Text mining project

mindu931

16 January 2019

Abstract

Wine sensory audition and assessment is never be an easy problem, even with a wine specilist. However, the description which is printed on each bottle can bring us some helpful information. Base on that ideal, I decided to analyse the data about wine review (take form Kaggle) and help the consumers make they decision on two different aspects. The first one is which is the good and excellent wine. The second aspect is which is the worth the price bottle (high value with a limit amount of money). I use different methords to process the text data, then when compare Naive bayse, SVm linear and SVM RBF kenel when building the best model for predicting. The final result is quite impressive with accurency aa%. Base on these model, some suggestions is provide for consumer to choose a satified bottle.

Introduction

Motivtion

I alway confuse when standing in a wine cellar. How can I choose a excelelnt wine for my family? What is the best wine I can choose with my budget? The problem can be solved easier if I had a good senses. But, sadly, I don't, like many people in this world. The problems still continue when I listen to the advices of the salesman and even taste some wine. So, I usually choose randomly one bottle based on these advices for my event. Unluckily, sometime my relatives don't think that it is a good wine. ## Aim I guess that the information from provider (description, country, region, designation...) should bring some important imformation. So, I try to solve the problems base on text mining ideal: Take the wine data which is already reviewd and marked by some specilists, then process these data together with the point and build a model to prediction in order to figure out which bottle is good or excellent, which one have high benefit or just medium. ## Content of this report In this report, Firstly, I will provided the information about the data as well as the way I define the class for each bottle. Then, I briefly summary relevent theory. The method I used will be presented in details. It include the steps for preporcees, compare and choosing the best model, my suggessions for choosing wine. Finnaly is some discussions and my conclusion.

Data

Wine review data is taked form Kaagle link. I choose the second .csv version of this dataset. In this version, duplicated data is removed. The data include 120975 samples about wine. Each sample contains the following information: country, description, designation, points, price, province, region, more specific region (region2), tester name, taster_twitter handle, title, variety and winery. Here is the overlook of the review:

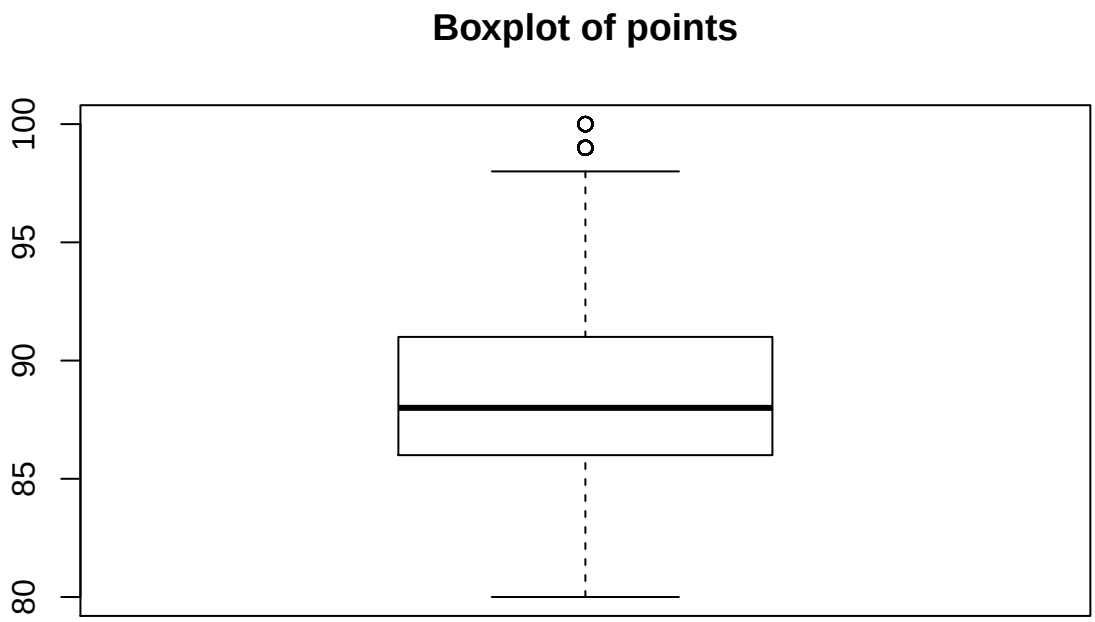
```
wine <- read.csv("C:/Users/Duong Minh Duc/Documents/GitHub/Text-Mining-Project/wine.csv")
head(wine)
```

```
##   X  country
## 1 0    Italy
```

```
## 2 1 Portugal
## 3 2       US
## 4 3       US
## 5 4       US
## 6 5    Spain
##
## 1                                                                                           Aromas in
## 2                                   This is ripe and fruity, a wine that is smooth while still struc
## 3                                                                                    Tart and snappy, the fl
## 4                                                                       Pineapple rind, lemon pith and orang
## 5             Much like the regular bottling from 2012, this comes across as rather rough and tannic
## 6 Blackberry and raspberry aromas show a typical Navarran whiff of green herbs and, in this case, ho
##                          designation points price          province
## 1                        VulkÃ  Bianco     87    NA Sicily & Sardinia
## 2                             Avidagos     87    15             Douro
## 3                                          87    14            Oregon
## 4                Reserve Late Harvest     87    13          Michigan
## 5 Vintner's Reserve Wild Child Block     87    65            Oregon
## 6                         Ars In Vitro     87    15    Northern Spain
##             region_1          region_2        taster_name
## 1               Etna                    Kerin Oâ\200\231Keefe
## 2                                                 Roger Voss
## 3   Willamette Valley Willamette Valley       Paul Gregutt
## 4 Lake Michigan Shore                     Alexander Peartree
## 5   Willamette Valley Willamette Valley       Paul Gregutt
## 6            Navarra                    Michael Schachner
##   taster_twitter_handle
## 1          @kerinokeefe
## 2            @vossroger
## 3          @paulgwineÂ
## 4
## 5          @paulgwineÂ
## 6           @wineschach
##                                                                                   title
## 1                                                     Nicosia 2013 VulkÃ  Bianco  (Etna)
## 2                                          Quinta dos Avidagos 2011 Avidagos Red (Douro)
## 3                                          Rainstorm 2013 Pinot Gris (Willamette Valley)
## 4                      St. Julian 2013 Reserve Late Harvest Riesling (Lake Michigan Shore)
## 5 Sweet Cheeks 2012 Vintner's Reserve Wild Child Block Pinot Noir (Willamette Valley)
## 6                                 Tandem 2011 Ars In Vitro Tempranillo-Merlot (Navarra)
##             variety               winery
## 1       White Blend              Nicosia
## 2    Portuguese Red Quinta dos Avidagos
## 3        Pinot Gris            Rainstorm
## 4          Riesling           St. Julian
## 5        Pinot Noir         Sweet Cheeks
## 6 Tempranillo-Merlot              Tandem
```

Points is an important information of this data. Points is marked by wine specilist, which make it realiable. The owner only public data which have at least 80 point (out of 100) which means the data contain only good and excellent wine. Here is the distribution of the point

```r
boxplot(wine$points, main="Boxplot of points")
```




```r
#bar plot
```

and some plot about other informations:

Theory

In this section I will brefly talk about some theory in nature leanguage processing(NLP) and machine learning(ML) in a simple way for reviewing. For some terms, only the aspect that relevent to this project is presented. I asump that the reader already have some basic knowleadge about NLP and ML before(this part is not a guide for a person with a blank background). I also noticed that we are processing the text writen in English.

-Terms in preprocess: + Stop words: This is the term for useless words in text. To be specific, these words usually have no (or very low) meaning but repreent a lot. For example: a, an, the... When processing text data. We usually remove these words. + Steamming: This is the process of reducing a word to it root from. For example: processs, processing and processed are 3 different words, but it just 3 different represent of the word "processs". An other example is evaluate and evaluation. + tokenization: the process of spliting text into smaller parts. Each part can considers as a features when tranning in machine learning kenels. If the smaller part here is the single word (split the text to single words) then all words we have will become the bag of words. + Corpus: You can understand simply that corpus is a collection of all features in NLP, used for trainning. For example, all words in the bag of words is a corpus. + Document term matrix(Dtm) is a matrix, which has column is all words of the corpus and column is the document ID. cross betwwen row

and column is the point for that word in that document. Points can be calculed in some different ways. In this project, I use tf-idf. + tf-idf: Stand for term frequency-inverse document frequency. This is a statistic method for calculating the importance of each word following this formular.... To be specific, If a word is represent many time in almost documents, it is not so importance, and vice versa.

