CSE440: Natural Language Processing II

Lab Assignment 2

1. Download the IMDB movie review dataset and preprocess the text by tokenizing, converting it to lowercase, and removing punctuation. Next, apply a TF-IDF vectorizer (`sklearn.feature_extraction.text.TfidfVectorizer`) to transform the corpus into TF-IDF embeddings. Split the dataset into 70% training and 30% testing data, ensuring stratification. Train a Logistic Regression model using the scikit-learn library and evaluate its performance by computing the F1 score. Report the confusion matrix along with the classification report.

2. Obtain the GloVe embeddings (`glove.6B.100d.txt`). Perform analogy tasks such as "Queen – Woman + Man" and check whether the resulting vector is closest to "King" using the GloVe embeddings.

3. Select Reuters 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. Select the most frequent 30 words from your train corpus. Apply a dimensionality reduction technique (PCA: `sklearn.decomposition.PCA`) to the embedding vectors for these words. Plot the resulting 2D projection. Label each point with its corresponding word to observe clusters or semantic groupings. Compare the results of both models.