Programming Assignment-1

Due Date: Wednesday October 6, 2021

Total Points: 100

1. In this programming assignment, your task is to conduct exercise on the dataset that you are intended to use in your course project. Please mention the dataset before working on the programming assignment and show output for the specific tasks in a separate file. [**points 30**]
    a. Please write down a code to show basic statistics of the datasets –
        i. number of vocabularies,
        ii. Percentage of training and test dataset and show number of sentences, number of tokens before and after normalization,
        iii. Number of annotations in training and testing. The number of samples in each class for training and testing set.
        iv. Write code to perform normalizations (e.g., tokenization, sent. segmentation, removing stop-words, case-folding, etc.) on your datasets
2. Compute feature vectors for your dataset using following methods [**points 20**]
    a. Compute co-occurrence matrix for the vocabulary assuming context window size ± 5.
    b. Compute TF-IDF feature vectors.
3. Extract feature vector from existing pre-built vector representation such as fasttext, word2vec, globe. Use one from the sources for your exercise. Please follow the lecture slide for corresponding links. [**points 10**]
4. Calculate feature vector for each sentence in the training and testing set. Use question 3 feature vectors to perform this computation. Find the top 10 similar sentence pair such that each pair consists of one training sentence, and one testing sentence. Pair can be presented as $Sent_{pair}$ (S_{train}, S_{test}). Compute similarity using cosine distance. [**points 40**]

    Note: *Please write functions for each specific task you are performing so that you can use these codebases for your project later.*