TP-5: Word Embeddings

1. Word2Vec

1) Install required libraries (gensim, nltk, sklearn, plotly).
2) Download NLTK Brown corpus.
3) Load Brown sentences.
4) Preprocess: lowercase words, remove punctuation, keep non-empty sentences.
5) Train a Word2Vec model using: vector_size = 100, window = 5, min_count = 5, sg = 1 (Skip-gram), epochs = 10
6) Print the vector for "king" (first 10 values).
7) Show 5 most similar words to "woman".
8) Perform analogy: king – man + woman.
9) Check if "government" is in the vocabulary.
10) Print the vocabulary size.
11) Select sample word list for visualization.
12) Filter words that exist in the Word2Vec vocabulary.
13) Apply PCA to reduce vectors to 2D.
14) Create a Plotly scatter plot with labels.
15) Print the first 500 characters of the Plotly JSON.

2. TF-IDF

1) Prepare Corpus: Collect documents, lowercase, remove punctuation, tokenize.
2) Compute TF: Count word frequency per document; normalize by total words.
3) Compute IDF: Count in how many documents each word appears.
4) Compute TF-IDF: Multiply TF × IDF for each term in each document.
5) Create TF-IDF Matrix: Rows = documents, Columns = terms, values = TF-IDF scores.
6) Explore Results: High TF-IDF → important/unique words; Low TF-IDF → common words.
7) Dataset: https://archive.ics.uci.edu/dataset/331/sentiment+labelled+sentences