Tarea 2. Análisis Textual de Libros de Project Gutenberg

Autor: *Iván Gabriel Salinas Castillo*

23 de enero de 2025

Resumen

Este reporte presenta un análisis textual obtenido de Project Gutenberg, enfocado en aspectos como la frecuencia de palabras, n-gramas, análisis de sentimientos, y uso de signos de puntuación. También se implementó un modelo avanzado basado en BERT para la vectorización y análisis de características semánticas. El objetivo es explorar la estructura y características lingüísticas del texto a través de técnicas computacionales y estadísticas, así como comparar enfoques.

1. Introducción

El análisis de texto es una herramienta poderosa para extraer patrones, tendencias y características relevantes de los datos textuales. En este proyecto, se seleccionó el repositorio de Project Gutenberg, y se aplicaron tanto técnicas tradicionales de procesamiento de lenguaje natural (NLP) como un enfoque basado en modelos avanzados como BERT (Bidirectional Encoder Representations from Transformers). El objetivo principal es ilustrar cómo diferentes métodos revelan aspectos lingüísticos y semánticos de los textos literarios.

Este reporte se centra en dos obras clásicas: *Don Quijote de la Mancha* y *La Odisea*, y compara sus patrones lingüísticos, estructuras narrativas y polaridad emocional.

2. Metodología

El análisis se llevó a cabo en varias etapas, integrando técnicas tradicionales y modernas:

2.1. Obtención del texto

Los textos fueron descargados del repositorio de Project Gutenberg utilizando la librería `gutenbergpy`. Se eliminó el encabezado y pie de página específicos del formato Gutenberg para limpiar el contenido [2].

2.2. Vectorización con BERT

Se utilizó el modelo BERT preentrenado, disponible a través de la librería `Hugging Face Transformers`, para generar representaciones vectoriales profundas de las oraciones en ambos textos. BERT es un modelo basado en transformadores bidireccionales que captura el contexto semántico de las palabras en un texto [6] [3].

2.3. Preprocesamiento

El texto fue preprocesado para:

- Convertir todas las palabras a minúsculas.
- Eliminar caracteres no alfabéticos.
- Remover *stopwords* utilizando la librería `nltk` [1].
- Aplicar la técnica de lematización.

2.4. Análisis tradicional

En paralelo, se aplicaron técnicas tradicionales como la extracción de n-gramas, análisis de sentimientos con `TextBlob`, análisis de signos de puntuación y estadísticas descriptivas para caracterizar las obras desde un enfoque más clásico [4] [5].

2.5. Comparaciones

Finalmente, se compararon los resultados obtenidos con BERT frente en el análisis de sentimientos general de las obras y se compararon con lo que opinan los lectores.

3. Resultados

3.1. Don Quijote de la Mancha

3.1.1. Estadísticas descriptivas

- Total de palabras: **444,430**
- Total de oraciones: **9,513**
- Longitud promedio de palabras: **3.64**
- Longitud promedio de oraciones: **220.44**

Las oraciones se identifican utilizando la función $nltk.sent_tokenize(texto)$ de la librería NLTK (Natural Language Toolkit). Esta función toma en cuenta los puntos finales (.), signos de exclamación (!), signos de interrogación (?), y otros delimitadores de oración para dividir el texto en oraciones individuales.

3.1.2. Palabras más frecuentes

Las palabras más comunes en el texto (después de eliminar *stopwords*) 1 fueron:

- **Señor**: 1,064
- **ser**: 1,064
- **bien**: 1,042

3.1.3. Trigramas más frecuentes

Los trigramas (fig 2) más comunes fueron:

- **'caballero', 'triste', 'figura'**: 36
- **'vuesa', 'merced', 'señor'**:29
- **'bachiller', 'sansón', 'carrasco'**: 25

