Sentiment Analysis over Yelp Business Reviews

paperezq

19 de noviembre de 2015

Contents

Introduction

This report aims at describing a set of classifiers in which a review vote or qualification is given based-on a set of words identified on a review. This is part of the Data Science capstone and the dataset used comes from the Yelp Dataset Challenge.

With the idea of connecting people to great local business, yelp provides five datasets describing business, checkin information, customers' reviews of businesses, tips and user data.

My question of interest for this project is how well can a tool guess a review's rating from its text alone?. Basically, the idea is measure the how well a classifier performs based-on some positive and negative words from text reviews. Nevertheless, words of other kind of lexicons were used. We will see that the models built used more than positive and negative words.

The following sections describes the analysis made over the reviews dataset in order to build classifiers of punctuations or stars qualifications. In the first section some basic concepts are mentioned. The second section explores tools and methods used over the dataset. An explanation and some performance measures of the classifiers built are describe in the third section. Finally, results are discussed and some conclusions are given.

Some concepts: Sentiment Analysis

Natural Language Processing (NLP) is a technology in which text, in a manner of documents, are processed aiming at to find new knowledge and to answer questions related to patterns that can be found. This is also named text mining. Some of the tasks in NLP includes words and sentence tokenazition, prediction on typing words, text classification, information extraction, and some others.

Text mining could be consider an extension of data mining because it includes the use of methods for classification and clustering. Thus, the variables that are likely manipulated in this kind of mining are counts of some words previously identified in a preprocessing step.

Inside this technology, exists a task so called sentiment analysis. This tasks has the challenge of predicting an election or rate based on some texts given by users of a particular matter. i.e. movies reviews, ease of use of an electronic device, qualifications of a service, etc. In sentiment analysis, the use of lexicons is important in order to identify affective states like emotions, attitudes, pleasure, pain, virtue, vice, and so on. For more information, see the coursera of NLP at standord.

In the next section, is described the text mining tools and the dataset of lexicons got for this project.

Data is explored: tools and methods used to get an exploratory data analysis

In order to get a good classifier model for ratings based on text, R tool is appropiate for such as job. The datasets are in json file format. The library *jsonlite* is suitable for treating this kind of files. Primeraly, jsonlite library has to main functions: toJSON and fromJSON. However, the dataset of reviews is the largest of the five datasets with 1.32GB. Because of its size and in terms of performance, the best function for reading these json files is the **stream_in** with the name of the file as a parameter. This function lets read the json file faster than the fromJSON function does. Although, it considers a slightly different format.

```r
require(jsonlite)
#data_business <- stream_in(file("yelp_academic_dataset_business.json"))
#data_checkin <- stream_in(file("yelp_academic_dataset_checkin.json"))
data_review <- stream_in(file("yelp_academic_dataset_review.json"))
#data_tip <- stream_in(file("yelp_academic_dataset_tip.json"))
#data_user <- stream_in(file("yelp_academic_dataset_user.json"))
```

The following sections descibres the process of reading, sampling and transforming the dataset.

Datasets

The whole Yelp dataset is composed of five data frames in which business, users, reviews, chechins and tips are registered.

Dataset of business has 15 variables for 61184 observations, many of those with categories and schedules of operation. The checkin dataset is composed of three other dataframes for 45166 observations. Tips dataset is described by texts for each customer, business and date and the likes given for each tip. The users dataset has features of 366715 customers and 11 features. Those features are the votes (funny, useful and cool) made by the customer, the count of reviews, names and list of friends.

This project focus on business reviews dataset. It has for each customer, business and date a text review which will turn our gold mine to be exploited. A summary of this dataset is presented as follows:

```
##     userid             reviewid              stars           date
##  Length:1569264     Length:1569264     Min.   :1.000   Length:1569264
##  Class :character   Class :character   1st Qu.:3.000   Class :character
##  Mode  :character   Mode  :character   Median :4.000   Mode  :character
##                                        Mean   :3.743
##                                        3rd Qu.:5.000
##                                        Max.   :5.000
##      text               type            businessid
##  Length:1569264     Length:1569264     Length:1569264
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##    votesfunny         votesuseful        votescool
##  Min.   :  0.0000   Min.   :  0.000   Min.   :  0.0000
##  1st Qu.:  0.0000   1st Qu.:  0.000   1st Qu.:  0.0000
##  Median :  0.0000   Median :  0.000   Median :  0.0000
##  Mean   :  0.4789   Mean   :  1.072   Mean   :  0.5942
##  3rd Qu.:  0.0000   3rd Qu.:  1.000   3rd Qu.:  1.0000
##  Max.   :141.0000   Max.   :166.000   Max.   :137.0000
```

