1. Using the bigram and unigram language models trained on the training data from the Assignment 1, compute the PMI scores for all the bigrams in the validation and testing sets created from Assignment 1.
2. Vectorize all the sentences in the training, validation, and testing data that you tokenized from Assignment-1 using TF-IDF. For the validation and testing data, use the IDF scores learned from the train data.
3. For each sentence in the validation and testing sets, find its nearest neighbor in that set using TF-IDF vectors.
4. **Bonus Question:** Can you extend question number 2 to the training data? For each sentence in the validation and testing sets, find its nearest neighbor in the training set using TF-IDF vectors. How will you compute the number of operations (as the training data size is huge) effectively?