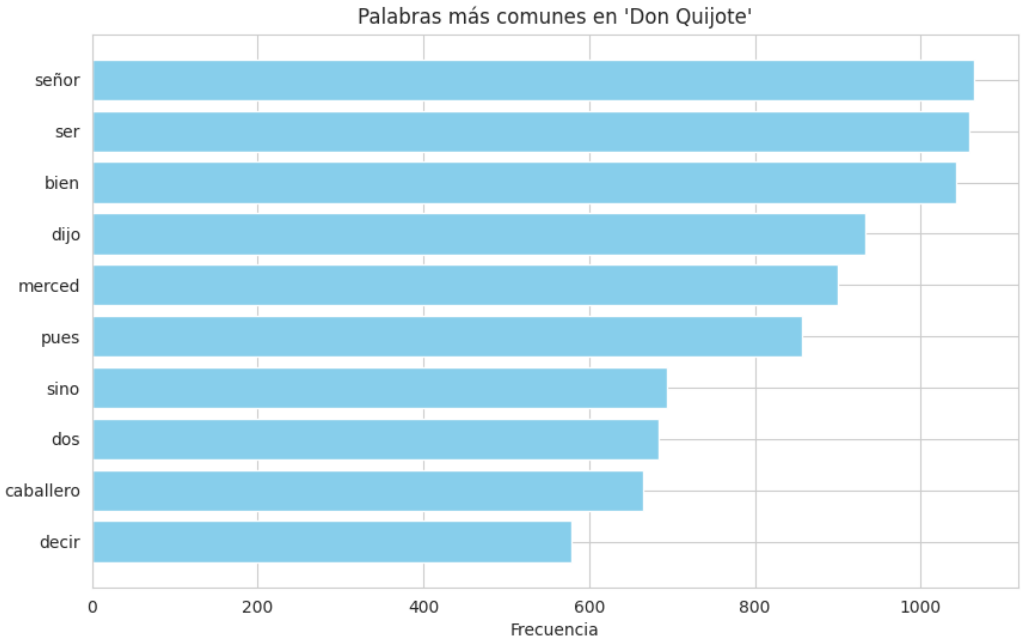

Figura 1: Lematización de Don Quijote.

3.1.4. Bigramas más frecuentes

Los bigramas (fig 3) más comunes fueron:

- **'vuesa', 'merced'**: 198
- **'caballeros', 'andantes'**: 129
- **'caballero', 'andante'**: 116

3.1.5. Análisis de sentimientos

Análisis de sentimientos para personajes de 'Don Quijote':

- Don Quijote: Positivo
- Sancho: Positivo
- Dulcinea: Positivo

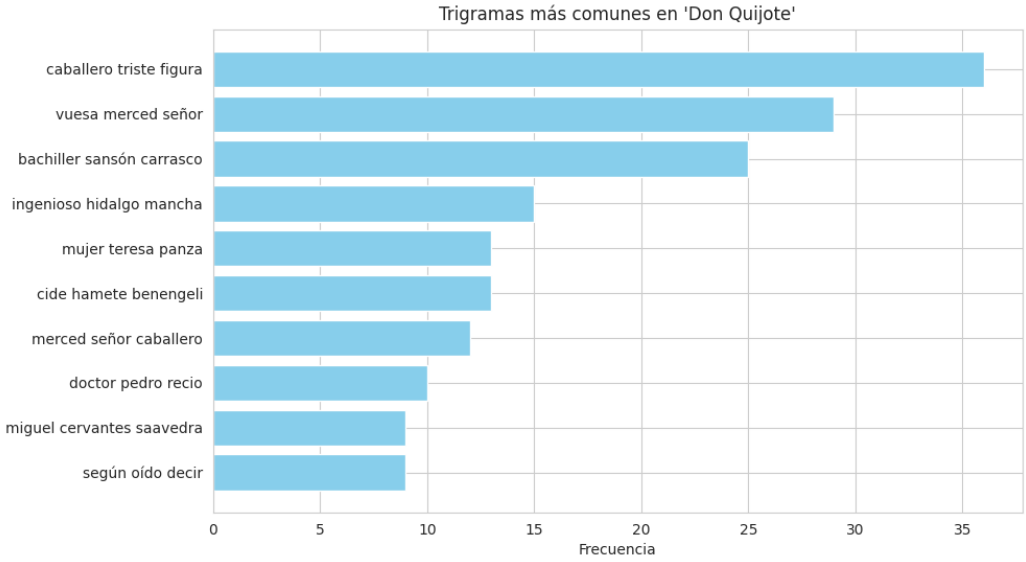

Figura 2: Trigramas de Don Quijote.