Notice that some other covariates exists in the reviews dataset. This features will be used in the model. Features like votes.funny, votes.cool, votes.useful. These last features came in a form of data frame. This means that *jsonlite* tool read some structures that involves data frames inside another. To reduce this overload of data in memory, the function used to enhanced the reading is **flatten**.When flatten a data frame, the variables of their included data frames becomes part of the main and their names are concatenated with a prefix of their original dataframe name.

```
#flatten data frame to reduce memory and enhance performance
sDataReview <- flatten(sDataReview)
#cleaning variable names, removing all punctuation characters
colnames(sDataReview) <- gsub("[^A-Za-z]", "", colnames(sDataReview))
```

Summarizing stats of each dataset:

dataset	size_MB	total_obs	total_vars
Reviews	1393	1569264	8
Business	53	61184	15
Checkin	20	45166	3
Tips	94	495107	6
User	159	366715	11

Stratified sampling. . .

with a sample size of 5% over the 1569264 observations, a stratified sample (by class) is created in order to get an analysis and to get faster results, though. The sampling was **without replacement**.

Text mining tool

A powerful tool to process text in R is **tm**.With tm we can build a corpus in which a set of task can be applied. In linguistics, a corpus (plural corpora) is a large and structured set of texts in which statistical analysis and hypothesis testing can be done.

The following, is a list of some other tools loaded in the project in order to do the sentiment analysis: - tm.plugin.sentiment : score of lexicons tool - tm.lexicon.GeneralInquirer : Harvard open lexicons - SnowballC : stemming tool - slam : manipulating term document matrix objects - bigmemory : converting term document matrix objects into matrix - stringi : manipulating strings

Preprocessing

A corpus is loaded from the sample of data reviews previously extracted.

```r
#header of the corpus
m <- list(id = "id", content = "text")
myReader <- readTabular(mapping = m)
#loading corpus from random sample of reviews
corpus <- Corpus(DataframeSource(sDataReview), readerControl = list(reader = myReader))
```

Some functions are defined to get data cleaned from characters that are not in the english alphabet, i.e. just letters. Functions that catch english stop words, urls, emails, twitter tags, duplicated quotes, non-ascii characters, non english character, duplicated letters and duplicated words.

```r
### preprocessing functions #####
skipWords <- function(x) removeWords(x, stopwords("english"))
# removing URLs
urlPat<-function(x) gsub("(ftp|http)s?:(\\/)+[\\d\\w.]*\\b", " ", x, perl = T)
# removing Emails
emailRgx <- "[a-z0-9!#$%&'*+/=?^_`{|}~-]+(?:\\.[a-z0-9!#$%&'*+/=?^_`{|}~-]+)*@(?:[a-z0-9](?:[a-z0-9-]*[
emlPat<-function(x) gsub(emailRgx, " ", x)
# removing Twitter tags
tags<-function(x) gsub("(RT )|via", " ", x)
#replace curly brace with single quote
singleQuote <- function(x) gsub("\u2019|`", "'", x, perl = T)
#replace non printable, except ' - and space with empty string
nonprint <- function(x) gsub("[^\\p{L}\\s'-]", "", x, perl = T)
#remove non English words
nonEng <- function(x) gsub("[^A-Za-z']", " ", x)
#remove duplicated letters in words
dupLetters <- function(x) gsub("(\\w)\\1+", "\\1", x, perl = T)
#remove duplicated consecutive words
dupWords <- function(x) gsub("(\\w+)\\s+\\1", "\\1", x, perl = T)
```

At the end, a list of this functions is created and an additional tm function is append in order to stem words. The latter is done loading the Snowballc library.

```r
#list of functions to preprocess the corpus
funcs <- list(tolower, urlPat, emlPat, tags, singleQuote, nonprint, nonEng, dupLetters, dupWords, remov
#cleansing process is done
rcorpus <- tm_map(corpus, FUN = tm_reduce, tmFuns = funcs)
#fix a bug of tm library
rcorpus <- tm_map(rcorpus, PlainTextDocument)
```

Next, frequent words are identified.

The wordcloud for each rate

A document term matrix is created. This is also done with a tm function. As control parameters, the set of words to be extracted might be those no larger than 10 characters and with a minimum frequencie of total 50 counts.

