# Informe de Análisis de Texto con TF-IDF, NLTK y Visualización de Frecuencias

## Introducción

El presente trabajo tiene como objetivo realizar un análisis de un corpus textual mediante técnicas de procesamiento de lenguaje natural (NLP). Se aplican diversos procesos como limpieza del texto, lematización, eliminación de palabras vacías, y posteriormente, se utiliza el método TF-IDF (Term Frequency - Inverse Document Frequency) para obtener la relevancia de los términos. Finalmente, se visualiza la frecuencia de las palabras más comunes con gráficos.

## Herramientas Utilizadas

- nltk: para tokenización, lematización y eliminación de stopwords.

- scikit-learn: para generar la matriz TF-IDF.

- matplotlib: para graficar las frecuencias de palabras.

## Proceso de Canalización (Pipeline)

Paso 1: Corpus Original  
Se parte de una lista de 10 oraciones sobre lenguajes de programación.

Paso 2: Tokenización  
Se separan las palabras del texto original, eliminando signos de puntuación y transformando todo a minúsculas con word\_tokenize.

Paso 3: Eliminación de Stopwords  
Se eliminan las palabras vacías del idioma inglés (por ejemplo: "the", "is", "and") usando nltk.corpus.stopwords.

Paso 4: Lematización  
Se normalizan las palabras usando WordNetLemmatizer, lo que permite reducirlas a su forma base (por ejemplo, "languages" → "language").

Paso 5: Construcción del Corpus Preparado  
Cada oración es procesada y reconstruida con palabras limpias y lematizadas.

## TF-IDF: Frecuencia e Importancia de Términos

Se utiliza el modelo TF-IDF a través de TfidfVectorizer() para transformar el corpus en una matriz numérica, donde:  
- Cada fila representa una oración.  
- Cada columna representa una palabra distinta.  
- Los valores representan la importancia de cada palabra en una oración respecto al resto del corpus.

Resultados:  
- Matriz TF-IDF: muestra los valores de cada término.  
- Vocabulario generado: muestra las palabras únicas analizadas.

## Análisis del Corpus

Top 6 Palabras Más Frecuentes:

|  |  |
| --- | --- |
| Palabra | Frecuencia |
| python | 6 |
| javascript | 6 |
| used | 3 |
| data | 3 |
| rust | 3 |
| cplus | 3 |

Palabra Menos Utilizada:  
Se identificó la palabra menos repetida del corpus, que aparece solo una vez. Ejemplo: "strong" (esto puede variar si se modifica el corpus).

Palabras Más Repetidas en una Misma Oración:  
Se analizó la segunda oración. Se procesó individualmente y se encontraron las palabras más repetidas. Por ejemplo:  
  
['javascript', 'run', 'web', 'browser', 'python', 'used', 'various', 'application', 'including', 'data', 'science', 'artificial', 'intelligence']  
  
Ninguna palabra se repite en la misma oración, ya que cada término aparece una sola vez.

## Visualización de Resultados

Se realizaron tres tipos de gráficos para representar la frecuencia de las palabras más comunes:  
  
Gráfico de Barras:  
Muestra las 10 palabras más frecuentes.

