Tarea 1. Análisis Textual de Libros de Project Gutenberg

Autor: *Iván Gabriel Salinas Castillo*

January 16, 2025

Abstract

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. El objetivo es explorar la estructura y características lingüísticas del texto a través de técnicas computacionales y estadísticas.

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 técnicas de procesamiento de lenguaje natural (NLP) para obtener estadísticas descriptivas, identificar las palabras más frecuentes, analizar n-gramas, evaluar sentimientos y estudiar el uso de signos de puntuación.

El propósito de este análisis es ilustrar cómo los métodos computacionales pueden revelar información útil sobre la estructura y el contenido de textos literarios.

2 Metodología

El análisis se llevó a cabo en varias etapas:

2.1 Obtención del texto

El texto fue descargado 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 Preprocesamiento

El texto fue preprocesado para:

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

2.3 Análisis descriptivo

Se calcularon las siguientes estadísticas del texto:

- Número total de palabras y oraciones.
- Longitud promedio de palabras y oraciones.
- Frecuencia de palabras más comunes.

2.4 Análisis de n-gramas

Se generaron n-gramas (secuencias de n palabras consecutivas) para identificar patrones lingüísticos. En este caso, se analizaron trigramas (n = 3) y bigramas (n=2).

2.5 Análisis de sentimientos

Utilizando la librería `TextBlob`, se evaluó la polaridad general del texto, donde:

- Un valor positivo indica un sentimiento positivo.
- Un valor negativo indica un sentimiento negativo.
- Un valor cercano a cero indica neutralidad.

2.6 Análisis de signos de puntuación

Se contabilizó el uso de los principales signos de puntuación para evaluar su frecuencia y estilo narrativo.

3 Resultados

3.1 Don Quijote de la mancha

3.1.1 Estadísticas descriptivas

- Total de palabras: **444430**
- Total de oraciones: **9513**
- Longitud promedio de palabras: **3.8667866705667935**
- Longitud promedio de oraciones: **220.44150110375276**

3.1.2 Palabras más frecuentes

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

- **Don**: 2714
- **Si**: 1959
- **Quijote**: 1719

3.1.3 Trigramas más frecuentes

Los trigramas 2 más comunes fueron:

- **'don', 'quijote', 'mancha'**: 144
- **'señor', 'don', 'quijote'**: 142
- **'don', 'quijote', 'sancho'**: 76

3.1.4 Bigramas más frecuentes

Los bigramas 3 más comunes fueron:

- **'don', 'quijote'**: 1710
- **'sancho', 'panza'**: 277
- **'vuesa', 'merced'**: 198

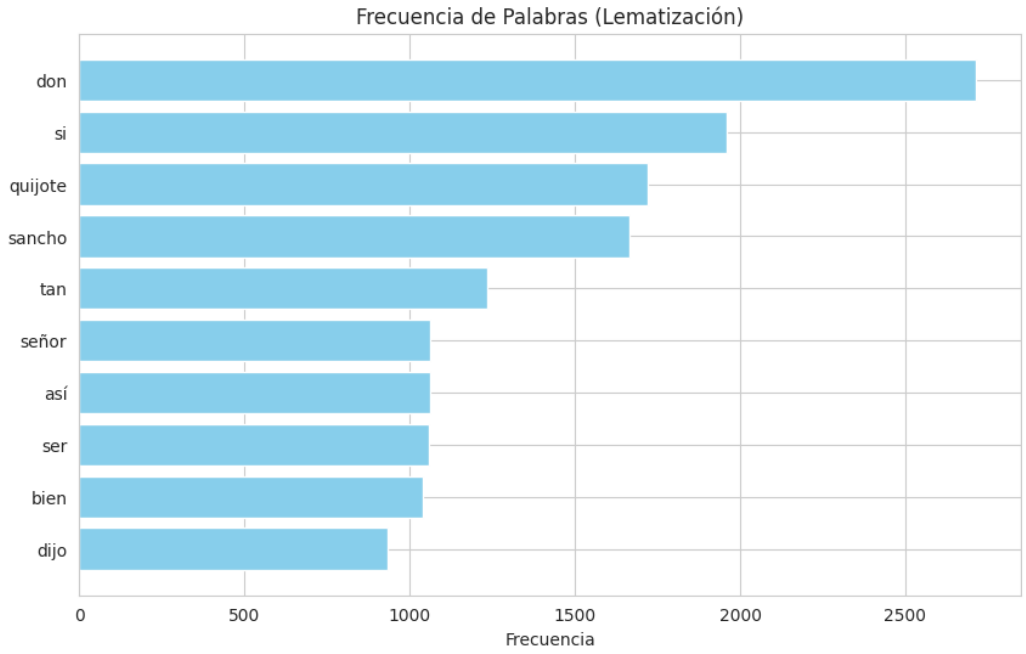

Figure 1: Lematización de Don Quijote.