- Machine leaning kenels: a kenel can understand as a core algorithim used when trainning a model in ML. Here is the three kenels that I use in this report:

+Naive Bayes: Let initial with we have a vector X of a document, which build in document term matrix. The probality of is vector belong to c class is p(c|X). From this, we have c = arg max p("good"|X) base on Bayes theory, we have c =... (5) Because p(X|c) is hard to calculate, we usually assump that X is independent. so ...

The asumption is not realitic, some how, it work well. +SVM: stand for support vector machine. In the previous kenel, we have vector X for each document. ++SVM linear...

++SVM RBF...

-Ngram:

-Pos of tag:

Method

In this part, I'll talk about the method for the first problem of this project: Which is the "good" and "excellent" wine. The second problem about wine that have "high" and "medium" value is solved using similar method but will be disscuss in details in a different part. - Define the class: The first step is define the class for each document. Base on the distribution of the point, I will take the point 88 as the bounary because it is the median value, which will make the data become balance. So, the wine that have higher than 88 will be the excellent wine, while the rest is the good wine. - Preprocess data + Split train and test: I use 80% of the data as the trainning set and 20% as the testing set. Then, the trainning data is process as follow: + Make the text ready: I decided that the information about country, designation, province, region and wine variety is important. So, I decided to merge that information to the description as a paragraph. That paragraph is the one that I will process. + Language convert: In the data set. there is some name of variety that are not unified (same type but written in different language). So I convert it to the most famous word as follow: ++ Replace german names with English names: "weissburgunder" is replaced as "chardonnay". "spatburgunder" is replaced as "pinot noir". "grauburgunder" is replaced as "pinot gris". ++ Replace the Spanish "garnacha"" with the french "grenache". ++ Replace the Italian "pinot nero"" with the french "pinot noir". ++ Replace the Portugues "alvarinho"" with the spanish "albarino".

- Remove non-ASCII characters.
- Remove punction.
- Make words to lower from.
- Remove number.
- Remove stop words: I use the default stop words lists in the `tm` library and then adjust it. Firstly, the lists have the word "very". In my opinion, very is a valuable word in this data. Because in the descriptions, for example, "sweet" and "very sweet" are different levels of flavor. So, I remove that word in the list. Then, I add three words "the", "and" and "wine" to the stop word list because in this data set, it don't have any meanning. Here is the final stop word list:

```
stopwords <- stopwords("english")
stopwords <- stopwords[!stopwords=="very"]
stopwords <- c("the", "and", "wine", stopwords)

stopwords
```

```
##   [1] "the"        "and"        "wine"       "i"          "me"
##   [6] "my"         "myself"     "we"         "our"        "ours"
##  [11] "ourselves"  "you"        "your"       "yours"      "yourself"
##  [16] "yourselves" "he"         "him"        "his"        "himself"
##  [21] "she"        "her"        "hers"       "herself"    "it"
##  [26] "its"        "itself"     "they"       "them"       "their"
##  [31] "theirs"     "themselves" "what"       "which"      "who"
##  [36] "whom"       "this"       "that"       "these"      "those"
##  [41] "am"         "is"         "are"        "was"        "were"
##  [46] "be"         "been"       "being"      "have"       "has"
##  [51] "had"        "having"     "do"         "does"       "did"
##  [56] "doing"      "would"      "should"     "could"      "ought"
##  [61] "i'm"        "you're"     "he's"       "she's"      "it's"
##  [66] "we're"      "they're"    "i've"       "you've"     "we've"
##  [71] "they've"    "i'd"        "you'd"      "he'd"       "she'd"
##  [76] "we'd"       "they'd"     "i'll"       "you'll"     "he'll"
##  [81] "she'll"     "we'll"      "they'll"    "isn't"      "aren't"
##  [86] "wasn't"     "weren't"    "hasn't"     "haven't"    "hadn't"
##  [91] "doesn't"    "don't"      "didn't"     "won't"      "wouldn't"
##  [96] "shan't"     "shouldn't"  "can't"      "cannot"     "couldn't"
## [101] "mustn't"    "let's"      "that's"     "who's"      "what's"
## [106] "here's"     "there's"    "when's"     "where's"    "why's"
## [111] "how's"      "a"          "an"         "the"        "and"
## [116] "but"        "if"         "or"         "because"    "as"
## [121] "until"      "while"      "of"         "at"         "by"
## [126] "for"        "with"       "about"      "against"    "between"
## [131] "into"       "through"    "during"     "before"     "after"
## [136] "above"      "below"      "to"         "from"       "up"
## [141] "down"       "in"         "out"        "on"         "off"
## [146] "over"       "under"      "again"      "further"    "then"
## [151] "once"       "here"       "there"      "when"       "where"
## [156] "why"        "how"        "all"        "any"        "both"
## [161] "each"       "few"        "more"       "most"       "other"
## [166] "some"       "such"       "no"         "nor"        "not"
## [171] "only"       "own"        "same"       "so"         "than"
## [176] "too"
```

+Steamming: I steam the word by using SnowballC steam.
+Tonkenzine and making bag of words as the corpus. Then I only keep 99% spare term. Which mean that only the words that appear in at lest 1% of documents is keeped. It make sense because there are some rare words which only appear in one or a few documents. That words are not helpful for training because maybe we never meet it again in the testing set. +Create the document terms matrix base on tf-idf. +Now the data is ready for training. For testing data, the preprocess is similar, except the Dtm. Dtm of testing data will be create base on the corpus of trainning data (which means some words that don't appear in the tranning data will be drop)

-Compare kenels: At this step, I'll use the ready Dtm for trainning model, then test with the testing Dtm. There kenel Naive-bayse, SVM linear and SVM RBF are used to compare. In this step, because the limit in my resouse (Laptop core i5 7th gen, 8GB) and time. I only use the sample 20% of the data (16% - 19356 samples for training and 4% - 4839 sample for tesing) to compute. Different parameter used for training is also reported. Notice that the time for training NB and SVM linear is quite fast (a few minutes) but It take a long time for training SVM RBF (~14 hours, for 505 terms in Dtm inlcude time for tunning) - Improve the best model: After comparing, the best model is selected. Then I continue improve the model when using n-gram and adjust weight for some terms using Part-Of-Speech Tagger (POS Tagger). - Final result: Final model is selected and run again with 100% data. The final result is reported based on this model.

