**Informe: Análisis del Corpus "CorpusLenguajes.txt"**

**1. Jerarquía de las 6 palabras más utilizadas en el corpus:**

Después de aplicar la tokenización, remoción de stopwords, lematización y análisis de frecuencia, las 8 palabras más utilizadas en todo el corpus fueron:

| **Puesto** | **Palabra** | **Frecuencia Absoluta** |
| --- | --- | --- |
| 1 | python | 7 apariciones |
| 2 | javascript | 7 apariciones |
| 3 | cplus | 5 apariciones |
| 4 | rust | 5 apariciones |
| 5 | interpreted | 3 apariciones |
| 6 | used | 3 apariciones |
| 7 | required | 3 apariciones |
| 8 | data | 3 apariciones |

Estas palabras reflejan una fuerte presencia dónde dichas palabras son asociadas a conceptos de la programación y desarrollo de softwares.

**2. Palabra menos utilizada en el corpus:**

La palabra con **menor frecuencia absoluta** en el corpus es:

**Palabra:** *various*

**Frecuencia absoluta:** 1 aparición.

Lo que dice que esta palabra aparece solo una vez en todo el corpus, dónde se refiere a la facilidad de aprendizaje de Python para los programadores que recién empiezan.

**3. Palabras más repetidas dentro de la misma oración:**

Analizando la frecuencia interna en cada oración, se observó que:

**Palabra:** javascript

**Cantidad máxima de repeticiones en una misma oración:** 3 veces.

Refleja la importancia de JavaScript en las frases analizadas, dónde se compara el lenguaje con otros cómo Python, CPlus y Rust.

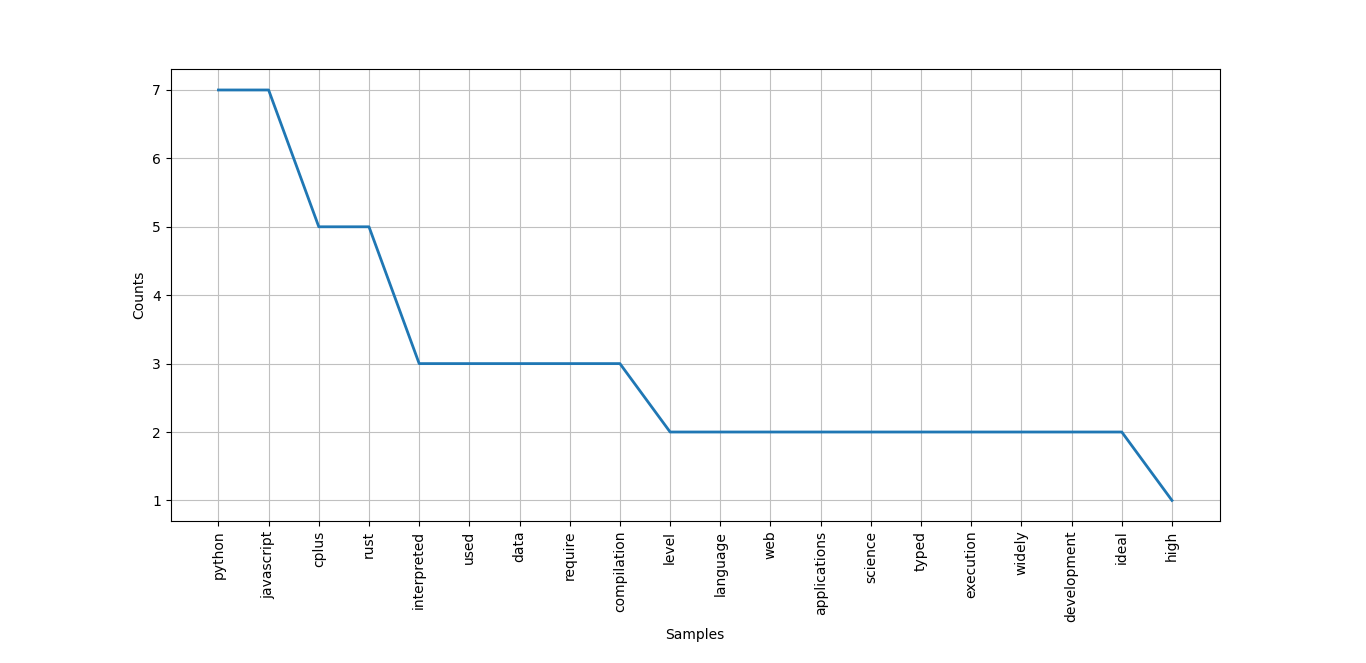
**4. Visualización de los datos:**

* Se construyó una **matriz TF-IDF** que permite observar el peso específico de cada término en las distintas oraciones del corpus.
* Se generó un **gráfico de frecuencias** para visualizar las 20 palabras más comunes, lo que facilita la identificación de términos clave de manera rápida e intuitiva.

**Conclusiones Finales**

El análisis del corpus "CorpusLenguajes.txt" permitió destacar:

* **Python** y **JavaScript** como protagonistas principales del contenido.
* La comparación entre lenguajes compilados como *CPlus* y *Rust* e interpretados como *Python* y *JavaScript*.
* El preprocesamiento textual que incluye separación de tokens, remoción de stopwords, aplicación de técnicas de lematización, combinado a su vez con técnicas de TF-IDF y análisis de frecuencia, nos otorga resultados más precisos para realizar interpretaciones más profundas del texto trabajado.

**Gráfico impreso**