Project 4 Report

Overview of the model:

Background:
We are provided with a dataset consists of IMDB movie reviews, where each review is labelled as positive or negative.

Problem statement:
We need to predict the sentiment of a movie review.

Objective:
Evaluation metric is AUC on the test data. Full credit for submissions with minimum AUC over three test data equal to or bigger than 0.96.

Input and output of your model:
Input is a document-term-matrix, each row is one document, each column is a word, each entry is the frequency of that word in the document.
Output is the probability of the positive sentiment for each document.

Customized vocabulary:

After reading the data,
first, I remove all the html tags and punctuations.
Second, the movie review has been tokenized and transformed to lower case. Start from all the data to build the customized vocabulary, first, create the vocabulary with ngram of size 2 and remove the stop words.
Third, prune the vocabulary, only keep those words which at least appear 10 times in all documents, at least 0.1% of all documents contain the word, at most 50% of all documents contain the word.
Fourth, we use the pruned vocabulary to build the vectorizer.
Fifth, create the document-term matrix of the third train data and test data based on the vectorizer.
Sixth, calculated the t-statistics for each vocabulary from different sentiment groups and extracted the first 2000 words with the largest absolute value of t-statistics. Save the vocabulary list for the first and second train, test data.

Technical details:

Type of algorithms/models used/explored:
We used the logistic regression with L2 penalty.

Pre-processing:
Only use the vocabularies from the selected vocabulary list.

Training process:

Use the cv.glmnet with 10 folds to get the min lambda. And use the min lambda to fit the model with train data.

Tuning parameters:
Min lambda used.

Model validation:

Performance on the three data sets:

Performance for split1	0.9659285
Performance for split2	0.965234
Performance for split3	0.9634743
Vocab Size	2000

Discussion on model limitations:
Bag of words can't catch the sequential information. Logistic regression with L2 penalty can't be further tuned. It is not quite flexible.

Explanation on errors:
There are two examples of misclassified examples:
"I desperately want to give this movie a 10 I really do Some movies especially horror movies are so budget that they are good A wise cracking ninja scarecrow who can implement corn cobs as lethal weaponry definitely fits this budget to brilliance system The depth of the movie is definitely its strong point and the twists and turns it implements keeping the audience at the edge of their seats really drives the creepy ninja puberty stricken pre thirty year old student non cowboy drawing wise cracking son of a bitch scarecrow into the limelight as the creepiest horror icon of the year All I can really say is can you dig it and recommend watching movies such as Frankenfish if you enjoy this sort of hilarious horror WHAT THE HELL WERE THEY SMOKING"

"It would be wrong and reprehensible of me to advise you to watch Killjoy 2 you must have better things to do washing the car throwing stones in a stream but at the same time it s nowhere near as awful as you probably think it is It s almost a proper film which a lot more than most straight to DVD sludge can manage Killjoy 2 is helped a great deal by Trent Haaga s manic turn as the eponymous clown he throws himself into the role with such fevered abandonment that he almost tips the scales in the movie s favour but of course it takes more than one man in big shoes Tammi Sutton gives the most entertaining director cameo since Roger Corman in Creature from the Haunted Sea and the whole thing is nearly destroyed by the rushed sugary ending All over the place and almost good fun"

These are two negative reviews which are misclassified. If we read the review, we will realize the author just scoffs the film. But we simply use the bag of words model which doesn't

consider the sequential information. As the review does have many positive words. It is reasonable to have this error.

Future steps you could take to improve your model:
Further manually prune the vocabulary, get rid of the words which have similar meanings.
Use the recurrent neural network with LSTM layer which can detect the sequential information.

Run time:
1.29230313301086 mins

Computer System:
Macbook Pro, 2.3GHz, i5, 8GB memory.