3.1.5 Análisis de sentimientos

El sentimiento general del texto fue:

- **Polaridad**: 0.00875050004706322 (Positivo)

3.1.6 Análisis de signos de puntuación

La frecuencia de los signos de puntuación 4 fue:

- Comas: **40277**
- Puntos: **8212**
- Puntos y comas: **4802**
- Signos de interrogación: **960**

3.2 La Odisea

3.2.1 Estadísticas descriptivas

- Total de palabras: **203915**

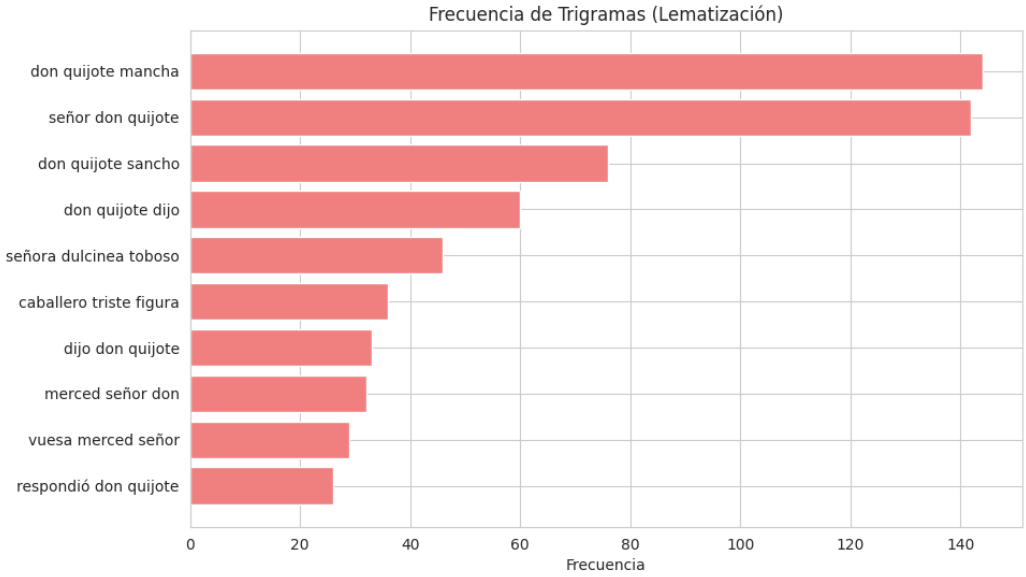

Figure 2: Trigramas de Don Quijote.

- Total de oraciones: **4637**
- Longitud promedio de palabras: **3.904744623985484**
- Longitud promedio de oraciones: **210.16670260944576**

3.2.2 Palabras más frecuentes

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

- **á**: 7038
- **ulises**: 1688
- **telémaco**: 716

3.2.3 Trigramas más frecuentes

Los trigramas 6 más comunes fueron:

- **'respondióle', 'ingenioso', 'ulises'**: 38
- **'paciente', 'divinal', 'ulises'**: 35
- **'deidad', 'brillantes', 'ojos'**: 32