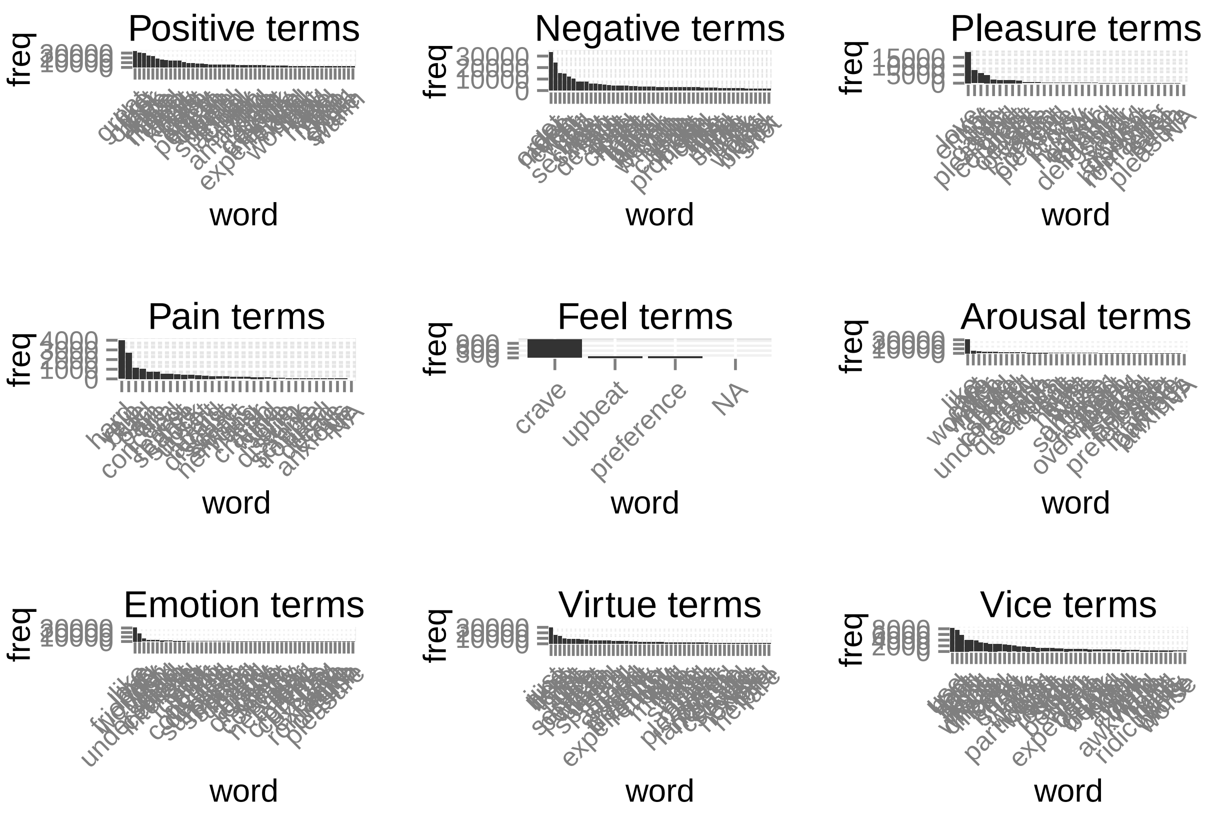
```
dtm <- DocumentTermMatrix(rcorpus, control = list(wordLengths = c(3,10), minDocFreq=50))
```

```
## Warning: Removed 16 rows containing missing values (position_stack).
```

```
## Warning: Removed 17 rows containing missing values (position_stack).
```

```
## Warning: Removed 47 rows containing missing values (position_stack).
```

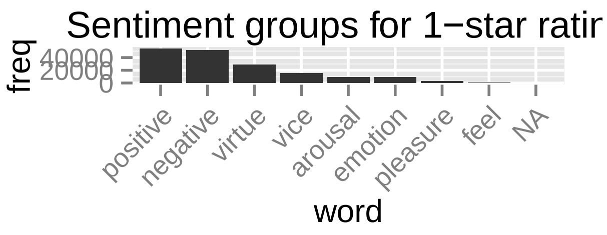
```
## Warning: Removed 11 rows containing missing values (position_stack).
```

