Informe del Corpus TP1

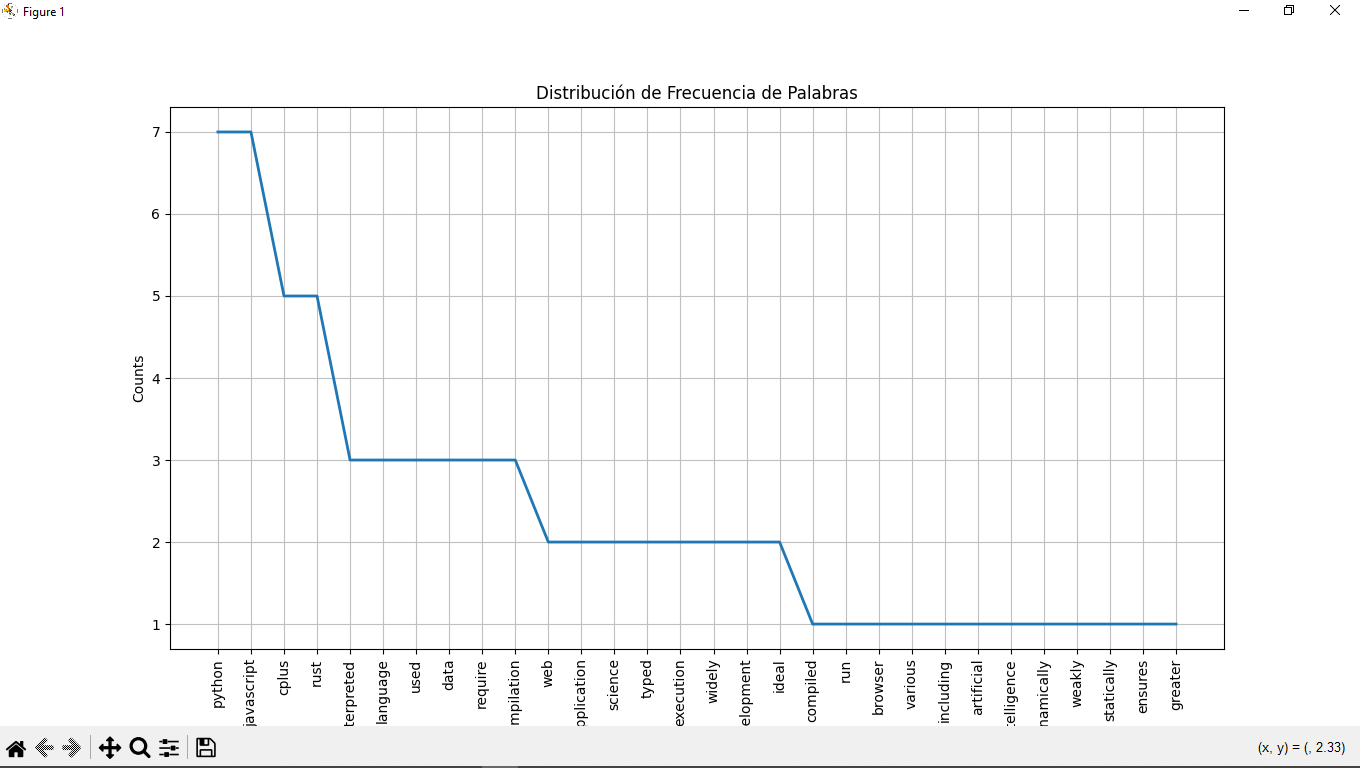
Se ha analizado el siguiente corpus, con las correspondientes técnicas de NLP

Empezando por un procesamiento del Corpus (tokenización, stopwords, lematización y vectorización Tf-Ifd) utilizando las librerías nltk, sklearn y string

Seguido de un Analisis utilizando las bibliotecas collections y matplotlib

Las 6 palabras mas usadas del corpus fueron: ‘python’,’javascript’,’cplus’,’rust’,’interpreted’,’lenguaje’.

En lo que las oraciones pueden ser consideradas de un ámbito técnico enfocado en distintos lenguajes de programación y su funcionamiento



El gráfico de frecuencia mostró una distribución típica de un corpus técnico:

-Algunas palabras con mucha repetición (términos clave).

-Una larga cola de palabras con muy poca aparición.