Result and explain

- Firstly, Let't take a look at 5 original paragrahps and the paragraphs after being preprocess to see how it work:

```
# Use for check the not available and duplicate data, but not necessary
#wine <- na.omit(wine)
#wine[duplicated(wine),]
wine$quality <- wine$points > 88
wine$quality[wine$quality == TRUE] <- "excellent"
wine$quality[wine$quality == FALSE] <- "good"
wine$quality <- as.factor(wine$quality)
wine$value <- wine$points/log(wine$price)
wine$benefit <- wine$value > 27
wine$benefit[wine$benefit == TRUE] <- "high"
wine$benefit[wine$benefit == FALSE] <- "medium"
wine$benefit <- as.factor(wine$benefit)
wine$description <- paste(wine$description, wine$country, wine$designation, wine$province, wine$region_
wine$description <- as.character(wine$description)

wine$description <- gsub("weissburgunder", "chardonnay", wine$description)
wine$description <- gsub("spatburgunder", "pinot noir", wine$description)
wine$description <- gsub("grauburgunder", "pinot gris", wine$description)

#Replace the Spanish garnacha with the french grenache
wine$description <- gsub("garnacha", "grenache", wine$description)

#Replace the Italian pinot nero with the french pinot noir
wine$description <- gsub("pinot nero", "pinot noir", wine$description)

#Replace the Portugues alvarinho with the spanish albarino
wine$description <- gsub("alvarinho", "albarino", wine$description)

#clean function
wine$description <- iconv(wine$description, from = "UTF-8", to = "ASCII", sub = "")
```

```
n = dim(wine)[1]
set.seed(12345)
id2 = sample(1:n, floor(n*0.2))
wine_sample <- wine[id2,]
wine2 <- wine[id2,]
n2 = length(id2)
id_test = sample(1:n2, floor(n2*0.8))
train = wine_sample[id_test,]
test = wine_sample[-id_test,]

##clean function
clean <- function(text_vector)
  {
    wine_corpus = VCorpus(VectorSource(text_vector))
    wine_corpus = tm_map(wine_corpus, removePunctuation)
    wine_corpus = tm_map(wine_corpus, content_transformer(tolower))
    wine_corpus = tm_map(wine_corpus, removeNumbers)
    wine_corpus = tm_map(wine_corpus, removeWords, stopwords )
```

```
    #wine_corpus = tm_map(wine_corpus, stripWhitespace)
    wine_corpus <- tm_map(wine_corpus, stemDocument)


    return(wine_corpus)
  }

##create the train set
wine_train_set <- clean(train$description)

train$description[1:5]
```

```
## [1] "Fresh green herbs on the nose accumulate fruit and floral notes on the palate of this pretty bu
## [2] "White peach and flower petal notes on the nose gain rich honeydew and fresh herb flavors on the
## [3] "This whopper of a wine sees 24 months in wood and is a powerhouse blend of Cabernet, Merlot and
## [4] "Featuring and named after a painting by renowned Santa Barbara artist James Jarvaise, this intri
## [5] "Leather meets plum sauce meets red licorice in this full-bodied, robustly crafted wine that's pl
```

```
wine_train_set[1:5]
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 5
```

- Then, Here is the Dtm for the first sentence of training set:

```
train_dtm_tfidf <- DocumentTermMatrix(wine_train_set, control = list(weighting = weightTfIdf))
#train_dtm_tfidf <- DocumentTermMatrix(wine_train_set, control = list( tokenize=NLP_tokenizer))
#train_dtm_tfidf <- DocumentTermMatrix(wine_train_set)
train_dtm_tfidf <- removeSparseTerms(train_dtm_tfidf, 0.99)

#wine_train_set <- cbind(wine_train_set, train$quality)

#create the test set
wine_test_set <- clean(test$description)
wine_test_set <- DocumentTermMatrix(wine_test_set, control = list(dictionary = Terms(train_dtm_tfidf) ,
#wine_test_set <- DocumentTermMatrix(wine_test_set, control = list(dictionary = Terms(train_dtm_tfidf)

#create matrix for training
wine_train_set <<- as.matrix(train_dtm_tfidf)
wine_test_set <- as.matrix(wine_test_set)
wine_test_set <- wine_test_set[,Terms(train_dtm_tfidf)]
#create the test result
wine_testing_result <- test$quality

#tagPos
tagPOS <-  function(x, ...) {
  s <- as.String(x)
  word_token_annotator <- Maxent_Word_Token_Annotator()
  a2 <- Annotation(1L, "sentence", 1L, nchar(s))
  a2 <- annotate(s, word_token_annotator, a2)
  a3 <- annotate(s, Maxent_POS_Tag_Annotator(), a2)
```

```
  a3w <- a3[a3$type == "word"]
  POStags <- unlist(lapply(a3w$features, `[[`, "POS"))
  POStagged <- paste(sprintf("%s/%s", s[a3w], POStags), collapse = " ")
  list(POStagged = POStagged, POStags = POStags)
}

#extract nouns and adj
tag <- tagPOS(Terms(train_dtm_tfidf))
tag <- tag$POStags
noun_id <- which( tag=="NN")
nouns <- colnames(wine_train_set)[noun_id]
adj_id <- which( tag=="JJ")
adj <- colnames(wine_train_set)[adj_id]



train_dtm_tfidf[1,]
```

```
## <<DocumentTermMatrix (documents: 1, terms: 505)>>
## Non-/sparse entries: 28/477
## Sparsity           : 94%
## Maximal term length: 15
## Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
```

