Joseph Shaffer

shaffer.567

CSE 5243 Programming Assignment 1

Report

The goal of this assignment was to construct a feature vector for each sentence in the data set. Thus for M sentences and N words in the data set, a M X N matrix D must be constructed, where D_i_j means the count of word j in sentence i.

My program uses a nested dictionary to resemble the feature vector that needs to me constructed. The key values of the outer dictionary are the sentence numbers of the data set and the value paired with each key (sentence number) is an inner dictionary. Each inner dictionary has the key values as each word that appears in the current sentence and the values paired with each key is the count for that word in the current sentence. Thus, we resemble the M X N matrix D with a nested dictionary, where we can find the count of word j in sentence i with dict[i][j] in preprocessing.py.

Thus, the program completes the goal of this assignment. The program will output a sample of the data, where it outputs the count for each word that is found in each one of the following sentences: 500, 1000, 1500, 2000, 2500, and 3000. The output can be formatted in different ways to see different outputs such as seeing the most popular word in each sentence. Thus, different things can be learned or looked at for editing the output of the program.

I only filtered the data by making sure that a word is not added twice in the same dictionary, but instead the count for that word is just updated. I did not filter the data in any other way, because without knowing what in the data I am trying to look for, I do not know how to filter the data to search for more relevant information than the count for each word in each sentence.

If we have a more detailed idea of what we are hoping to find or looking for in the data, the data can be filtered using the current feature vector of a nested dictionary in different ways to help find relevant information.