Lab Report on Exercise 2 - To Embed or Not to Embed

1. Decisions and Methodology

Preprocessing

For both the Hotel Reviews and Sci-Fi datasets, the preprocessing steps were crucial to ensure the quality of the word embeddings. The steps included:

Lowercasing: All text was converted to lowercase to maintain consistency and reduce the vocabulary size.

Punctuation Removal: All punctuation marks were removed to focus on words only.

Tokenization: The text was split into individual words using NLTK's word_tokenize function.

Class Structure

The Continuous Bag of Words (CBOW) model was implemented using PyTorch. Key aspects of the model included:

Embedding Layer: Transforms word indices into dense vectors of fixed size.

Linear Layers: Two linear layers to map the embedded words to the context.

Activation Function: ReLU activation was used between linear layers for non-linearity.

Output: The final layer outputs the log probability of words.

2. Differences Between CBOW2 and CBOW5

The primary difference between CBOW2 and CBOW5 lies in the width of the context window. CBOW2 considers two words before and after the target word, whereas CBOW5 extends this to five words in each direction. This variation leads to:

CBOW2: Tends to capture more immediate and specific context.

CBOW5: Captures a broader context, potentially understanding more abstract relations but may include irrelevant information.

Conclusion: Both models have their merits and are suited for different applications. The choice between CBOW2 and CBOW5 depends on the desired granularity and the specific use case of the word embeddings.