And let see the the first line of Dtm for tesing set see make sure that the prepared data is corrrect. As we can see, the term is remain. Just the points is different.

```
wine_test_set[1,]
```

```
##          accent            acid             add          africa
##      0.00000000      0.00000000      0.00000000      0.00000000
##       aftertast             age         alcohol          almond
##      0.00000000      0.15791012      0.00000000      0.00000000
##          almost           along        alongsid         alreadi
##      0.00000000      0.00000000      0.00000000      0.00000000
##           alsac            also        although            ampl
##      0.00000000      0.00000000      0.00000000      0.00000000
##            anis          appeal            appl        approach
##      0.00000000      0.00000000      0.00000000      0.00000000
##         apricot       argentina           aroma          aromat
##      0.00000000      0.00000000      0.00000000      0.00000000
##          around         astring         attract       australia
##      0.00000000      0.28850553      0.00000000      0.00000000
##         austria            back            bake          balanc
##      0.00000000      0.00000000      0.00000000      0.00000000
##          barolo          barrel            bean          beauti
##      0.00000000      0.00000000      0.00000000      0.00000000
##           berri            best          better             big
##      0.00000000      0.00000000      0.00000000      0.00000000
##             bit          bitter           black      blackberri
##      0.00000000      0.00000000      0.00000000      0.00000000
##           blanc           blend         blossom            blue
```

```
##      0.00000000      0.00000000      0.00000000      0.00000000
##       blueberri           boast            bodi            bold
##      0.00000000      0.00000000      0.00000000      0.00000000
##        bordeaux    bordeauxstyl           bottl         bouquet
##      0.00000000      0.00000000      0.00000000      0.00000000
##          bright           bring           brisk         brunello
##      0.00000000      0.00000000      0.00000000      0.00000000
##           brut         burgundi          butter             cab
##      0.00000000      0.00000000      0.00000000      0.00000000
##         cabernet      california             can           candi
##      0.00000000      0.00000000      0.00000000      0.00000000
##         caramel           carri           cassi           cedar
##      0.00000000      0.00000000      0.00000000      0.00000000
##          cellar         central         champagn            char
##      0.00000000      0.00000000      0.00000000      0.00000000
##         charact      chardonnay          cherri           chewi
##      0.00000000      0.00000000      0.00000000      0.00000000
##         chianti           chile           chocol          chunki
##      0.00000000      0.00000000      0.00000000      0.00000000
##         cinnamon          citrus         citrusi         classic
##      0.00000000      0.00000000      0.00000000      0.00000000
##         classico           clean           close           clove
##      0.00000000      0.00000000      0.00000000      0.00000000
##           coast           cocoa           coffe            cola
##      0.00000000      0.00000000      0.00000000      0.00000000
##           color         columbia          combin            come
##      0.00000000      0.00000000      0.00000000      0.00000000
##         complex         concentr            cool            core
##      0.00000000      0.18776852      0.00000000      0.00000000
##          counti       cranberri           cream          creami
##      0.00000000      0.00000000      0.00000000      0.00000000
##           creek           crisp             cru           crush
##      0.00000000      0.00000000      0.00000000      0.00000000
##            ctes         currant             cut            cuve
##      0.00000000      0.00000000      0.00000000      0.00000000
##            dark            deep             del           delic
##      0.00000000      0.00000000      0.00000000      0.00000000
##          delici           deliv            dens           depth
##      0.00000000      0.00000000      0.00000000      0.00000000
##          despit         develop          doesnt           domin
##      0.00000000      0.00000000      0.00000000      0.00000000
##             dri           drink           dusti           earth
##      0.00000000      0.09459364      0.00000000      0.00000000
##          earthi            easi             edg            eleg
##      0.00000000      0.00000000      0.00000000      0.00000000
##         element             end           enjoy          enough
##      0.00000000      0.00000000      0.00000000      0.00000000
##         espresso           estat            even           excel
##      0.00000000      0.00000000      0.00000000      0.00000000
##            exot         express           extra         extract
##      0.00000000      0.00000000      0.00000000      0.00000000
##            fair          famili          featur            feel
##      0.00000000      0.00000000      0.00000000      0.00000000
##         ferment            fill            find            fine
```

```
##      0.00000000      0.00000000      0.00000000      0.00000000
##          finger          finish           firm           first
##      0.00000000      0.00000000      0.00000000      0.00000000
##          flavor          fleshi          floral          flower
##      0.00000000      0.00000000      0.00000000      0.00000000
##           focus          follow            food         foothil
##      0.00000000      0.00000000      0.00000000      0.00000000
##          forest         forward        fragrant           frame
##      0.00000000      0.00000000      0.00000000      0.00000000
##           franc          french           fresh           front
##      0.00000000      0.00000000      0.00000000      0.00000000
##           fruit          fruiti            full        fullbodi
##      0.00000000      0.00000000      0.00000000      0.00000000
##         generous           gentl         germani             get
##      0.00000000      0.00000000      0.00000000      0.00000000
##            give           glass            good           grand
##      0.00000000      0.00000000      0.00000000      0.00000000
##           grape      grapefruit         graphit           great
##      0.00000000      0.00000000      0.00000000      0.00000000
##           green         grenach          grigio           grill
##      0.00000000      0.00000000      0.00000000      0.00000000
##            grip            gris           grner           grown
##      0.00000000      0.00000000      0.00000000      0.00000000
##            hard           heavi            herb          herbal
##      0.00000000      0.00000000      0.00000000      0.00000000
##            here            high            hill            hint
##      0.00000000      0.00000000      0.00000000      0.00000000
##            hold           honey         impress          includ
##      0.00000000      0.00000000      0.00000000      0.00000000
##          integr          intens        interest         intrigu
##      0.00000000      0.20959543      0.00000000      0.00000000
##           itali             jam           jammi            juic
##      0.13174236      0.00000000      0.00000000      0.00000000
##           juici            just            keep            lack
##      0.00000000      0.00000000      0.00000000      0.00000000
##            lake            last           layer            lead
##      0.00000000      0.00000000      0.00000000      0.00000000
##            leaf            lean         leather            leav
##      0.00000000      0.00000000      0.00000000      0.00000000
##           lemon          lemoni            lend          length
##      0.00000000      0.00000000      0.00000000      0.00000000
##             les         licoric            lift           light
##      0.00000000      0.00000000      0.00000000      0.00000000
##            like            lime          linger           littl
##      0.00000000      0.00000000      0.00000000      0.00000000
##            live            load            loir            long
##      0.00000000      0.00000000      0.00000000      0.20931672
##             lot            love            lush            made
##      0.00000000      0.00000000      0.00000000      0.00000000
##            make          malbec           mango            mani
##      0.00000000      0.00000000      0.00000000      0.00000000
##            mark           matur            meat          medium
##      0.00000000      0.00000000      0.00000000      0.00000000
##      mediumbodi           melon         mendoza          merlot
```

```
##      0.00000000      0.00000000      0.00000000      0.00000000
##          midpal            mild           miner            mint
##      0.00000000      0.00000000      0.00000000      0.00000000
##             mix           mocha           moder      montalcino
##      0.00000000      0.00000000      0.00000000      0.00000000
##           month        mountain           mouth       mouthfeel
##      0.00000000      0.00000000      0.00000000      0.00000000
##           much            napa           natur        nebbiolo
##      0.00000000      0.00000000      0.00000000      0.00000000
##         nectarin            need            new            next
##      0.00000000      0.00000000      0.19067469      0.00000000
##           nice            noir           north    northeastern
##      0.00000000      0.00000000      0.00000000      0.00000000
##         northern            nose            note             now
##      0.00000000      0.00000000      0.00000000      0.13745893
##           nuanc             nut             oak            oaki
##      0.00000000      0.00000000      0.14252758      0.00000000
##           offer             old            oliv             one
##      0.00000000      0.00000000      0.00000000      0.00000000
##            open            opul           orang          oregon
##      0.00000000      0.00000000      0.00000000      0.00000000
##          overal            pack            pair           palat
##      0.00000000      0.00000000      0.00000000      0.00000000
##            paso           peach            pear            peel
##      0.00000000      0.00000000      0.00000000      0.00000000
##          pepper         pepperi         perfect          perfum
##      0.00000000      0.00000000      0.00000000      0.00000000
##         persist           petit             pie        piedmont
##      0.00000000      0.00000000      0.00000000      0.00000000
##         pineappl            pink           pinot            play
##      0.00000000      0.00000000      0.00000000      0.00000000
##         pleasant           plenti            plum           plump
##      0.00000000      0.00000000      0.00000000      0.00000000
##           plush          polish          portug       portugues
##      0.00000000      0.00000000      0.00000000      0.00000000
##          potenti           power         premier          pretti
##      0.00000000      0.00000000      0.00000000      0.00000000
##           price          produc         provenc          provid
##      0.00000000      0.00000000      0.00000000      0.00000000
##          provinc           prune            pure         qualiti
##      0.00000000      0.00000000      0.00000000      0.00000000
##            quit            raci          raisin       raspberri
##      0.00000000      0.00000000      0.00000000      0.00000000
##          rather             raw           readi             red
##      0.00000000      0.00000000      0.22555658      0.08862907
##         refresh          region          remain          reserv
##      0.00000000      0.30103098      0.00000000      0.00000000
##         reserva          reveal            rhne        rhnestyl
##      0.00000000      0.00000000      0.00000000      0.00000000
##            rich           riesl           right           rioja
##      0.12218768      0.00000000      0.00000000      0.00000000
##            ripe         riserva           river           roast
##      0.00000000      0.00000000      0.00000000      0.00000000
##            robl             ros            rose           round
```

```
##      0.00000000      0.00000000      0.00000000      0.00000000
##         russian          rustic            sage       sangioves
##      0.00000000      0.00000000      0.00000000      0.00000000
##           santa        sardinia       sauvignon          savori
##      0.00000000      0.00000000      0.00000000      0.00000000
##           scent            seem          select            sens
##      0.00000000      0.00000000      0.00000000      0.00000000
##              set           sharp            show          sicili
##      0.00000000      0.00000000      0.00000000      0.00000000
##             side          sierra           silki           simpl
##      0.00000000      0.00000000      0.00000000      0.00000000
##              sip           sirah            skin          slight
##      0.00000000      0.00000000      0.00000000      0.00000000
##            smell           smoke           smoki          smooth
##      0.00000000      0.00000000      0.00000000      0.00000000
##             soft          soften            soil           solid
##      0.00000000      0.00000000      0.00000000      0.00000000
##         somewhat          sonoma            soon            sour
##      0.00000000      0.00000000      0.00000000      0.00000000
##            sourc           south        southern       southwest
##      0.00000000      0.00000000      0.00000000      0.00000000
##            spain          sparkl           spice           spici
##      0.00000000      0.00000000      0.00000000      0.00000000
##            start           still           stone straightforward
##      0.00000000      0.21821272      0.00000000      0.00000000
##       strawberri          streak          strong        structur
##      0.00000000      0.00000000      0.00000000      0.00000000
##            style           subtl           sugar         suggest
##      0.00000000      0.00000000      0.00000000      0.00000000
##         superior           suppl         support           sweet
##      0.00000000      0.00000000      0.00000000      0.00000000
##            syrah            take         tangerin           tangi
##      0.00000000      0.27212675      0.00000000      0.00000000
##           tannic          tannin            tart            tast
##      0.00000000      0.09538360      0.00000000      0.00000000
##              tea     tempranillo          textur            that
##      0.00000000      0.00000000      0.00000000      0.00000000
##            there           thick          though           tight
##      0.00000000      0.00000000      0.00000000      0.20987526
##             time           toast          toasti         tobacco
##      0.00000000      0.00000000      0.00000000      0.00000000
##           togeth          tomato            tone             top
##      0.00000000      0.00000000      0.00000000      0.00000000
##          toscana           touch          tropic            turn
##      0.00000000      0.00000000      0.00000000      0.00000000
##          tuscani             two      underbrush          valley
##      0.00000000      0.00000000      0.00000000      0.00000000
##          vanilla          variet         varieti         veltlin
##      0.00000000      0.00000000      0.00000000      0.00000000
##          velveti          veneto          verdot            veri
##      0.00000000      0.26183469      0.00000000      0.00000000
##          vibrant            vine        vineyard          vintag
##      0.00000000      0.00000000      0.00000000      0.00000000
##         viognier          violet           walla            warm
```

```
##      0.00000000      0.00000000      0.00000000      0.00000000
##      washington             way          weight            well
##      0.00000000      0.00000000      0.00000000      0.00000000
##             wet           whiff           white            wild
##      0.00000000      0.00000000      0.00000000      0.00000000
##            will       willamett            wine         winemak
##      0.19004151      0.00000000      0.22178657      0.00000000
##          wineri         without            wood            wrap
##      0.00000000      0.00000000      0.00000000      0.00000000
##            year          yellow             yet            york
##      0.00000000      0.00000000      0.00000000      0.00000000
##           young         zealand            zest           zesti
##      0.00000000      0.00000000      0.00000000      0.00000000
##       zinfandel
##      0.00000000
```

