TP-4: CBOW Model for Word Embeddings

Introduction

The Continuous Bag of Words (CBOW) model is a neural network-based approach for learning word embeddings. In CBOW, the context words are used to predict the target word. This exercise allows you to train a CBOW model on a given text corpus (any dataset) and learn embeddings for each word in the vocabulary.

Implementation Details

The CBOW model in this task is implemented in Python using NumPy. The key components of the implementation include:

- **Split sentences:** Break text into sentences.
- **Make vocabulary:** List all unique words.
- **One-hot encode:** Turn each word into a simple vector with 1 and 0.
- **Prepare training data:** Make pairs of a word and the words around it.
- **Initialize weights:** Start with random numbers for the model.
- **Forward pass:** Use surrounding words to guess the target word.
- **Calculate loss:** Check how wrong the guess is.
- **Update weights:** Adjust numbers to improve guesses.