Práctica 1: Análisis de textos

César Alejandro Hernández Orozco
Universidad Autónoma de Nuevo León
Maestría en Ciencia de Datos
Procesamiento y Clasificación de Datos

February 10, 2025

1 Introducción

El análisis de textos es una herramienta clave en el procesamiento de lenguaje natural y la investigación literaria. Para la realización de esta investigación se decidió comparar dos obras clásicas del repositorio de libros: Project Gutenberg. Los textos elegidos son: *Laws* de Platón y *The Common Law* de Oliver Wendell Holmes Jr. Ambas obras tocan el tema de las leyes, pero desde distintos contextos. Ambas son fundamentales en sus respectivos campos, la filosofía política y el derecho. Su comparación permite entender cómo el lenguaje, la estructura y el estilo varían según el propósito del texto y la época en que fueron escritos.

Comparar estos dos textos es relevante porque representan distintas tradiciones intelectuales y momentos históricos clave:

- **Platón** (siglo IV a.C.) escribe en forma de diálogo filosófico, con un estilo argumentativo y denso.
- **Holmes** (siglo XIX) desarrolla un análisis jurídico con una estructura más concisa y expositiva.

2 Análisis Estadístico de los Textos

Se realizaron diversos análisis estadísticos sobre ambos textos, incluyendo conteo de palabras, distribución de puntuación y n-gramas.

2.1 Dimensión del Texto

Característica	*Laws*	*The Common Law*
Total de palabras	241,746	135,766
Promedio letras por palabra	4.38 caracteres	4.32 caracteres

Table 1: Comparación de Características Básicas de los Textos

2.2 Distribución de Puntuación

Se analizó la frecuencia de distintos signos de puntuación para evaluar la estructura del lenguaje:

- **Platón**: uso abundante de comas y punto y coma, lo que indica oraciones largas y complejas.
- **Holmes**: predominancia de puntos finales, lo que sugiere un estilo conciso.

2.3 Frecuencia de Palabras y N-gramas

Se identificaron las palabras más comunes y combinaciones de dos palabras (bigramas) en ambos textos. Para esta investigación se encontró que los primeros bigramas están relacionados con la licencia de Project Gutenberg, esto nos da a entender que en caso de querer continuar usando estos datos, tenemos que realizar una limpieza profunda.

3 Visualización de Datos

A continuación, se presentan gráficos que ilustran los resultados del análisis, permitiendo una mejor comprensión de las diferencias estilísticas y estructurales entre los textos.

3.1 Nube de Palabras

Las nubes de palabras permiten visualizar las palabras más frecuentes en cada texto. Nos ayudan a ver los datos de forma más fácil de entender y más visual.

Figure 1: Nube de palabras - Laws by Plato

En la Figura 1, se observa que los términos más frecuentes en *Laws* incluyen "law", "state" y "virtue". Esto indica un enfoque filosófico, normativo y además muy relevante a su tiempo y su sociedad.

Figure 2: Nube de palabras - The common law by Oliver

Por otro lado, la Figura 2 muestra que en *The Common Law* las palabras más frecuentes son "law", "case", "court" y "contract". Esto confirma el enfoque jurídico y técnico del texto de Holmes, pero manteniendo algunas similitudes en cada texto.

3.2 Distribución de Frecuencia de Palabras

Para analizar la estructura del lenguaje en cada texto, se generaron histogramas con las palabras más frecuentes, excluyendo términos vacíos (stopwords) como "the", "and" o "of".

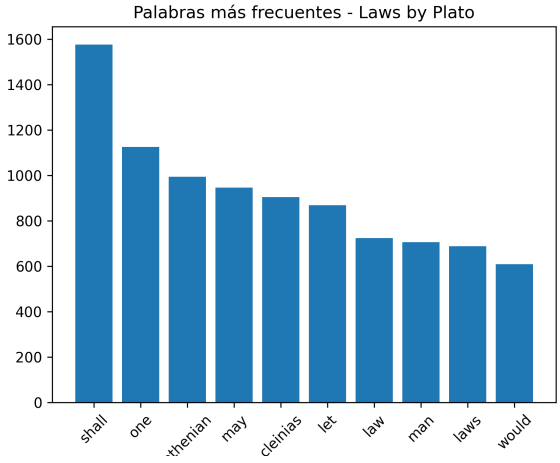

Figure 3: Palabras más frecuentes - Laws by Plato

En la Figura 3, se observa que las palabras más comunes en *Laws* son "state", "law" y "virtue".

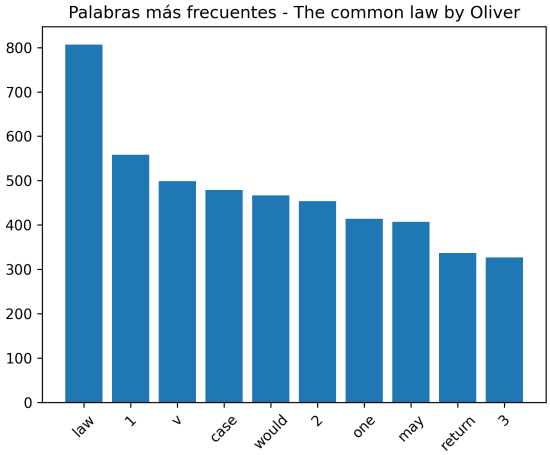

Figure 4: Palabras más frecuentes - The common law by Oliver

En la Figura 4, las palabras predominantes en *The Common Law* incluyen "law", "case" y "court".

El análisis de estas frecuencias permite entender cómo cada autor estructura su discurso y enfatiza distintos aspectos del concepto de ley. Mientras Platón desarrolla un marco filosófico, Holmes construye un argumento basado en precedentes legales y casos concretos.

4 Conclusiones

El análisis de estos dos textos ha permitido observar diferencias estructurales y estilísticas notables:

- *Laws* presenta un estilo más elaborado, con oraciones largas y puntuación densa.

- *The Common Law* es más directo, con oraciones más cortas y separaciones frecuentes.
- La comparación entre textos de diferentes tradiciones permite entender mejor las estrategias discursivas de cada autor y su impacto en la comprensión del lector.

Este análisis inicial es un buen inicio para aprender a trabajar con el análisis de textos.

References

[1] Plato. *The Laws*. 350 BC. Project Gutenberg.

[2] Oliver Wendell Holmes Jr. *The Common Law*. 1881. Project Gutenberg.