Now, the data is ready. I train that dataset with threes different kenels as mentioned. And here is the results:

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.5.2
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:NLP':
##
##     annotate
```

```
train_nb_model <- train(x= wine_train_set, y=train$quality , method = 'naive_bayes')
train_nb_model
```

```
## Naive Bayes
##
## 20795 samples
##   505 predictor
##     2 classes: 'excellent', 'good'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 20795, 20795, 20795, 20795, 20795, 20795, ...
## Resampling results across tuning parameters:
##
##   usekernel  Accuracy   Kappa
##   FALSE      0.7394611  0.4799649
##    TRUE      0.5538528  0.1479704
```

```
##
## Tuning parameter 'laplace' was held constant at a value of 0
##
## Tuning parameter 'adjust' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were laplace = 0, usekernel =
##  FALSE and adjust = 1.
```

```
model_nb_result <- predict(train_nb_model, newdata = wine_test_set)

conf_nb_train <- table(model_nb_result, wine_testing_result)
names(dimnames(conf_nb_train)) <- c("Predicted class", "Actual class")
confusionMatrix(conf_nb_train)
```

```
## Confusion Matrix and Statistics
##
##                Actual class
## Predicted class excellent good
##       excellent      1608  561
##       good            832 2198
##
##                Accuracy : 0.7321
##                  95% CI : (0.7198, 0.7441)
##     No Information Rate : 0.5307
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4586
##  Mcnemar's Test P-Value : 4.684e-13
##
##             Sensitivity : 0.6590
##             Specificity : 0.7967
##          Pos Pred Value : 0.7414
##          Neg Pred Value : 0.7254
##              Prevalence : 0.4693
##          Detection Rate : 0.3093
##    Detection Prevalence : 0.4172
##       Balanced Accuracy : 0.7278
##
##        'Positive' Class : excellent
##
```

```
#Here is the SVM Linear kenel
train_svmLinear_model <- train(x= wine_train_set, y=train$quality , method = 'svmLinear3')
train_svmLinear_model
```

```
## L2 Regularized Support Vector Machine (dual) with Linear Kernel
##
## 20795 samples
##   505 predictor
##     2 classes: 'excellent', 'good'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
```

```
## Summary of sample sizes: 20795, 20795, 20795, 20795, 20795, 20795, ...
## Resampling results across tuning parameters:
##
##   cost  Loss  Accuracy   Kappa
##   0.25  L1    0.7769075  0.5523296
##   0.25  L2    0.7799786  0.5585181
##   0.50  L1    0.7774427  0.5535199
##   0.50  L2    0.7800307  0.5586457
##   1.00  L1    0.7774043  0.5534831
##   1.00  L2    0.7795170  0.5576407
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were cost = 0.5 and Loss = L2.
```

```
model_svmLinear_result <- predict(train_svmLinear_model, newdata = wine_test_set)

conf_svmLinear_train <- table(model_svmLinear_result, wine_testing_result)
names(dimnames(conf_svmLinear_train)) <- c("Predicted class", "Actual class")
confusionMatrix(conf_svmLinear_train)
```

```
## Confusion Matrix and Statistics
##
##                Actual class
## Predicted class excellent good
##       excellent      1717  444
##       good            723 2315
##
##                Accuracy : 0.7755
##                  95% CI : (0.7639, 0.7868)
##     No Information Rate : 0.5307
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5464
##  Mcnemar's Test P-Value : 4.024e-16
##
##             Sensitivity : 0.7037
##             Specificity : 0.8391
##          Pos Pred Value : 0.7945
##          Neg Pred Value : 0.7620
##              Prevalence : 0.4693
##          Detection Rate : 0.3303
##    Detection Prevalence : 0.4157
##       Balanced Accuracy : 0.7714
##
##        'Positive' Class : excellent
##
```

```
# #Here is the SVM RBF kenel
# # Because the trainning time is long. I'll not run this code and only attach images of the previous r
# train_svmRBF_model <- train(x= wine_train_set, y=train$quality , method = 'svmRadial')
# train_svmRBF_model
# model_svmRBF_result <- predict(train_svmRBF_model, newdata = wine_test_set)
#
```

```
> train_svmRBF_model
Support Vector Machines with Radial Basis Function Kernel

19356 samples
  505 predictor
    2 classes: 'excellent', 'good'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 19356, 19356, 19356, 19356, 19356, 19356, ...
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.25  0.8033515  0.6050840
  0.50  0.8072213  0.6130414
  1.00  0.8111750  0.6211173

Tuning parameter 'sigma' was held constant at a value of 0.001066893
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.001066893 and C = 1.
```

Figure 1: SVMRBF kenel

```
# conf_svmRBF_train <- table(model_svmRBF_result, wine_testing_result)
# names(dimnames(conf_svmRBF_train)) <- c("Predicted class", "Actual class")
# confusionMatrix(conf_svmRBF_train)
```

As we can see in the result, The accurancy when training data for NB, SVM Linear and SVM RBF is aa%, bb% and 81.11%. As a results, I expected that the accurancy when testing with test data will be similar. But it just 74.68% for NB, 78.59% for SVM Linear and 71.36%. SVM Linear is the kenel wich have higest value of accurancy. No information rate (NIR) 0.5323 mean that a class take 53.23%(Good class),which mean the data is balanced. We can judge that the model is actually worked. But, accurecy is just one side of the story. Let see about the classification: We have two classes "good" and "excellent". If the class is predicted exacttly, it's perfect. Obviously, excellent wine is better than good wine. So, if a good wine is predicted as excellent wine, it's hard to eccept (Similar with False negative). On contrast, If you pretend to buy a good wine but have the excellent wine. You are just lucky and nothing happend. (similar with False positive) SVM RBF kenel show the lowest number of False negative. But, I also see that the predicted result is biased to the "good class": The number of good wine is predicted is triple as the number of excellent wine. It's quite hard for understand. On the other hand, SVM Linear is better than NB in all indicators.

Afterall, I consider the accurancy, the number of confusion matrix and the traning time for choosing the best model. In my opinion, It's SVM Linear. (SVM RBF is interesting but it's hard when I try improve the model with that high trainning time, consider the scope of this project)

-Improve the model: I'll try to improve the SVM linear model with the folling methods: + Improve with n-gram: Firtsly, I try to train the model with 2-gram, (the preprocess still remain) here is the first line of Dtm:

```
NLP_tokenizer <- function(x) {
  unlist(lapply(ngrams(words(x), 2:2), paste, collapse = "_"), use.names = FALSE)
}

wine_train_set <- clean(train$description)

train_dtm_tfidf <- DocumentTermMatrix(wine_train_set, control = list(weighting = weightTfIdf, tokenize=
#train_dtm_tfidf <- DocumentTermMatrix(wine_train_set, control = list( tokenize=NLP_tokenizer))
```

```
> confusionMatrix(conf_train)
Confusion Matrix and Statistics

               Actual class
Predicted class excellent good
      excellent       962   85
      good           1301 2491