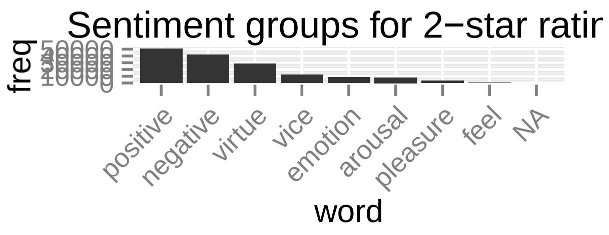


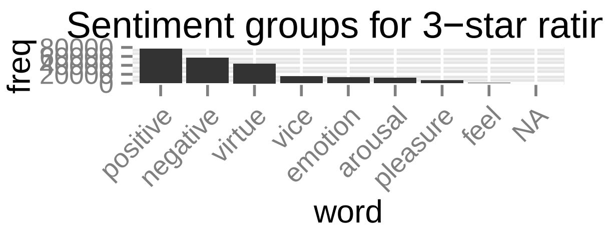
By star rating, the next plot shows the distribution of sentiment word groups. Notice that 5-star rating has more positive words thant the others. It decreases proportionaly with lower star rating.

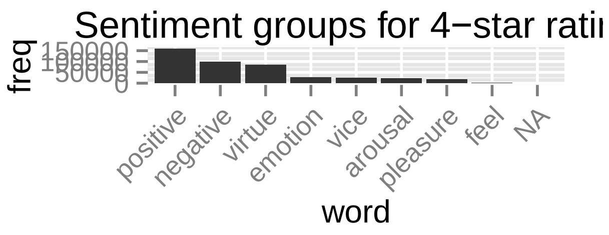
```
## Warning: Removed 42 rows containing missing values (position_stack).
```

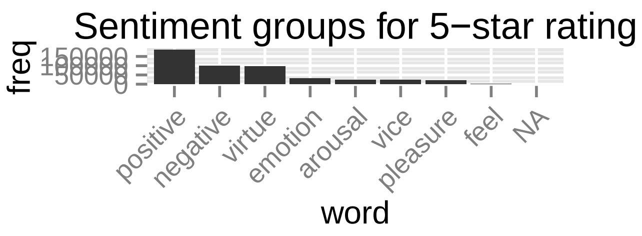
```
## Warning: Removed 42 rows containing missing values (position_stack).
## Warning: Removed 42 rows containing missing values (position_stack).

## Warning: Removed 42 rows containing missing values (position_stack).

## Warning: Removed 42 rows containing missing values (position_stack).
```











Also, notice that when star rating is increasing (from 1 to 5) the emotions group begins to scale up, i.e. the frequencies rises up instead of vice group. There's no insidences of pleasures and vice.

Building classifiers: how well a tool can predict a 5-star rating based on any review

A dataset is created merging all sentiment words matched above. From these, the most frequent (quantile above 85% of frequencies) are filtered. Finally, the new dataset is merged with the original variables from the reviews dataset, excluding text, ids and type covariates.

For modelling, we take advantage of caret package for the modelling process. It provides a set of functions in which the mining process is divided in cross-validation step, transformation, training algorithms and evaluation. The following sections descibres each step.

Splitting step

Because of the dataset size and the number of variables found (316!), some problems arouse like the curse of dimensionality. To face this problem, the use of parallel processing is imperative. For that, we use doSNOW library that is applied by the train function of the caret package (see parallel processing).

So, the dataset is divided into training and testing sets, holding a random sample of 75% of size for the training set.

```
##Cross Validation
set.seed(1234)
inTrain <- createDataPartition(y=data.fimp$stars, p=3/4, list=FALSE)
training = data[inTrain,]
testing = data[-inTrain,]
```

