**CMPE 255 - PROGRAM1**

Name – Nishchay Madan

SID -011824501

Rank- 7

Accuracy- 0.82

**Approach and Methodology**

* Combining the two data sets i.e. train.dat and test.dat into one list.
* Filtering the list for the words which are shorter than 4
* Building a csr sparse matrix.
* Decreasing the popular words with the help of idf.
* Normalizing the matrix.
* Computation of cosine similarity by splitting the matrix into Xs and Ys.
* Sorting the matrix be np.argsort.
* Rank the top values based on their neighbors.
* Output is stored on format.dat file.