Carrera: Ciencia de Datos e IA
Materia: Técnicas del Procesamiento del Habla
Alumno: Rojas Milagros

Informe: Análisis del Lenguaje-Técnicas - Pipeline

Se aplicaron los siguientes pasos de procesamiento:

a) Tokenización

Se usó word_tokenize de NLTK para dividir cada oración en palabras individuales.

b) Limpieza

Se eliminaron:

- Stopwords en inglés (stopwords.words("english"))
- Signos de puntuación (string.punctuation)
- Símbolos residuales como "''", '``', '--', entre otros.

c) Lematización

Se utilizó WordNetLemmatizer con la función get_wordnet_pos() para una lematización precisa según el POS tag (adjetivo, verbo, sustantivo o adverbio).

d) Unificación y análisis de frecuencia

Las palabras procesadas fueron unidas y analizadas con FreqDist() para calcular las palabras más frecuentes y generar una visualización gráfica con matplotlib.

e) Menciones específicas por lenguaje

Se extrajeron las frecuencias de palabras clave como "python", "javascript", "rust", "cplus", "java" y "go".

f) Palabras únicas

Se identificaron aquellas palabras que aparecen solo una vez en todo el corpus, ya que pueden aportar información específica o interesante.

5. Resultados

- Las 10 palabras más frecuentes se visualizan mediante un gráfico de barras.
- Python y JavaScript fueron los lenguajes más mencionados, reflejando su popularidad.
- Se identificaron varias palabras únicas, lo que sugiere diversidad léxica en el corpus.

---

7. Bibliografía / Herramientas

- Biblioteca NLTK
- Matplotlib
- Python
- Clases teóricas de Técnicas de PLN