

MLE Modelo de estimación de máxima verosimilitud



17 de septiembre de 2025

alumno: rodriguez aguilar francisco jesus

GRUPO: V9

# 1. EL CORPUS TOKENIZADO

* Se debe mostrar cómo quedó el texto después de aplicar:
  1. **Conversión a minúsculas: :** Este paso es crucial para asegurar que las palabras sean tratadas de manera uniforme, evitando duplicados como "Palabra" y "palabra".

Pantalla de computadora con letras

El contenido generado por IA puede ser incorrecto.

* 1. **Eliminación de puntuación y caracteres extraños:** Se eliminan caracteres como signos de puntuación (.,!?), números y símbolos no alfabéticos que no aportan al análisis del texto.

Una captura de pantalla de un celular con texto e imagen

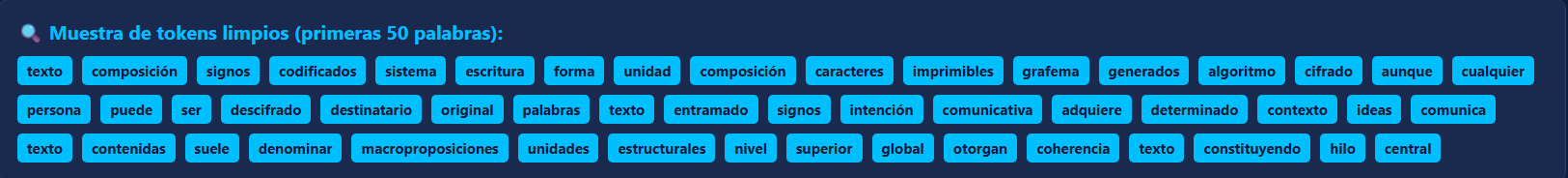
El contenido generado por IA puede ser incorrecto.

* 1. **Eliminación de stopwords:** Las stopwords son palabras comunes (como "y", "el", "de") que no aportan significado significativo al análisis. Su eliminación ayuda a resaltar términos más relevantes.

Logotipo

El contenido generado por IA puede ser incorrecto.

* **Salida del listado de tokens limpios**, primeros 50 tokens .



Esta captura sirve para ilustrar **cómo quedó el corpus antes de generar n- gramas**.

# 2. TABLAS DE FRECUENCIAS Y PROBABILIDADES CONDICIONALES

## 2.1 Palabras más frecuentes (frecuencias originales y limpias)

* **“📈 Análisis Original” y “✨ Análisis Limpio (sin stopwords)”**
* Muestra claramente:
  + **Palabra**
  + **Frecuencia**
  + **Frecuencia relativa**
  + **Barra de progreso**: Para visualizar el porcentaje de cada palabra respecto al total.
* Esta tabla muestra la **distribución de palabras antes y después de la limpieza**, permitiendo comparar el efecto de eliminar stopwords.

Pantalla de un video juego

El contenido generado por IA puede ser incorrecto.Pantalla de un video juego

El contenido generado por IA puede ser incorrecto.

## 2.2 N-gramas más frecuentes

**Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.**

* **“🔗 N-gramas más frecuentes”** de la página.
* Incluir la tabla que muestra:
  + **(posición)**
  + N-grama
  + Frecuencia
* **Probabilidad condicional MLE**:
  + S**ección “📐 Probabilidades Condicionales (MLE)”** de la página.
  + Mostrar tabla con:
    - N-grama
    - Frecuencia
    - Probabilidad
  + Aquí también se puede mostrar la **barra de progreso** si implementaste la visualización de porcentajes para la frecuencia relativa.

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

# 3. COMPARACIÓN CON Y SIN FRONTERAS DE ORACIÓN

* Se generan n-gramas bajo dos condiciones:
  + **Sin fronteras**: Se analizan n-gramas que no consideran los marcadores de inicio y fin de oración.
  + **Con fronteras de oración**: Se incluyen los marcadores <s> y </s>, lo que ayuda a capturar mejor el contexto de cada palabra dentro de la estructura de la oración.

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

# 4. CONCLUSIONES

El análisis del texto ha mostrado la importancia del preprocesamiento en el procesamiento de lenguaje natural (PLN). La tokenización del corpus permitió descomponer el texto en palabras individuales, facilitando la identificación de patrones significativos. La limpieza del texto, que incluyó la eliminación de caracteres no relevantes y stopwords, fue esencial para resaltar las palabras que realmente aportan valor al análisis.

Las tablas de frecuencias revelaron cómo la eliminación de stopwords afecta el conteo total, destacando que términos como artículos y preposiciones, que son muy frecuentes, desaparecieron, permitiendo así que se evidencien los términos más significativos. La comparación de probabilidades condicionales con y sin fronteras de oración demostró diferencias notables en la interpretación de los n-gramas, sugiriendo que la inclusión de estos marcadores mejora la comprensión del contexto en el que aparecen las palabras.

En resumen, este ejercicio subraya la necesidad de un preprocesamiento cuidadoso en cualquier proyecto de PLN. La limpieza del texto y la consideración de la estructura oracional no solo hacen que las métricas obtenidas sean más representativas, sino que también permiten realizar comparaciones más precisas entre diferentes enfoques de modelado. Este análisis no solo proporciona métricas cuantitativas, sino que también resalta cómo el tratamiento de los datos influye en la interpretación de los patrones lingüísticos, reforzando la relevancia de un enfoque metódico en el análisis de textos.

# 5. Captura de GitHub

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.