Tarea: Word embeddings

Descarga una de las siguientes bases de datos propuestas:

medical_data.csv
`import kagglehub`
`# Download latest version`
`path = kagglehub.dataset_download("akashadesai/clinical-notes")`
`print("Path to dataset files:", path)`
Usando únicamente la columna “texto” construye un corpus linguístico con todas las notas clínicas presenten en medical_data.csv

Corpus: Movie Reviews disponible en NLTK
Corpus: Inaugural Address Corpus disponible en NLTK
GENIA corpus

Una vez seleccionado el corpus, realiza los siguientes pasos:

1. Limpia y tokeniza los textos
2. Genera la gráfica de la Ley de Zipf.
3. Construye los siguientes modelos:
    a. CBOW model y Skip-Gram model
    b. Glove
    c. FastText
4. Visualiza los word-embeddings generados por cada modelo.
5. Calcula la similitud coseno de las 10 palabras más frecuentes.
6. Compara los resultados entres los cuatro modelos generados, ¿observas alguna diferencia relevante entre ellos?