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 **80% training** and **20% testing data**, ensuring stratification. Train a **K-Nearest Neighbors (KNN) classifier** model using the **scikit-learn** library and evaluate its performance by computing the **F1 score**.

2. Obtain the **GloVe embeddings** (`glove.6B.200d.txt`). Perform analogy tasks such as **"Queen – Female + Male"** and check whether the resulting vector is closest to **"King"** using the GloVe embeddings.

3. Select **Reuters/Gutenberg** corpus and load the text data. Preprocess the text by tokenizing, converting to lowercase. Train a **Word2Vec** model on your chosen corpus (`gensim.models.Word2Vec`). **Evaluate** the trained model on word similarity tasks and the same analogy tasks from the previous questions. Select the most frequent 100 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.