3.1.6. Análisis de signos de puntuación

La frecuencia de los signos de puntuación (fig 4) fue:

- Comas: **40,277**
- Puntos: **8,212**
- Puntos y comas: **4,802**

3.2. La Odisea

3.2.1. Estadísticas descriptivas

- Total de palabras: **196,950**
- Total de oraciones: **4,634**
- Longitud promedio de palabras: **3.90**
- Longitud promedio de oraciones: **210.16**

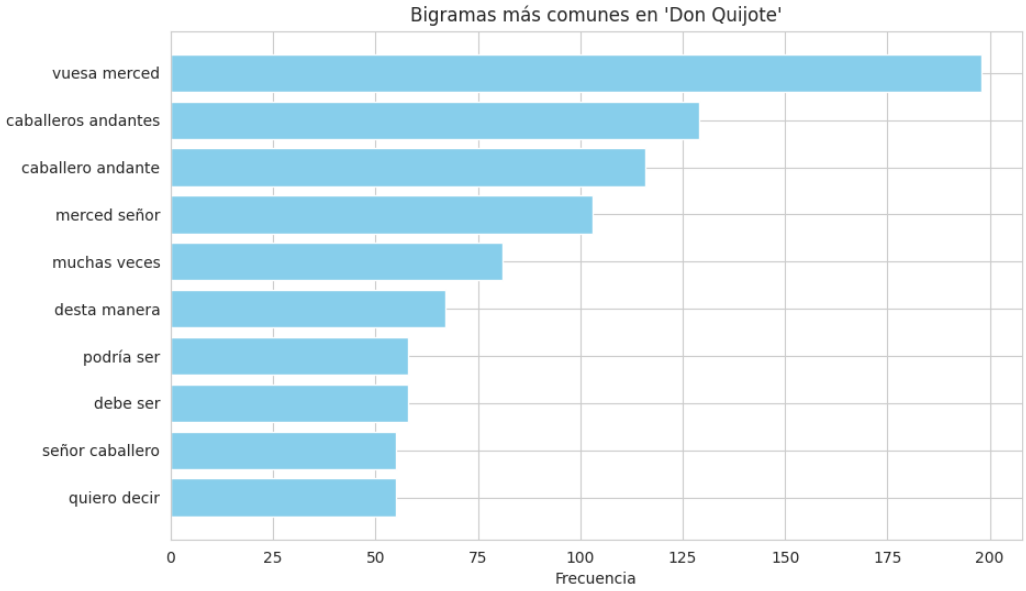

Figura 3: Bigramas de Don Quijote.

3.2.2. Palabras más frecuentes

Las palabras más comunes en el texto (después de eliminar *stopwords*) (fig 5) fueron:

- **pues**: 7,038
- **pretendientes**: 501
- **júpiter**: 481

3.2.3. Trigramas más frecuentes

Los trigramas (fig 6) más comunes fueron:

- ('deidad', 'brillantes', 'ojos'): 32
- ('minerva', 'deidad', 'brillantes'): 30
- ('aurora', 'rosceos', 'dedos'): 22

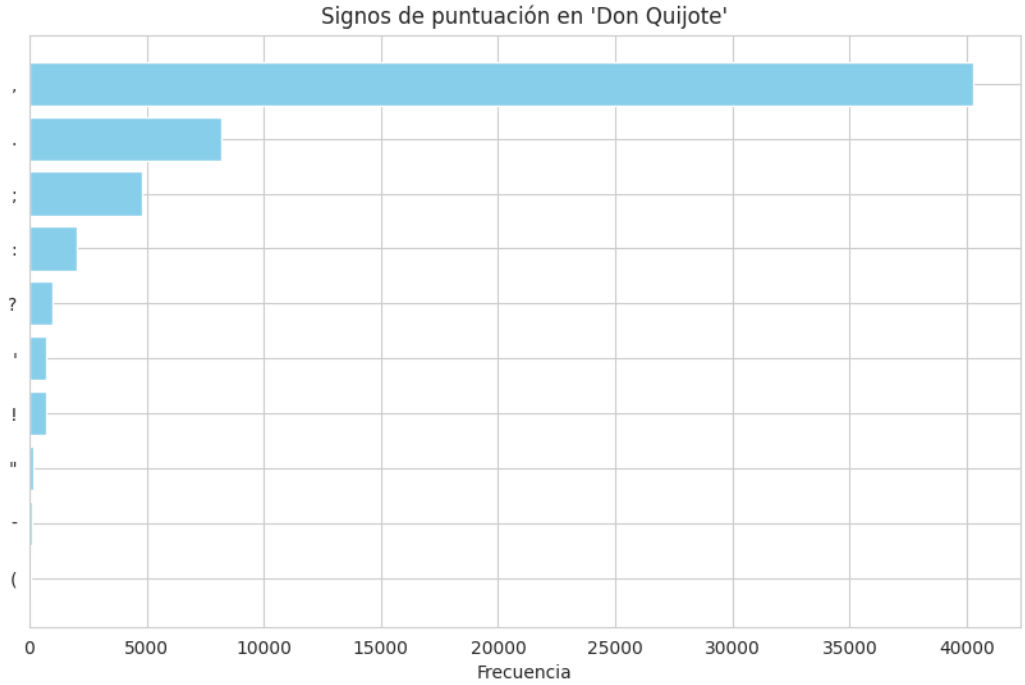

Figura 4: Signos de puntuación de Don Quijote

3.2.4. Bigramas más frecuentes

Los bigramas (fig 7) más comunes fueron:

