Movie Review Classification

Name: Meenakshi Paryani **SJSU ID:** 011819392

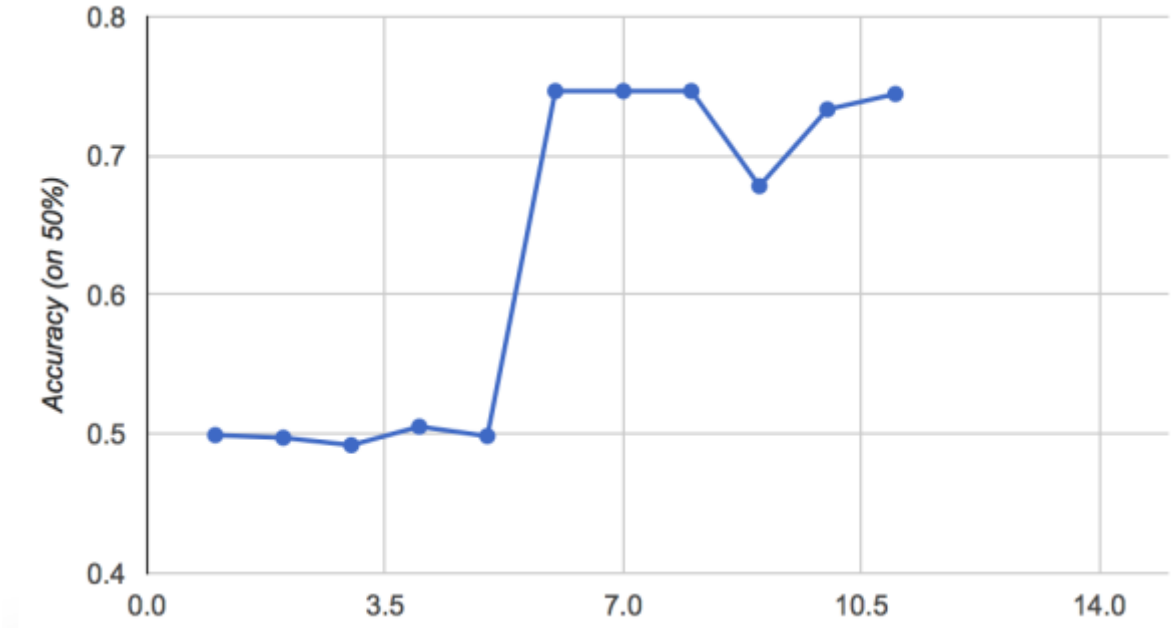
Rank: 18
Accuracy: 0.7445

Approach: The classification of the movie reviews datasets was done using K nearest neighbor classification. There was a training data which was fed to the model which contained 25000 reviews and their rating. The model was then used to compute the sentiments of the test dataset review into positive or negative reviews/classes. On a high level, following steps were followed:

- **Pre-Processing:** Removed the unwanted content from the dataset(test+train) which will actually act as noise if present and doesn't actually contribute to the classification. Following pre-processing steps were followed:
    - Removed digits from the data as they don't add any value
    - Removed words which are less than 4 characters long
    - Used the Lemmetizer to reduce all the morphs of the same word to their root ex- 'watch', 'watched', 'watching' are reduced to the same name
    - Used the Stemmer to stem the words like 'mixed' reduced to 'mix'
    - Converted all the words to lower case for accurate comparison
- **Building CSR Matrix:** computed the CSR matrix of training and test data to be able to compute cosine similarities of the data
- **Inverse Document Frequency transformation:** Applied the IDF methodology on matrix to achieve better accuracy
- **Normalization:** Normalized the values to have accurate results
- **Classification:** As a final step, classifying the test data records as positive/negative review, we perform the following steps:
    - Use the computed cosine similarities of training and test matrices
    - Compute the dot product of the cosine similarities of a given test record with all the cosine similarities of the train dataset
    - Decide the value of k, value of k should not be a multiple of number of classes as it can produce a tie, here as there are 2 classes, the value of k is selected as odd

- Determine the k nearest neighbors(similarities) of the test record cosine similarity with the training data similarities
- Take majority vote of the neighbors and accordingly classify the test record as “Positive” or “Negative” review

Methodology for Approach: At first, the pre-processing was not involved and without that the classifier accuracy was reported very low in range of 0.49. After applying the pre-processing, the accuracy improved a little and came in the range of 0.55. After applying the IDF transform, the accuracy improved more and came in the range of 0.67. Initially the value of k taken was 71 after testing with few values of k. The accuracy was improving by lowering the value of k. Started testing with k in the range of 11 and 23 as suggested by ISA and then the maximum accuracy was found at k=23 as shown in below graph



Leaderboard:

Rank	Accuracy (on 50%)	User ID	Submission Count
1	0.8610	11428040	3
2	0.8594	11815440	11
3	0.8583	11430939	37
18	0.7466	11819392	11