               Accuracy : 0.7136
                 95% CI : (0.7006, 0.7263)
    No Information Rate : 0.5323
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4053
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4251
            Specificity : 0.9670
         Pos Pred Value : 0.9188
         Neg Pred Value : 0.6569
             Prevalence : 0.4677
         Detection Rate : 0.1988
   Detection Prevalence : 0.2164
      Balanced Accuracy : 0.6961

       'Positive' Class : excellent
```

Figure 2: SVMRBF result

```
#train_dtm_tfidf <- DocumentTermMatrix(wine_train_set)
train_dtm_tfidf <- removeSparseTerms(train_dtm_tfidf, 0.99)


NLP_tokenizer <- function(x) {
  unlist(lapply(ngrams(words(x), 2:2), paste, collapse = "_"), use.names = FALSE)
}

#create the test set
wine_test_set <- clean(test$description)
wine_test_set <- DocumentTermMatrix(wine_test_set, control = list(dictionary = Terms(train_dtm_tfidf) ,
```

```
## Warning in weighting(x): empty document(s): 11 41 44 48 51 56 99 101 117
## 119 128 142 147 160 167 168 172 195 208 219 220 238 241 243 249 262 291 310
## 320 344 359 413 416 450 500 522 532 567 585 604 630 648 680 694 698 699 730
## 732 733 744 748 769 778 794 801 840 867 883 910 920 927 980 1002 1009 1011
## 1013 1051 1061 1087 1097 1104 1115 1118 1128 1129 1133 1137 1141 1164 1169
## 1187 1201 1226 1233 1243 1245 1255 1292 1336 1337 1342 1359 1367 1369 1374
## 1393 1408 1413 1443 1499 1525 1543 1549 1554 1578 1601 1626 1635 1654 1701
## 1735 1736 1765 1782 1787 1796 1838 1904 1906 1990 2017 2019 2036 2051 2070
## 2072 2078 2102 2107 2123 2141 2148 2169 2177 2188 2197 2201 2212 2218 2229
## 2248 2270 2271 2296 2311 2314 2317 2332 2353 2380 2382 2414 2438 2439 2459
## 2506 2529 2575 2586 2629 2632 2648 2650 2678 2710 2727 2751 2762 2806 2812
## 2824 2846 2851 2858 2895 2911 2917 2935 2947 2951 2965 2983 2984 2986 3018
## 3019 3043 3080 3113 3126 3153 3157 3163 3205 3222 3228 3256 3259 3286 3287
## 3296 3306 3307 3314 3324 3357 3377 3385 3387 3391 3399 3420 3435 3439 3445
```

```
## 3480 3484 3498 3509 3511 3523 3526 3544 3546 3549 3562 3613 3614 3616 3622
## 3628 3634 3650 3675 3705 3724 3734 3748 3754 3802 3803 3828 3842 3845 3884
## 3919 3941 3956 3963 3984 4016 4028 4031 4047 4054 4055 4065 4071 4074 4085
## 4090 4106 4118 4136 4138 4146 4152 4198 4228 4234 4246 4250 4271 4283 4287
## 4296 4299 4311 4316 4335 4337 4351 4372 4377 4387 4421 4450 4453 4459 4467
## 4490 4491 4497 4520 4538 4551 4556 4564 4577 4588 4603 4604 4610 4630 4654
## 4679 4729 4731 4753 4763 4765 4771 4800 4843 4868 4877 4945 4949 4956 4991
## 4994 4996 4999 5011 5023 5029 5034 5054 5066 5071 5091 5128 5136 5170 5183
```

```
#wine_test_set <- DocumentTermMatrix(wine_test_set, control = list(dictionary = Terms(train_dtm_tfidf)

#create matrix for training
wine_train_set <<- as.matrix(train_dtm_tfidf)
wine_test_set <- as.matrix(wine_test_set)
wine_test_set <- wine_test_set[,Terms(train_dtm_tfidf)]
#create the test result
wine_testing_result <- test$quality

wine_train_set[1,]
```

```
##             age_drink           alsac_alsac             appl_pear
##             0.0000000             0.0000000             0.0000000
##          aroma_flavor            aroma_lead            bake_spice
##             0.0000000             0.0000000             0.0000000
##       barolo_nebbiolo           berri_aroma           berri_flavor
##             0.0000000             0.0000000             0.0000000
##           berri_fruit          black_cherri          black_currant
##             0.0000000             0.0000000             0.0000000
##           black_fruit          black_pepper             black_plum
##             0.0000000             0.0000000             0.0000000
##     blackberri_cherri        blend_cabernet     bordeaux_bordeaux
##             0.0000000             0.0000000             0.0000000
##      bordeauxstyl_red           bright_acid           brunello_di
##             0.0000000             0.0000000             0.0000000
##        cabernet_franc    cabernet_sauvignon california_california
##             0.1305604             0.0000000             0.0000000
##       california_napa       california_paso    california_russian
##             0.0000000             0.0000000             0.0000000
##      california_santa     california_sonoma         central_coast
##             0.0000000             0.0000000             0.0000000
##        central_valley        champagn_blend     champagn_champagn
##             0.0000000             0.0000000             0.0000000
##         cherri_flavor          cherri_fruit      cherri_raspberri
##             0.0000000             0.0000000             0.0000000
##      coast_chardonnay           coast_pinot          coast_sonoma
##             0.0000000             0.0000000             0.0000000
##       columbia_valley        counti_central            crisp_acid
##             0.0000000             0.0000000             0.0000000
##          cru_burgundi               ctes_de           dark_chocol
##             0.0000000             0.0000000             0.0000000
##         di_montalcino              dri_herb           drink_franc
##             0.0000000             0.0000000             0.0000000
##           drink_itali             drink_now          drink_portug
##             0.0000000             0.0000000             0.0000000
```

```
##      estat_california           finger_lake          finish_drink
##             0.0000000             0.0000000             0.0000000
##          finish_itali             finish_us           firm_tannin
##             0.0000000             0.0000000             0.0000000
##         flavor_finish             flavor_us        franc_bordeaux
##             0.0000000             0.0000000             0.0000000
##             franc_les            french_oak            fresh_acid
##             0.0000000             0.0000000             0.0000000
##          fruit_flavor             full_bodi            green_appl
##             0.0000000             0.0000000             0.0000000
##         grner_veltlin         itali_tuscani           lake_finger
##             0.0000000             0.0000000             0.0000000
##             lead_nose         linger_finish           loir_valley
##             0.0000000             0.0000000             0.0000000
##           long_finish       mendoza_provinc       merlot_cabernet
##             0.0000000             0.0000000             0.1623050
##  montalcino_sangioves         napa_cabernet           napa_valley
##             0.0000000             0.0000000             0.0000000
##              new_york           new_zealand    northeastern_itali
##             0.1406568             0.0000000             0.0000000
##        northern_spain            nose_palat             now_franc
##             0.0000000             0.0000000             0.0000000
##                now_us              old_vine            open_aroma
##             0.0000000             0.0000000             0.0000000
##      oregon_willamett           palat_deliv           palat_offer
##             0.0000000             0.0000000             0.0000000
##            palat_show             paso_robl           petit_sirah
##             0.0000000             0.0000000             0.0000000
##          petit_verdot       piedmont_barolo          pinot_grigio
##             0.0000000             0.0000000             0.0000000
##            pinot_gris             pinot_noir        portugues_red
##             0.0000000             0.0000000             0.0000000
##           premier_cru       provinc_mendoza      raspberri_cherri
##             0.0000000             0.0000000             0.0000000
##           readi_drink             red_berri             red_blend
##             0.0000000             0.1385762             0.0000000
##            red_cherri           red_currant             red_fruit
##             0.0000000             0.0000000             0.0000000
##     reserv_california          rhnestyl_red           ripe_fruit
##             0.0000000             0.0000000             0.0000000
##          river_valley          robl_central         russian_river
##             0.0000000             0.0000000             0.0000000
##       sauvignon_blanc       sicili_sardinia        sierra_foothil
##             0.0000000             0.0000000             0.0000000
##     sonoma_chardonnay          sonoma_coast         sonoma_counti
##             0.0000000             0.0000000             0.0000000
##          sonoma_pinot          south_africa       south_australia
##             0.0000000             0.0000000             0.0000000
##       southwest_franc           spain_rioja          sparkl_blend
##             0.0000000             0.0000000             0.0000000
##          spice_flavor           stone_fruit          tannin_drink
##             0.0000000             0.0000000             0.0000000
##          tropic_fruit      tuscani_brunello       tuscani_chianti
##             0.0000000             0.0000000             0.0000000
```