- ('aladas', 'palabras'): 63 veces
- ('brillantes', 'ojos'): 57 veces
- ('mas', 'ea'): 51 veces

3.2.5. Análisis de sentimientos

Análisis de sentimientos para personajes de 'La Odisea':

- Ulises: Negativo
- Penélope: Negativo
- Telémaco: Negativo

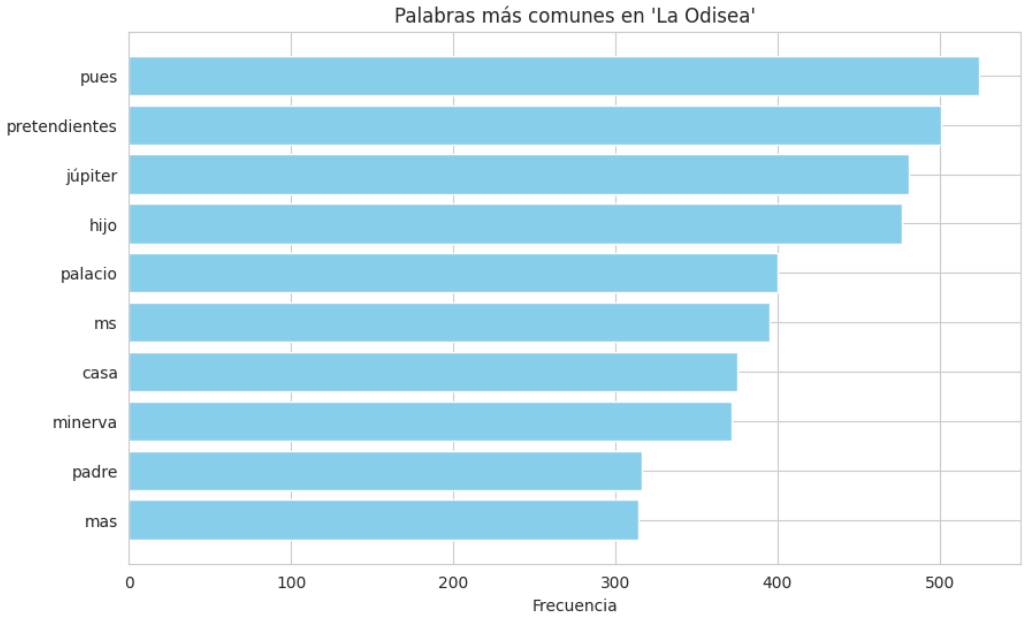

Figura 5: Lematización de La Odisea.

3.2.6. Análisis de signos de puntuación

La frecuencia de los signos de puntuación (fig 8) fue:

- Comas: **17,454**
- Puntos: **4,749**
- Puntos y comas: **3,410**

3.3. Comparación de sentimientos y polaridad en 'Don Quijote' y 'La Odisea'

3.3.1. Análisis de vectorización con BERT

El análisis de vectorización de ambos libros utilizando BERT mostró que el texto se convirtió en un vector con una dimensión de 384 componentes. Esta técnica permite capturar información contextual rica del texto, facilitando aplicaciones posteriores en el Procesamiento de Lenguaje Natural (PLN).