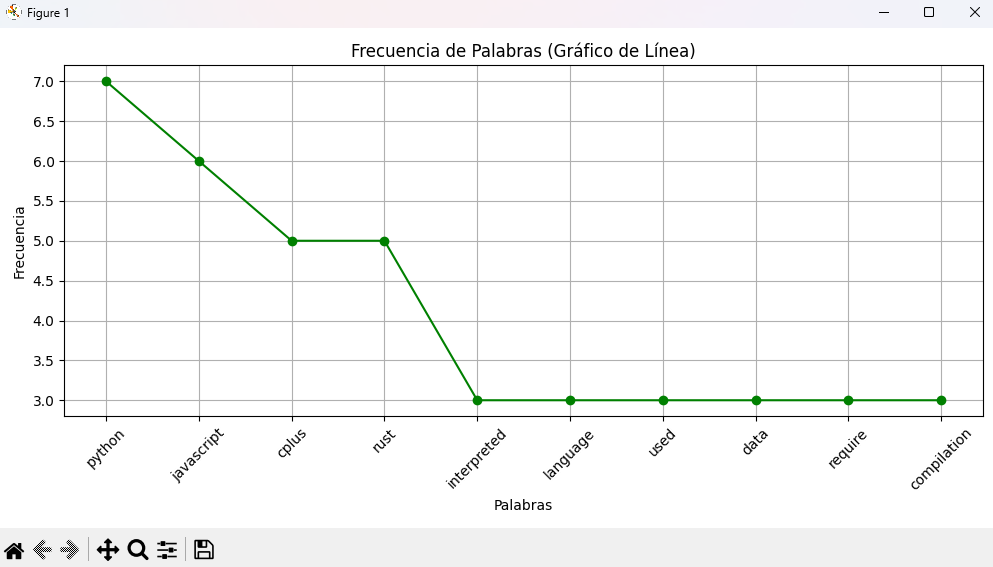


Gráfico de Torta:  
Representa la proporción de las 5 palabras más comunes respecto al total.

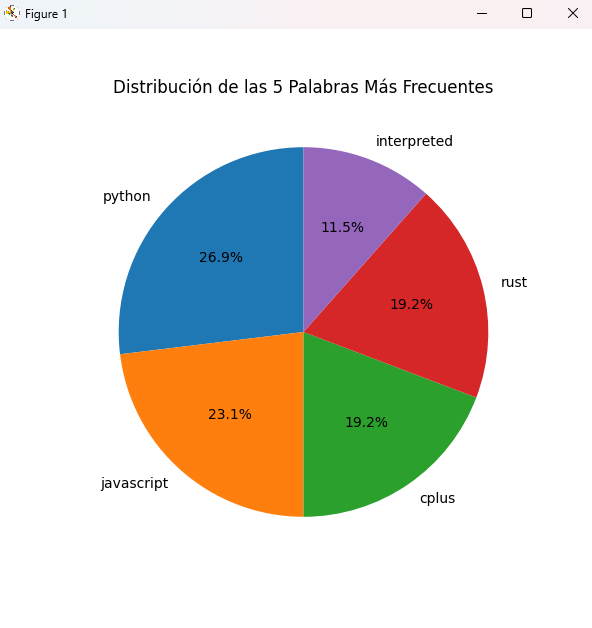
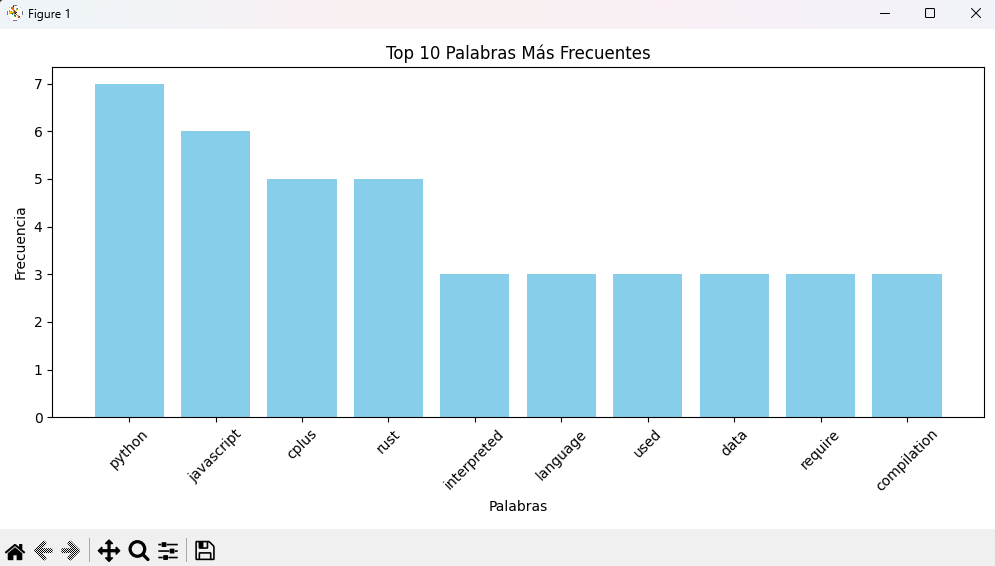


Gráfico de Línea:  
Permite observar el comportamiento de la frecuencia de palabras en forma secuencial.



## Conclusiones

- El lenguaje Python y JavaScript fueron los más mencionados en el corpus.  
- TF-IDF permitió identificar la importancia de palabras en función de su contexto en las oraciones.  
- La limpieza previa del texto es esencial para obtener resultados coherentes y evitar sesgos.  
- La visualización ayuda a interpretar rápidamente los términos más relevantes del corpus.  
  
Este análisis puede ser útil para estudios de procesamiento de lenguaje, motores de búsqueda o sistemas de recomendación.