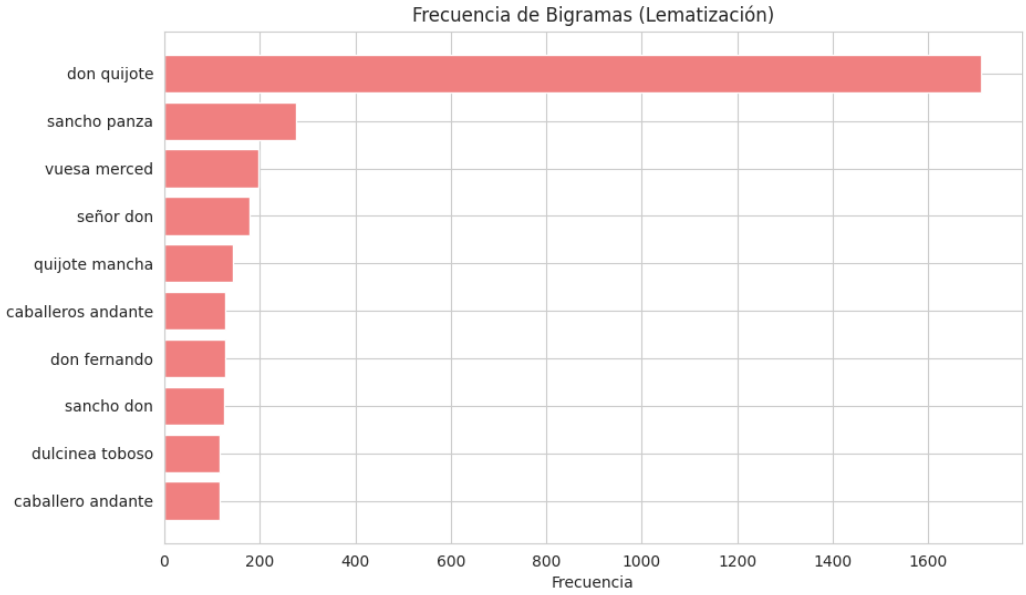

Figure 3: Bigramas de Don Quijote.

3.2.4 Bigramas más frecuentes

Los bigramas 7 más comunes fueron:

- **'á', 'ulises'**: 372
- **'á', 'telémaco'**: 240
- **'á', 'pretendientes**: 129

3.2.5 Análisis de sentimientos

El sentimiento general del texto fue:

- **Polaridad**: -0.34406647013667396 (Negativo)

3.2.6 Análisis de signos de puntuación

La frecuencia de los signos de puntuación 8 fue:

- Comas: **17454**
- Puntos: **4749**
- Puntos y comas: **3410**

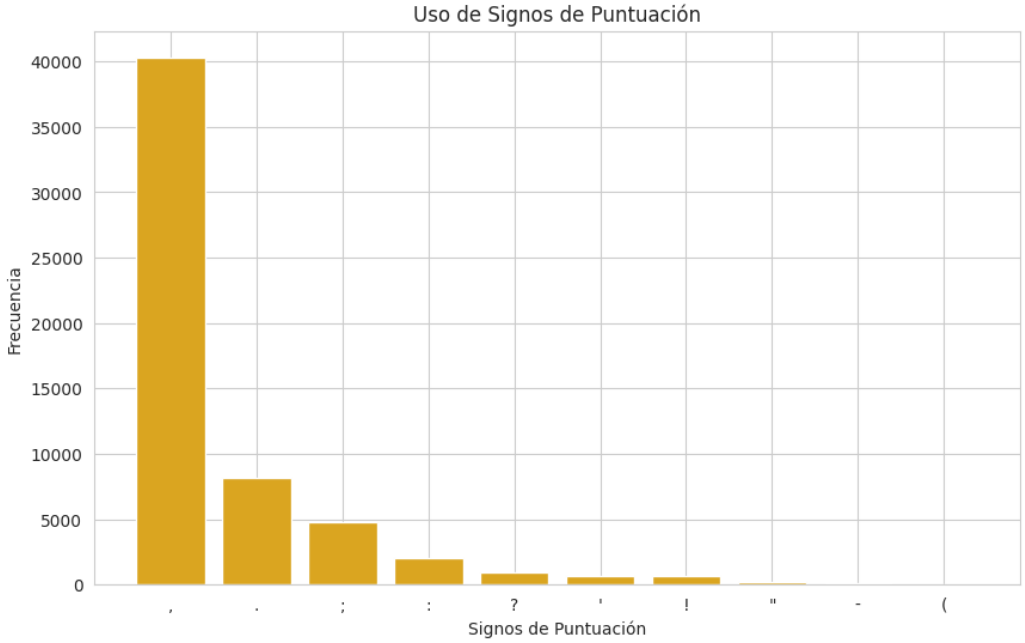

Figure 4: Signos de puntuación de Don Quijote

4 Conclusión

Si comparamos las dos obras de los autores, podemos decir que en Don Quijote se tiende a ser más neutral/positivo, y por otro lado, en La Odisea se tiene una escritura más negativa por parte del autor. También podemos decir, en base a los bigramas, trigramas y la frecuencia de las palabras, que en Don Quijote se centra mucho en los protagonistas debido a que se tiene una alta frecuencia en sus nombres. Por otro lado, en La Odisea el nombre de Ulises y otras palabras están más igualadas en su frecuencia de uso. Por último, también el uso de los signos de puntuación nos ayuda a inferir que en Don Quijote se suelen tener más interrogantes y, posiblemente, también los personajes sean más efusivos al momento de interactuar, por el uso de signos de interrogación y exclamación. Por otro lado, en La Odisea se usan muy poco.

