# Técnicas de procesamiento del lenguaje natural

# Valentina Pilar Garcia Centurion

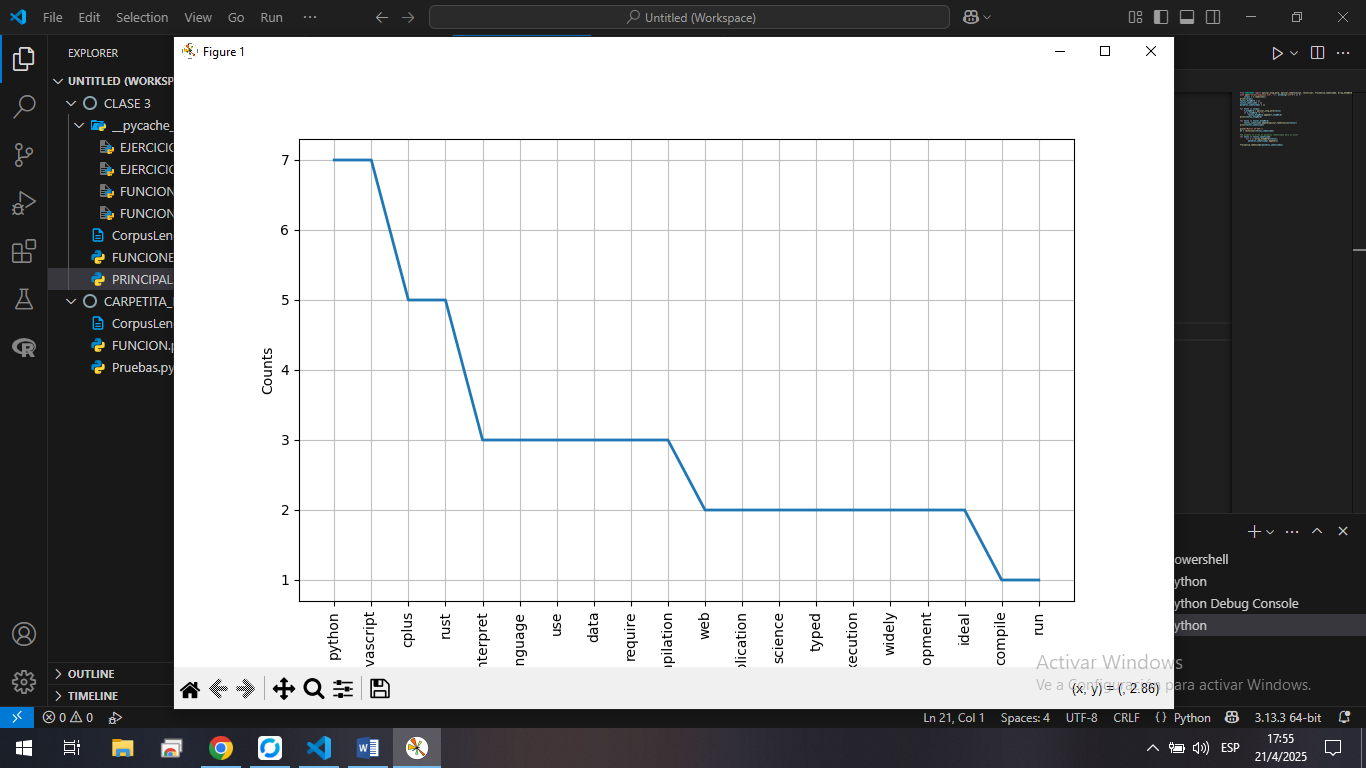
# Análisis del Lenguaje-Técnicas - Pipeline

El procesamiento de lenguaje natural (NLP, por sus siglas en inglés) es una rama de la inteligencia artificial que permite a las computadoras comprender, interpretar y generar lenguaje humano.

En este trabajo práctico se utilizaron las librerías NLTK, Scikit-learn, Pandas y Matplotllib y sus correspondientes herramientas para la ejecución de lo pedido. Además, hice una pestaña de funciones y otra principal donde está la ejecución de cada función.

Lo primero que hice fue una función para eliminar las stopwords (palabras vacias), los símbolos de puntuación y las palabras que estaban en el corpus que no hacían la diferencia para la ejecución del programa. Para esto, en la pestaña principal primero invoqué el corpus tal cual estaba e importé las funciones que iba a utilizar. Luego hice una función para hacer la lematización, donde las palabras pasan a su forma base, tanto verbos, como adverbios, como adjetivos y sustantivos.

Para saber las seis palabras más utilizadas en el corpus me base en el gráfico de frecuencias, donde las primeras seis son las que más se repiten. Estas son: Python, JavaScript, cplus, rust, interpret, language; y la palabra menos utilizada entre las primeras veinte es: compile.

El grafico de frecuencias es el siguiente:

Después hice una función para lograr el TF-IDF, donde se separan las oraciones en palabras y estas en tokens, donde cada uno se ubica en una fila. Sobre esto se aplica en dataframe, para que me arme una lista siguiendo las indicaciones anteriores, asi se calcula la frecuencia de aparición de cada palabra por sobre la cantidad de palabras del corpus.

Finalmente, se hace una función que me separa las palabras de cada oración en tokens, donde me dice las palabras que menos y más se repiten por oración, estas son:

* Oración 1: language
* Oración 3: typed
* Oración 8: require y compilation