CSE440: Natural Language Processing II

Lab Assignment 2

1. Download the IMDB movie review dataset. Preprocess the review texts by tokenizing, converting to lowercase, and removing punctuation. Then, use CountVectorizer to convert the preprocessed texts into Bag-of-Words feature vectors. Report the dimensions of the resulting feature matrix.

2. Using the same preprocessed IMDB dataset, apply TfidfVectorizer to generate TF-IDF embeddings. Identify the top 10 words with the highest TF-IDF scores.

3. Obtain the GloVe embeddings (glove.6B.100d.txt). Perform analogy tasks such as "Teacher - Educate + Heal" and check whether the resulting vector is closest to "Doctor" using the GloVe embeddings.

4. Select Brown corpus and load the text data. Preprocess the text by tokenizing, converting to lowercase. Train Word2Vec models using both Skipgram and CBOW on your chosen corpus (gensim.models.Word2Vec). Evaluate the trained models on word similarity tasks and the same analogy tasks from the previous questions.