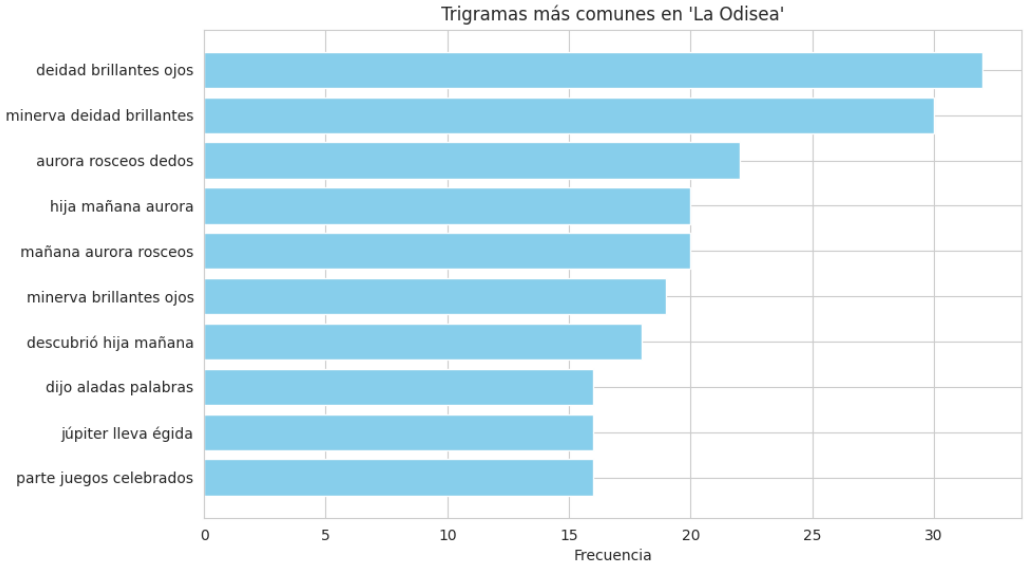

Figura 6: Trigramas de La Odisea.

3.3.2. Polaridad promedio

Un valor de polaridad cercano a 0 indica que el sentimiento general en el texto es neutro, sin tendencias marcadas hacia lo positivo o negativo.

Un valor de polaridad negativo (-0.344) sugiere que el sentimiento general en el texto tiende hacia lo negativo, indicando la presencia de emociones más sombrías o tristes.

- Polaridad promedio en 'Don Quijote': 0
- Polaridad promedio en 'La Odisea': -0.34
- **Don Quijote**: aventura, caballerosidad, confusión, desilusión, comedia, ironía
- **La Odisea**: anhelo, nostalgia, coraje, determinación, pérdida, tristeza

Los resultados del análisis de polaridad promedio son consistentes con las emociones atribuidas por los lectores. En "Don Quijote", la polaridad neutra refleja una mezcla de humor y desilusión, mientras que en "La Odisea", la polaridad negativa está en línea con los sentimientos de pérdida y anhelo presentes en la épica historia de Ulises.

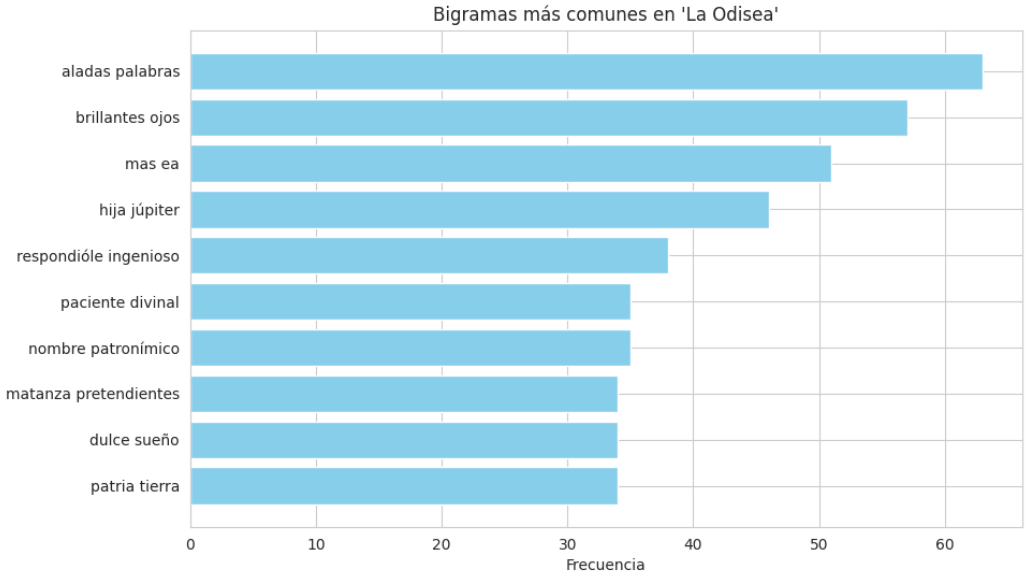

Figura 7: Bigramas de La Odisea.

