So for this part, we first focus on the review text, we make some change on the original stop word dictionary so that we wont miss some important words such as don’t can’t or haven’t, right now we just remove all the punctuation for convenient. We also lower the word and make lemmatization so that the single word with different form are the same.

Next, we transfer the text into bag of word and wordvector. We are trying to fit the bag of word with svm, lm, and logistic and fit the wordvector into cnn and gbm.

In general, we assume the more sample we have, the better model we can fit. But as it turns out, we found some model like svm perform much better in the small dataset. And we guess maybe the complexity of the language also increase at the sample grow up.

We will show more details in next week and our plan for this weekend is to combine the basic information and the text together, and also create some features by our prior knowledge.