**Computer Engineering Department**

**Course Name: CMPE255 – Data Mining**

**Student Name:** Sahitya Mullapudi

**Sjsu Id:** 011545404

**Program 1:** Movie review sentiment

**Semester:** Fall,2017

**Rank**: 8

**Accuracy**: 0.8164

**Approach:**

1. The train set is divided into classes and reviews.

2. The test set and train set reviews are then combined.

3. The data is preprocessed for html tags, special characters.

4. The documents are split into documents of words and changed to lower case.

5. The words of length less than 4 are filtered.

6. csr matrix is built for the processed data. Tf-idf and l2 normalization are done on csr matrix. (csr\_matrix, csr\_info, csr\_idf, csr\_l2normalisation in Activity\_data\_3 are used).

7. The test and train set are divided.

8. Cosine similarity is calculated by dot product of the two matrices as the matrices are already normalized.

9. k neighbors are calculated by sorting each row in the cosine similarity matrix.

10. K is tested for different values.

**Methodology:**

The data is preprocessed as it contained html tags and special characters. The test data and train data sets are combined, so as to get equal features or columns in both the sets. The words of length lesser than 4 are filtered as most of the words are like – ‘is’, ‘the’, ‘a’ and so on. Those words do not contribute to sentiment analysis. A sparse matrix is built from the processed data. The term frequency–inverse document frequency is applied, so that importance of frequently occurring words is decreased.

Cosine similarity is chosen because it is used commonly in text mining and comparing documents. Cosine similarity is calculated for every document of test set with all the documents of train set. The standardization was not applied as the values are close by and range in between 0 and 1. For calculating k nearest neighbors, argsort() is used, as it sorts based on the index which makes the process of accessing classes of respective reviews easy. K is tested for different values – 80, 100, 100, 150, 200. The movie sentiment prediction is more accurate when k=150. When k=200, the accuracy decreased significantly.