4. Conclusión

En este análisis textual de Don Quijote de la Mancha y La Odisea, se utilizaron diversas técnicas de procesamiento de lenguaje natural (NLP), que incluyen estadísticas descriptivas, extracción de n-gramas, análisis de sentimientos y vectorización mediante el modelo BERT. El estudio mostró que, a pesar de las diferencias en el contenido de los textos, ambas obras comparten patrones lingüísticos interesantes, como la alta frecuencia de ciertas palabras clave que reflejan sus temas centrales.

Los resultados del análisis de sentimientos indicaron que Don Quijote presenta una polaridad mayoritariamente neutra, reflejando su mezcla de comedia e ironía, mientras que La Odisea mostró una polaridad negativa, lo cual es consistente con sus temas de pérdida y sufrimiento. Estos hallazgos coinciden con las emociones atribuidas por los lectores a ambas obras.

La utilización de herramientas modernas, como BERT para la vectorización, permitió una comprensión más profunda del contexto semántico de los textos, revelando aspectos complejos que no serían evidentes mediante métodos tradicionales. Esta combinación de técnicas clásicas y avanzadas proporciona una visión más rica y matizada de las obras literarias, permitiendo una exploración profunda tanto de sus características lingüísticas como de sus elementos emocionales.

En resumen, el análisis comparativo entre estas dos obras clásicas resalta

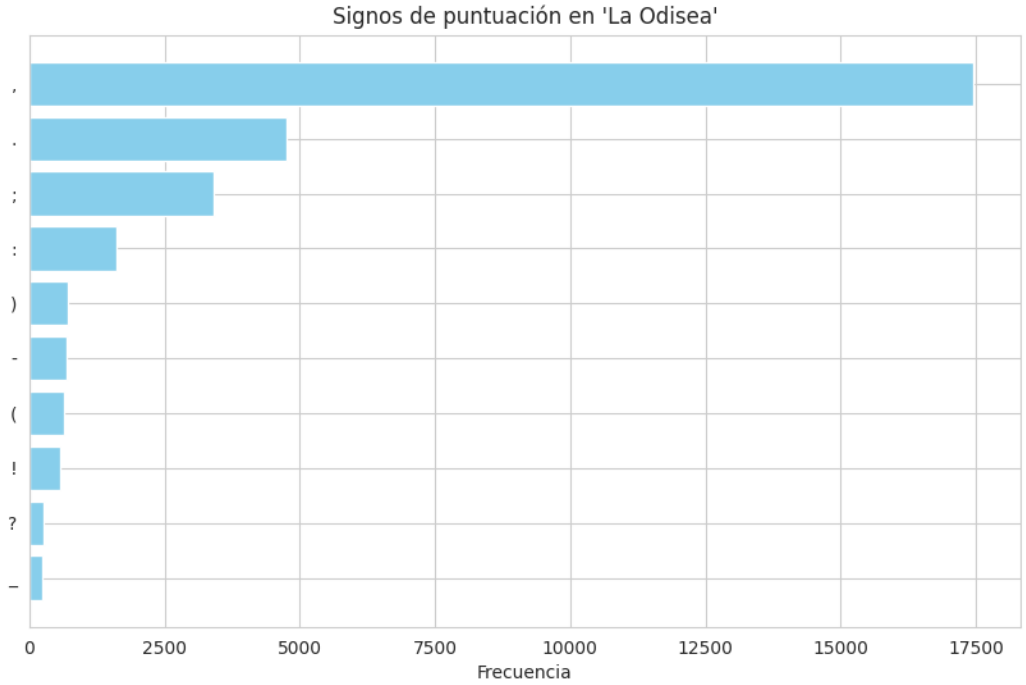

Figura 8: Signos de puntuación de La Odisea.

cómo las técnicas computacionales pueden enriquecer nuestra comprensión de textos literarios y cómo los enfoques tradicionales y modernos pueden complementarse para ofrecer una visión más completa de las obras estudiadas.

Referencias

[1] Steven Bird, Edward Loper, and Ewan Klein. Natural language toolkit (nltk). *NLTK Documentation*, 2009.

[2] Project Gutenberg Contributors. *Gutenbergpy: Access to Project Gutenberg eBooks via Python.* 2023.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Steven Loria. Textblob: Simplified text processing. *TextBlob Documentation*, 2018.

[5] Michael L. Waskom. Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[6] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art machine learning for pytorch, tensorflow, and jax. *arXiv preprint arXiv:1910.03771*, 2019.