Vocabulary List Construction

To create the vocabulary list, we first removed the HTML tags attached to each movie review. Then, we created the document term matrix (DTM) by using all the movie reviews inside “alldata.tsv.” To achieve this DTM, we first removed stop words, or commonly used words, such as “the,” “a,” and “they.” Then using R’s text2vec package, we tokenized the words into terms with minimum of 1 word and maximum of 4 words, including terms that appear at least ten times or more, that appear in at most 50% of the reviews, and appear in in at least 0.1% of the reviews. With these steps, the DTM is still quite large as we managed to shrink the vocabulary size from 50,000 reviews to 30,000 terms. To shrink down the DTM even more and to avoid overfitting and increase interpretability of the model, we fitted a logistic regression model (“family = binomial”) with Ridge regression (“alpha = 0”) on the DTM and picked the set of estimated term coefficients with the largest degree of freedom just below 1000. We than manually selected this set, which is the 36th row with 977 terms to get the vocabulary test for training the sentiment analysis binary classification model using glmnet.

```
library('text2vec')
```

```
## Warning: package 'text2vec' was built under R version 4.2.2
```

```
library('glmnet')
```

```
## Warning: package 'glmnet' was built under R version 4.2.2
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-4
```

```
library('data.table')
```

```
## Warning: package 'data.table' was built under R version 4.2.2
```

```
set.seed(5178)


# remove html tags
train = read.table("alldata.tsv",
                   stringsAsFactors = FALSE,
                   header = TRUE)
train$review = gsub('<.*?>', ' ', train$review)

# construct document term matrix
stop_words = c("i", "me", "my", "myself",
               "we", "our", "ours", "ourselves",
               "you", "your", "yours",
               "their", "they", "his", "her",
               "she", "he", "a", "an", "and",
               "is", "was", "are", "were",
               "him", "himself", "has", "have",
               "it", "its", "the", "us")
it_train = itoken(train$review,
                  preprocessor = tolower,
                  tokenizer = word_tokenizer)
tmp.vocab = create_vocabulary(it_train,
                              stopwords = stop_words,
                              ngram = c(1L,4L))
tmp.vocab = prune_vocabulary(tmp.vocab, term_count_min = 10,
                             doc_proportion_max = 0.5,
                             doc_proportion_min = 0.001)
dtm_train  = create_dtm(it_train, vocab_vectorizer(tmp.vocab))


# trim vocab size
tmpfit = glmnet(x = dtm_train,
                y = train$sentiment,
                alpha = 1,
                family='binomial')
tmpfit$df
```

```
##   [1]     0     1     2     3     4     4     6     7    11    15    18    22
##  [13]    25    39    48    57    67    83    97   113   131   152   174   208
##  [25]   238   272   303   337   388   437   489   558   638   740   857   976
##  [37]  1125  1279  1454  1684  1932  2219  2529  2872  3232  3592  4010  4393
##  [49]  4778  5165  5581  6021  6433  6818  7204  7547  7910  8242  8569  8867
##  [61]  9118  9391  9651  9881 10111 10332 10541 10698 10846 10971 11105 11259
##  [73] 11410 11508 11602 11708 11784 11875 11936 12006 12093 12150 12202 12266
##  [85] 12308 12362 12405 12449 12475 12585 12646 12690 12733 12773 12814 12827
##  [97] 12866 12894 12918 12941
```

```
myvocab = colnames(dtm_train)[which(tmpfit$beta[, 36] != 0)]

write.csv(myvocab,"myvocab.txt", row.names = FALSE)
```