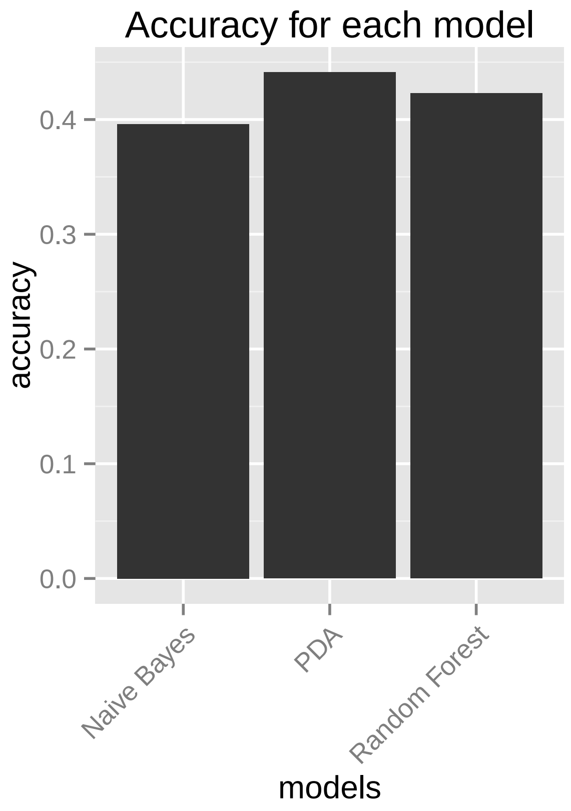
Some transformations before training

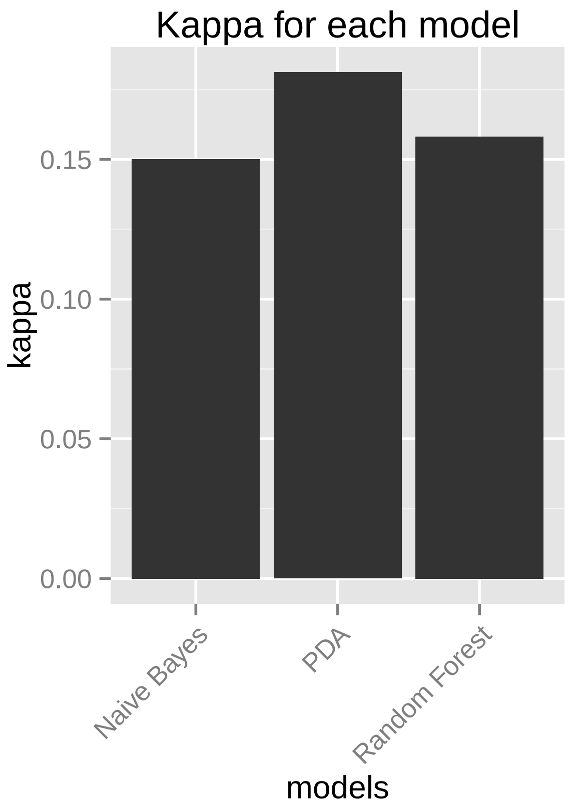
Before training a model, there is a main issue: curse of dimensionality. To face this problem, near zero variance and less important covariates must be removed. To do that, **klaR** and **FSelector** tools must be loaded in order to filter the less important variables: those that are colinear with others (correlated in a manner) and in which the gain of information over the classes (ratings-stars) is the least of all covariates.

Training classifiers

In order to get a competitive model, three types of techniques were compared: naive bayes, penalized discriminant analysis and random forest.

A summary of each trained model is the following presented:





It seems that the best model based on the accuracy and kappa measure over the same training set is the **Penalized Discriminant Analysis**. Next, each trained model will be evaluated with the testing set.

Evaluating trained models

Evaluation for Naive Bayes

```
##              logLoss              Mean_ROC              Accuracy
##           3.26816750            0.66657311            0.39815429
##                Kappa      Mean_Sensitivity      Mean_Specificity
##           0.15145090            0.31556627            0.82972587
##    Mean_Pos_Pred_Value   Mean_Neg_Pred_Value   Mean_Detection_Rate
##           0.32151481            0.83907842            0.07963086
## Mean_Balanced_Accuracy
##           0.57264607
```

Evaluation for Random forest

```
##              logLoss              Mean_ROC              Accuracy
##           2.66724967            0.66690300            0.42614592
##                Kappa      Mean_Sensitivity      Mean_Specificity
##           0.13097726            0.27839511            0.82492488
##    Mean_Pos_Pred_Value   Mean_Neg_Pred_Value   Mean_Detection_Rate
##           0.36366529            0.84628271            0.08522918
## Mean_Balanced_Accuracy
##           0.55165999
```