```
##         us_california              us_estat             us_oregon
##             0.0000000             0.0000000             0.0000000
##             us_reserv         us_washington       valley_cabernet
##             0.0000000             0.0000000             0.0000000
##        valley_central     valley_chardonnay           valley_napa
##             0.0000000             0.0000000             0.0000000
##          valley_pinot            valley_red         valley_sonoma
##             0.0000000             0.0000000             0.0000000
##          valley_syrah             valley_wa      valley_willamett
##             0.0000000             0.0000000             0.0000000
##   vineyard_california   vineyard_washington           wa_columbia
##             0.0000000             0.0000000             0.0000000
##          walla_valley           walla_walla   washington_columbia
##             0.0000000             0.0000000             0.0000000
##           white_blend           white_peach          white_pepper
##             0.0000000             0.0000000             0.0000000
##      willamett_valley              wood_age           york_finger
##             0.0000000             0.0000000             0.0000000
```

result:

```
train_svmLinear_model <- train(x= wine_train_set, y=train$quality , method = 'svmLinear3')
train_svmLinear_model
```

```
## L2 Regularized Support Vector Machine (dual) with Linear Kernel
##
## 20795 samples
##   159 predictor
##     2 classes: 'excellent', 'good'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 20795, 20795, 20795, 20795, 20795, 20795, ...
## Resampling results across tuning parameters:
##
##   cost  Loss  Accuracy   Kappa
##   0.25  L1    0.6468496  0.2859531
##   0.25  L2    0.6488347  0.2957380
##   0.50  L1    0.6495618  0.2947636
##   0.50  L2    0.6497968  0.2977784
##   1.00  L1    0.6489001  0.2961883
##   1.00  L2    0.6500568  0.2983729
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were cost = 1 and Loss = L2.
```

```
model_svmLinear_result <- predict(train_svmLinear_model, newdata = wine_test_set)

conf_svmLinear_train <- table(model_svmLinear_result, wine_testing_result)
names(dimnames(conf_svmLinear_train)) <- c("Predicted class", "Actual class")
confusionMatrix(conf_svmLinear_train)
```

```
## Confusion Matrix and Statistics
```

```
##
##                Actual class
## Predicted class excellent good
##       excellent      1344  756
##       good           1096 2003
##
##                Accuracy : 0.6438
##                  95% CI : (0.6306, 0.6568)
##     No Information Rate : 0.5307
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.2791
##  Mcnemar's Test P-Value : 3.344e-15
##
##             Sensitivity : 0.5508
##             Specificity : 0.7260
##          Pos Pred Value : 0.6400
##          Neg Pred Value : 0.6463
##              Prevalence : 0.4693
##          Detection Rate : 0.2585
##    Detection Prevalence : 0.4039
##       Balanced Accuracy : 0.6384
##
##        'Positive' Class : excellent
##
```

```
train_svm_model <- train(x= wine_train_set, y=train$quality , method = 'svmLinear3')
```

The result is quite bad, which means 2-gram is not usuful. Based on my observation. After preprocess and steaming, the description is quite discrate. I means the words is likely not connected to each other. It lead to 2-gram model not works.

- I continue with both bag of words and 2-gram (which means 1-2gram). Here is the first line of Dtm:

```
NLP_tokenizer <- function(x) {
  unlist(lapply(ngrams(words(x), 1:2), paste, collapse = "_"), use.names = FALSE)
}

wine_train_set <- clean(train$description)

train_dtm_tfidf <- DocumentTermMatrix(wine_train_set, control = list(weighting = weightTfIdf, tokenize=
#train_dtm_tfidf <- DocumentTermMatrix(wine_train_set, control = list( tokenize=NLP_tokenizer))
#train_dtm_tfidf <- DocumentTermMatrix(wine_train_set)
train_dtm_tfidf <- removeSparseTerms(train_dtm_tfidf, 0.99)


NLP_tokenizer <- function(x) {
  unlist(lapply(ngrams(words(x), 1:1), paste, collapse = "_"), use.names = FALSE)
}

#create the test set
wine_test_set <- clean(test$description)
wine_test_set <- DocumentTermMatrix(wine_test_set, control = list(dictionary = Terms(train_dtm_tfidf) ,
```

```
## Warning in weighting(x): unreferenced term(s): age_drink alsac_alsac
## appl_pear aroma_flavor aroma_lead bake_spice barolo_nebbiolo
## berri_aroma berri_flavor berri_fruit black_cherri black_currant
## black_fruit black_pepper black_plum blackberri_cherri blend_cabernet
## bordeaux_bordeaux bordeauxstyl_red bright_acid brunello_di cabernet_franc
## cabernet_sauvignon california_california california_napa california_paso
## california_russian california_santa california_sonoma central_coast
## central_valley champagn_blend champagn_champagn cherri_flavor cherri_fruit
## cherri_raspberri coast_chardonnay coast_pinot coast_sonoma columbia_valley
## counti_central crisp_acid cru_burgundi ctes_de dark_chocol di_montalcino
## dri_herb drink_franc drink_itali drink_now drink_portug estat_california
## finger_lake finish_drink finish_itali finish_us firm_tannin flavor_finish
## flavor_us franc_bordeaux franc_les french_oak fresh_acid fruit_flavor
## full_bodi green_appl grner_veltlin itali_tuscani lake_finger lead_nose
## linger_finish loir_valley long_finish mendoza_provinc merlot_cabernet
## montalcino_sangioves napa_cabernet napa_valley new_york new_zealand
## northeastern_itali northern_spain nose_palat now_franc now_us old_vine
## open_aroma oregon_willamett palat_deliv palat_offer palat_show paso_robl
## petit_sirah petit_verdot piedmont_barolo pinot_grigio pinot_gris
## pinot_noir portugues_red premier_cru provinc_mendoza raspberri_cherri
## readi_drink red_berri red_blend red_cherri red_currant red_fruit
## reserv_california rhnestyl_red ripe_fruit river_valley robl_central
## russian_river sauvignon_blanc sicili_sardinia sierra_foothil
## sonoma_chardonnay sonoma_coast sonoma_counti sonoma_pinot south_africa
## south_australia southwest_franc spain_rioja sparkl_blend spice_flavor
## stone_fruit tannin_drink tropic_fruit tuscani_brunello tuscani_chianti
## us_california us_estat us_oregon us_reserv us_washington valley_cabernet
## valley_central valley_chardonnay valley_napa valley_pinot valley_red
## valley_sonoma valley_syrah valley_wa valley_willamett vineyard_california
## vineyard_washington wa_columbia walla_valley walla_walla
## washington_columbia white_blend white_peach white_pepper willamett_valley
## wood_age york_finger
```

```
#wine_test_set <- DocumentTermMatrix(wine_test_set, control = list(dictionary = Terms(train_dtm_tfidf)

#create matrix for training
wine_train_set <<- as.matrix(train_dtm_tfidf)
wine_test_set <- as.matrix(wine_test_set)
wine_test_set <- wine_test_set[,Terms(train_dtm_tfidf)]
#create the test result
wine_testing_result <- test$quality

wine_train_set[1,]
```

```
##               accent                 acid                  add
##           0.00000000           0.00000000           0.00000000
##               africa            aftertast                  age
##           0.00000000           0.00000000           0.00000000
##            age_drink              alcohol               almond
##           0.00000000           0.00000000           0.00000000
##               almost                along              alongsid
##           0.00000000           0.05780700           0.00000000
##              alreadi                alsac           alsac_alsac
##           0.00000000           0.00000000           0.00000000
```

