Homework 2

1. Word2Vec
    a. Write Python program to implement Skip-gram Word2Vec algorithm.
    b. Write Python program that implements CBOW Word2Vec algorithm
    c. Apply both programs to the following text:
        i. The bank is located near the river.
        ii. The bank approved my loan application.
        iii. He rose from his chair to close the window.
        iv. The rose bloomed beautifully in the garden.
        v. The lead actor delivered a stunning performance.
        vi. Exposure to lead is harmful to health.
        vii. She is reading a book in the library.
        viii. The book mentioned a fascinating historical event.
        ix. I need to file a report for my manager.
        x. He lost the file containing important documents.
    d. What is the difference between the embeddings? explain the results.
    e. Can you find a text that its embedding will be similar in these two algorithms?
    f. Repeat step c with different window sizes. Is there a significant change?
    g. Can you compare these two models in terms of capturing the syntactic and the semantic relationship between words.
    h. Demonstrate the difference between CBOW and Skip-grams in terms of cosine similarity between the following words: bank, rose, lead, book and file.
    i. How can the subword embeddings be applied?

2. Create example sentences demonstrating how contextual embeddings handle words with multiple meanings (polysemy) differently than static embeddings like Word2Vec.
3. Propose metrics for evaluating word embeddings that can differentiate between syntactic and semantic relationships.

4. Use the Gensim library to train a Word2Vec model on a custom corpus.
    a. Evaluate the quality of embeddings by calculating the cosine similarity for the following word pairs:
        i. "king" and "queen"
        ii. "man" and "woman"
        iii. "apple" and "orange"
    b. Write a brief explanation of the results.
5. Train the GloVe model using the glove-python package on a subset of a publicly available dataset (e.g., Wikipedia, or a smaller custom corpus).
    a. Use t-SNE or PCA to visualize the embeddings in 2D.
    b. Analyze the clustering patterns observed in the visualization.