Evalution for Penalized Discriminant Analysis

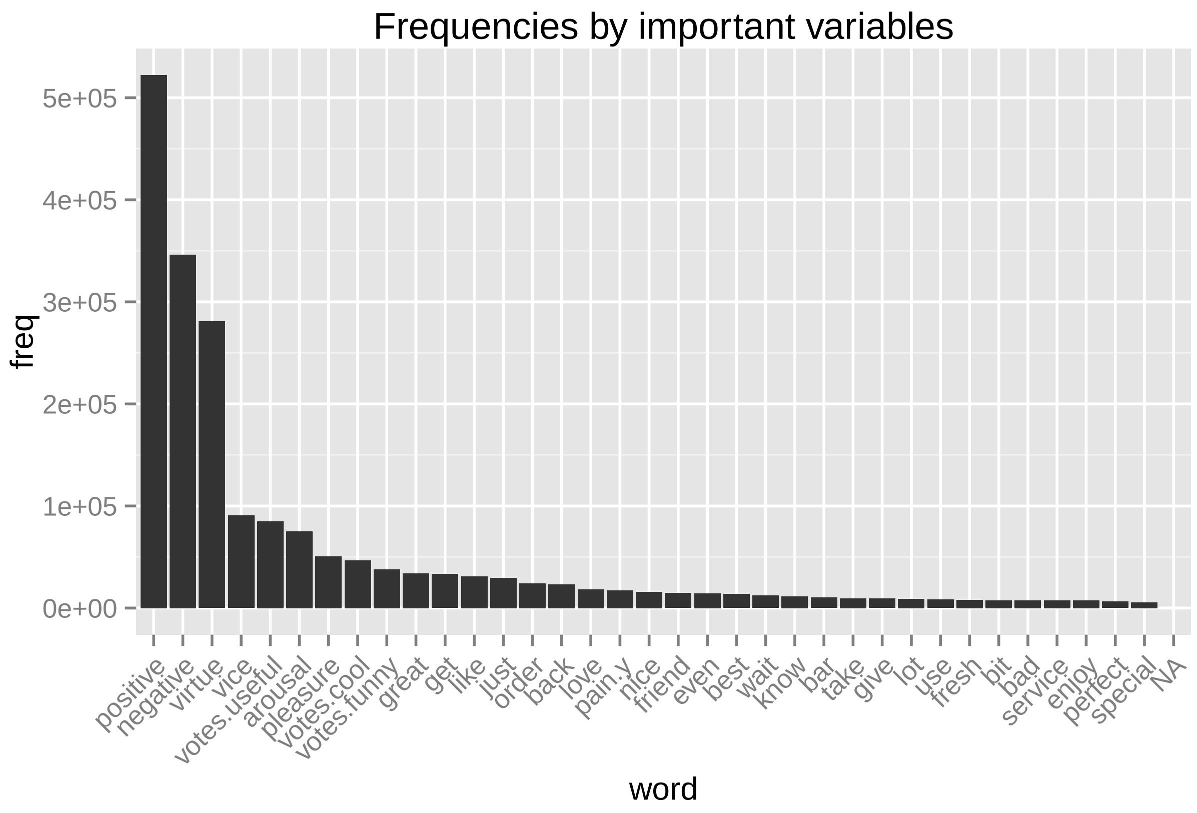
```
##              logLoss              Mean_ROC              Accuracy
##           1.28770296            0.72182841            0.44434814
##                Kappa      Mean_Sensitivity      Mean_Specificity
##           0.18566035            0.32074626            0.83601017
##    Mean_Pos_Pred_Value   Mean_Neg_Pred_Value   Mean_Detection_Rate
##           0.38392795            0.84915481            0.08886963
## Mean_Balanced_Accuracy
##           0.57837822
```

Results and discussion

The near zero variance reduction remove most of feelings and emotions. They had the lowest variance. The words with the highest gain of information with respect of the classes are plotted next:

```r
hs <- getWordHist(data.fimp[,-1], "Frequencies by important variables")
hs
```

```
## Warning: Removed 15 rows containing missing values (position_stack).
```



Notice that words like great, get, like, order, back, love, pain, nice and friend were more important than the others.

Definitely, the best model fit is the penalized discriminant analysis. These assertion is proved by the fact that the error behaves better rather in training set than in the test set, Despite the naive bayes model were trained with boostrapping with 25 bootstrap samples and 5 repetitions, it could not do better.

The best accuracy reached was 44% and with concordance level of 19%. These measures are average for all five classes of ratings.

These results demonstrate that issues like curse of dimensionality and a model for predicting a five-star rate based on text can be done. This is done thanks to reducing near zero variables and those with the lowest gain of information for all classes.

Also, the text mining tool helps in the way of getting a cleansed corpus and tokenized dataset. Parallel processing do not be missed considered.

Wrapping up

You can find the whole source code in my github: https://github.com/Okenfor/DSCapstone_YieldChallenge