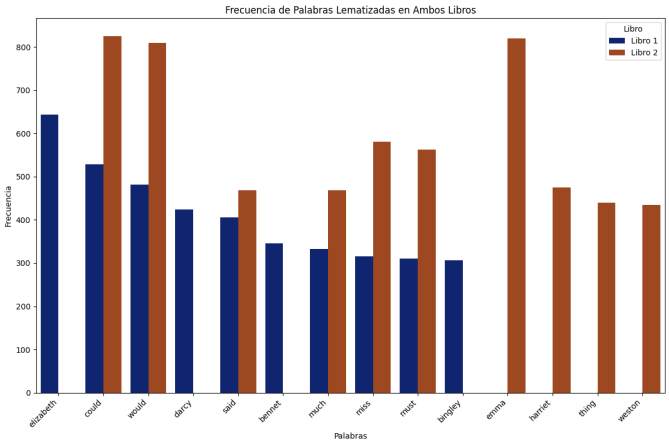

Fig. 7. Frecuencias de palabras Lematizadas en el Libro 1 vs Libro 2.

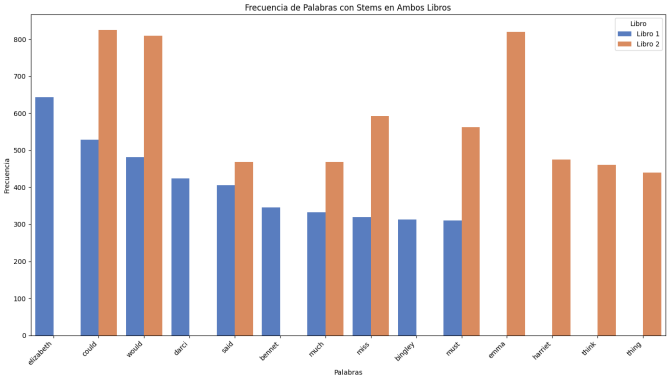

Fig. 8. Frecuencias de palabras Stenm en el Libro 1 vs Libro 2.

B. Bigramas más comunes

En ambos libros, el biagrama *young man* apareció con mayor frecuencia, aunque con distintas proporciones. Mientras que todas las otras 10 mas comunes no se repiten en el otro libro.

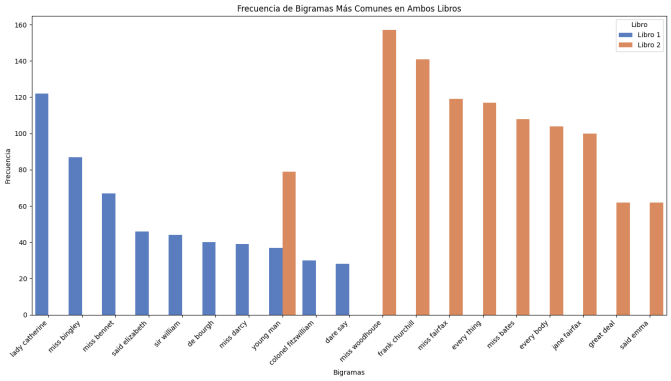

Fig. 9. Frecuencias de palabras bigramas en el Libro 1 vs Libro 2.

C. Signos de puntuación

Con ayuda de esta gráfica, podemos visualizar los 10 símbolos más utilizados en ambos libros. En el Libro 1 destacan símbolos como `-`, `_`, `:` `[ ]`, mientras que en el Libro 2 predominan los signos más comunes, como la coma (`,`), el punto (`.`), el signo de interrogación (`?`) y el signo de exclamación (`!`).

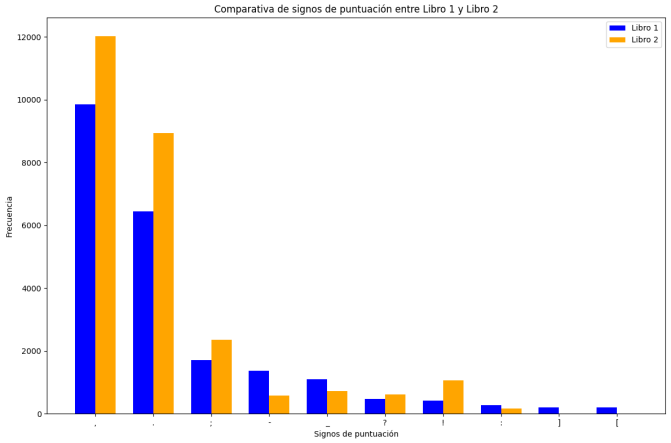

Fig. 10. Comparativa de signos de puntuacion entre Libro 1 y 2

V. Conclusión

El análisis destaca diferencias importantes entre los libros seleccionados, incluyendo la longitud de los textos, que podría ser un factor influyente en la diversidad de palabras y biagramas observados. Esto refuerza la necesidad de normalizar ciertas métricas al comparar textos de diferentes tamaños, considerando que tambien el autor tiene forma de describir similares aun cuando son libros diferentes, ya que podemos identificar que el nombre del personaje principal es el mas utilizado en ambos libros Elizabeth y Enma refiriendonos al libro 1 y 2 respectivamente. De la misma manera se identifica que en el caso de las palabras que son bigramas no sucede esto, ya que ninguna se repite a excepcion de young man. En el caso de los simbolos no tiene algo muy destacable ya que repite 8 de los 10 simbolos en ambos libros, sin embargo en el libro 1 se identifican los simbolos a la derecha, es decir los ultimos 2 el cual es utilizado para aclarar o dar una explicacion adicional.

Agradecimientos

Agradezco a mis compañeros y profesores por el aprendizaje compartido.

References

[1] NLTK Documentation. Disponible en: `https://www.nltk.org/`
[2] Pandas Documentation. Disponible en: `https://pandas.pydata.org/`
[3] Proyecto Gutenberg. Disponible en: `https://www.gutenberg.org/`
[4] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009.
[5] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
[6] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Pearson Education, 2019.
[7] J. Silge and D. Robinson, *Text Mining with R*. [Online]. Available: https://www.tidytextmining.com/