Este análisis demostró cómo las técnicas de procesamiento de lenguaje natural permiten extraer información valiosa de textos literarios. Las estadísticas descriptivas y los patrones lingüísticos revelaron aspectos estilísticos y temáticos del autor. Además, el análisis de sentimientos proporcionó una visión general de la tonalidad del texto, mientras que la evaluación de signos de puntuación destacó el estilo narrativo.

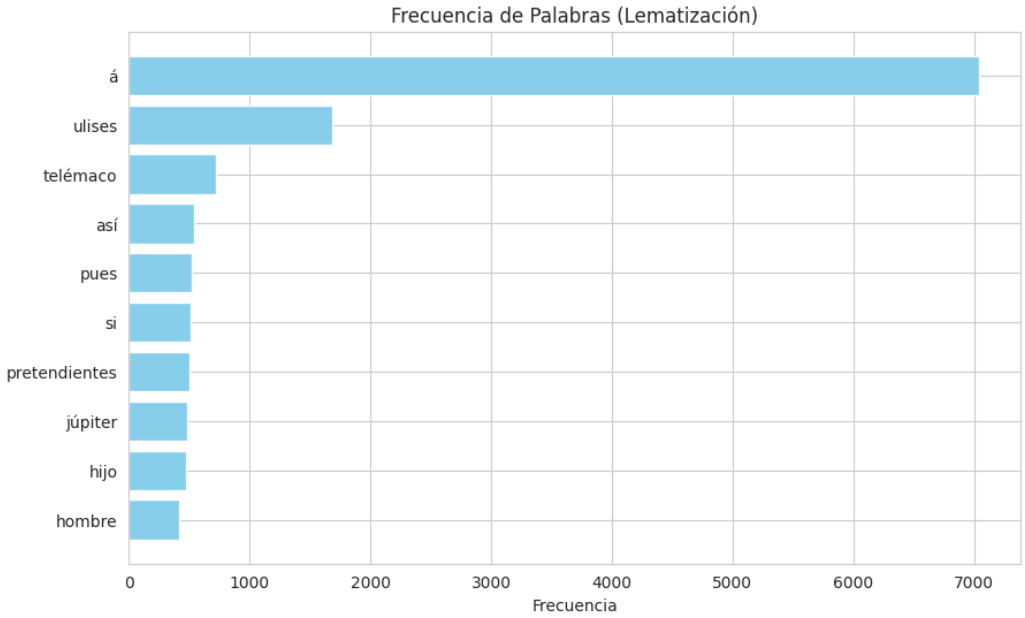

Figure 5: Lematización de La Odisea.

Futuras investigaciones pueden extender este análisis a múltiples textos para identificar tendencias entre diferentes autores o géneros literarios.

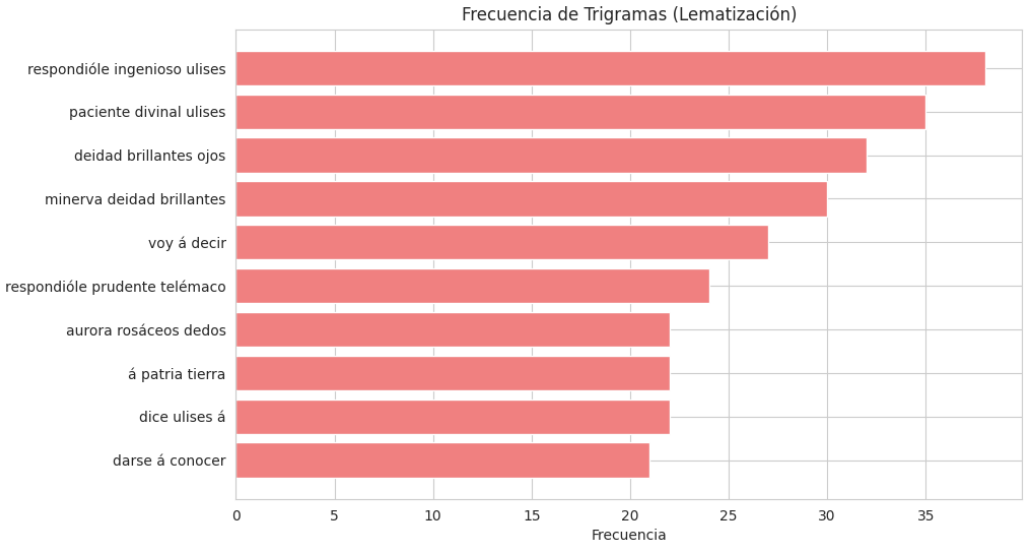

Figure 6: Trigramas de La Odisea.

