STAT 542 Project 4 Report

Ankush Agrawal (NetID : ankusha2)

I did the following for classifying the sentiments of a movie review:

1. I cleaned the entire corpus of texts by removing the stop words like ‘a’, ‘the’, ‘an’ etc. I created a bag of words models using 3000 most frequent words in the entire set of our data. This is used to map our reviews to create a vector for each review where the numerical value in the 3000-long dimension represent the number of times the word has occurred in the review.
2. I used term frequency and inverse document frequency concept which will help reflect how important a word is to a document in a collection of corpus. The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general.
3. I ran a simple logistic regression with ridge penalty using a 5 fold cross validation.
   Here are the results from my run using 3 test train splits

Performance for split 1	0.9428
Performance for split 2	0.9433
Performance for split 3	0.9446
Vocab size	3000

Here are the details of my session info :

```
R version 3.5.1 (2018-07-02)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows >= 8 x64 (build 9200)
```

I used the following packages :

1. pROC_1.13.0
2. tm
3. stringr
4. tidytext
5. glmnet