Project #3 for the Biomedical Information Retrieval Course

Name: Phan Ben

ID Student: P76127051

Project Overview:

In this project, I will implement the Word Embedding Technique, specifically using the Continuous Bag of Words (CBOW) model. CBOW involves utilizing a word window to predict the central word. My task is to apply this technique to a collection of approximately 1000 text documents related to the biomedical field and integrate it into a search engine to enhance search algorithm performance.

Key Features:

1. **CBOW Implementation:** The project exclusively focuses on implementing the Continuous Bag of Words (CBOW) model for word embedding.
2. **Neural Network Training:** The CBOW model will be trained using a neural network, with PyTorch as the chosen framework for this purpose.
3. **Data Collection and Preprocessing:** Gather and preprocess approximately 1000 text documents from the biomedical field to create a suitable dataset.
4. **Dimensionality Reduction:** Apply the Principal Component Analysis (PCA) algorithm to reduce the word embeddings to 2D, facilitating the visualization of the model's embeddings.
5. **Visualization:** Utilize the reduced-dimensional embeddings to visualize the relationships between words and their context in the biomedical domain.
6. **Integration into a Search Engine:** Integrate the CBOW model into a search engine to enhance the search algorithm's capabilities in retrieving relevant information.

Conclusion:

In this project, my primary goal is to apply the Continuous Bag of Words (CBOW) model to create word embeddings that improve our understanding of word relationships. These embeddings will be integrated into a search engine, enriching the system's capabilities to provide more precise and context-aware search results. This holistic approach will contribute to the development of a robust and efficient information retrieval system within the biomedical domain.