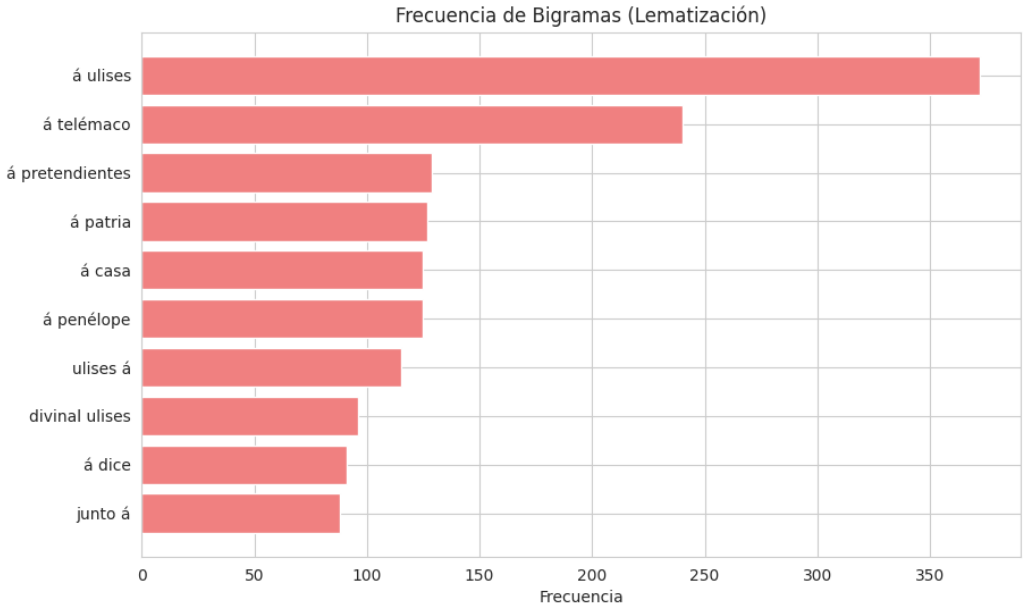

Figure 7: Bigramas de La Odisea.

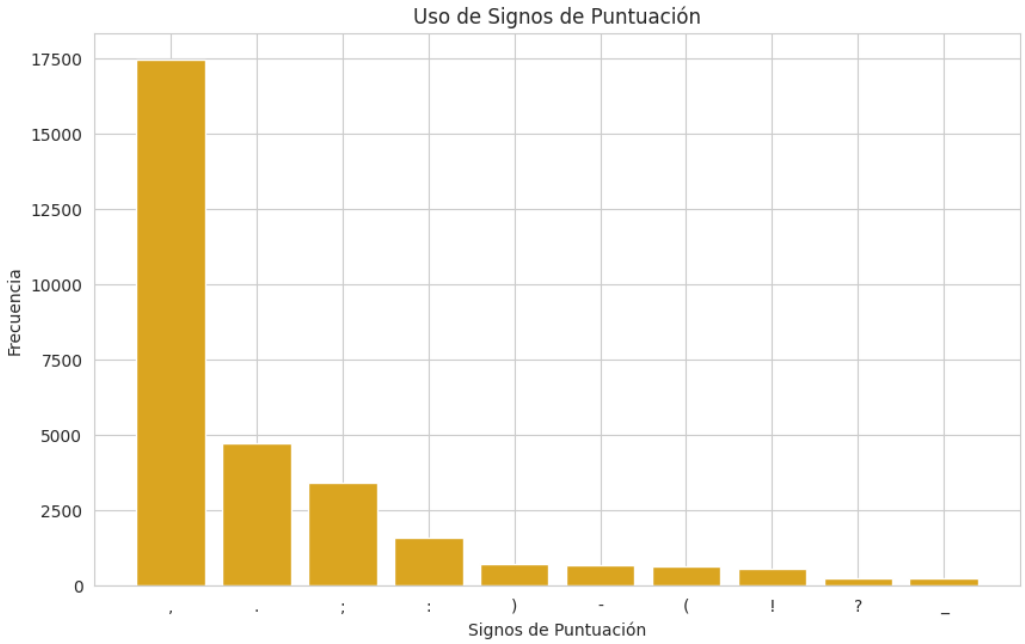

Figure 8: Signos de puntuación de La Odisea.