```
##                also            although               ampl
##          0.00000000          0.00000000          0.00000000
##                anis              appeal               appl
##          0.00000000          0.00000000          0.00000000
##           appl_pear            approach             apricot
##          0.00000000          0.00000000          0.00000000
##           argentina               aroma        aroma_flavor
##          0.00000000          0.00000000          0.00000000
##          aroma_lead              aromat              around
##          0.00000000          0.00000000          0.00000000
##             astring             attract           australia
##          0.00000000          0.00000000          0.00000000
##             austria                back                bake
##          0.00000000          0.00000000          0.00000000
##          bake_spice              balanc              barolo
##          0.00000000          0.00000000          0.00000000
##      barolo_nebbiolo              barrel                bean
##          0.00000000          0.00000000          0.00000000
##              beauti               berri         berri_aroma
##          0.00000000          0.03706390          0.00000000
##         berri_flavor         berri_fruit                best
##          0.00000000          0.00000000          0.00000000
##              better                 big                 bit
##          0.00000000          0.00000000          0.00000000
##              bitter               black         black_cherri
##          0.00000000          0.00000000          0.00000000
##        black_currant         black_fruit         black_pepper
##          0.00000000          0.00000000          0.00000000
##          black_plum           blackberri    blackberri_cherri
##          0.00000000          0.00000000          0.00000000
##               blanc               blend       blend_cabernet
##          0.00000000          0.02550924          0.00000000
##             blossom                blue            blueberri
##          0.00000000          0.00000000          0.00000000
##               boast                bodi                bold
##          0.00000000          0.00000000          0.00000000
##            bordeaux    bordeaux_bordeaux         bordeauxstyl
##          0.00000000          0.00000000          0.00000000
##     bordeauxstyl_red               bottl              bouquet
##          0.00000000          0.05089149          0.00000000
##              bright         bright_acid               bring
##          0.00000000          0.00000000          0.00000000
##               brisk             brunello          brunello_di
##          0.00000000          0.00000000          0.00000000
##                brut             burgundi              butter
##          0.00000000          0.00000000          0.00000000
##                 cab             cabernet       cabernet_franc
##          0.00000000          0.03687528          0.06528019
##    cabernet_sauvignon          california california_california
##          0.00000000          0.00000000          0.00000000
##       california_napa       california_paso    california_russian
##          0.00000000          0.00000000          0.00000000
##      california_santa     california_sonoma                 can
##          0.00000000          0.00000000          0.00000000
```

```
##                candi              caramel                carri
##           0.00000000           0.00000000           0.00000000
##                cassi                cedar                cellar
##           0.00000000           0.00000000           0.00000000
##               central         central_coast        central_valley
##           0.00000000           0.00000000           0.00000000
##              champagn        champagn_blend     champagn_champagn
##           0.00000000           0.00000000           0.00000000
##                 char              charact            chardonnay
##           0.00000000           0.00000000           0.04126152
##                cherri         cherri_flavor          cherri_fruit
##           0.00000000           0.00000000           0.00000000
##      cherri_raspberri                chewi               chianti
##           0.00000000           0.00000000           0.00000000
##                chile                chocol                chunki
##           0.00000000           0.00000000           0.00000000
##              cinnamon                citrus                citrusi
##           0.00000000           0.00000000           0.00000000
##               classic              classico                clean
##           0.00000000           0.00000000           0.00000000
##                close                clove                coast
##           0.00000000           0.00000000           0.00000000
##      coast_chardonnay          coast_pinot          coast_sonoma
##           0.00000000           0.00000000           0.00000000
##                cocoa                coffe                 cola
##           0.00000000           0.00000000           0.00000000
##                color              columbia       columbia_valley
##           0.00000000           0.00000000           0.00000000
##               combin                come               complex
##           0.00000000           0.00000000           0.00000000
##              concentr                cool                core
##           0.00000000           0.00000000           0.00000000
##               counti       counti_central            cranberri
##           0.00000000           0.00000000           0.00000000
##                cream               creami               creek
##           0.00000000           0.00000000           0.00000000
##                crisp           crisp_acid                cru
##           0.00000000           0.00000000           0.00000000
##          cru_burgundi                crush                ctes
##           0.00000000           0.00000000           0.00000000
##               ctes_de              currant                cut
##           0.00000000           0.00000000           0.00000000
##                cuve                 dark          dark_chocol
##           0.00000000           0.00000000           0.00000000
##                deep                  del                delic
##           0.00000000           0.00000000           0.00000000
##               delici                deliv                 dens
##           0.00000000           0.00000000           0.00000000
##                depth               despit              develop
##           0.00000000           0.00000000           0.00000000
##         di_montalcino               doesnt                domin
##           0.00000000           0.00000000           0.00000000
##                  dri             dri_herb                drink
##           0.00000000           0.00000000           0.00000000
```

```
##          drink_franc          drink_itali           drink_now
##           0.00000000           0.00000000          0.00000000
##         drink_portug                dusti               earth
##           0.00000000           0.00000000          0.00000000
##              earthi                 easi                 edg
##           0.00000000           0.00000000          0.00000000
##                 eleg              element                 end
##           0.00000000           0.00000000          0.00000000
##                enjoy               enough            espresso
##           0.00000000           0.00000000          0.00000000
##                estat     estat_california                even
##           0.00000000           0.00000000          0.00000000
##                excel                 exot             express
##           0.00000000           0.00000000          0.00000000
##                extra              extract                fair
##           0.00000000           0.00000000          0.00000000
##                famili               featur                feel
##           0.00000000           0.00000000          0.00000000
##               ferment                 fill                find
##           0.00000000           0.00000000          0.00000000
##                 fine               finger          finger_lake
##           0.00000000           0.00000000           0.00000000
##                finish         finish_drink         finish_itali
##           0.00000000           0.00000000           0.00000000
##              finish_us                firm          firm_tannin
##           0.00000000           0.00000000           0.00000000
##                first               flavor        flavor_finish
##           0.00000000           0.00000000          0.00000000
##              flavor_us               fleshi               floral
##           0.00000000           0.00000000          0.06381930
##                flower                focus               follow
##           0.00000000           0.00000000          0.00000000
##                 food              foothil               forest
##           0.00000000           0.00000000          0.00000000
##               forward             fragrant                frame
##           0.00000000           0.00000000          0.00000000
##                franc        franc_bordeaux             franc_les
##           0.02910668           0.00000000          0.00000000
##                french            french_oak                fresh
##           0.00000000           0.00000000          0.03314167
##              fresh_acid                front                fruit
##           0.00000000           0.00000000          0.01674002
##         fruit_flavor               fruiti                 full
##           0.00000000           0.04369158          0.00000000
##              full_bodi             fullbodi              generous
##           0.00000000           0.00000000          0.00000000
##                gentl              germani                  get
##           0.00000000           0.00000000          0.00000000
##                 give                glass                 good
##           0.00000000           0.00000000          0.00000000
##                grand                grape           grapefruit
##           0.00000000           0.00000000          0.00000000
##               graphit                great                green
##           0.00000000           0.00000000          0.04841613
```

```
##          green_appl             grenach              grigio
##          0.00000000          0.00000000          0.00000000
##               grill                grip                gris
##          0.00000000          0.00000000          0.00000000
##               grner        grner_veltlin               grown
##          0.00000000          0.00000000          0.00000000
##                hard               heavi                herb
##          0.00000000          0.00000000          0.04491553
##              herbal                here                high
##          0.00000000          0.00000000          0.00000000
##                hill                hint                hold
##          0.00000000          0.00000000          0.00000000
##               honey             impress              includ
##          0.00000000          0.00000000          0.00000000
##              integr              intens            interest
##          0.00000000          0.05463712          0.00000000
##             intrigu               itali        itali_tuscani
##          0.00000000          0.00000000          0.00000000
##                 jam               jammi                juic
##          0.00000000          0.00000000          0.00000000
##               juici                just                keep
##          0.00000000          0.00000000          0.00000000
##                lack                lake          lake_finger
##          0.00000000          0.00000000          0.00000000
##                last               layer                lead
##          0.00000000          0.00000000          0.00000000
##           lead_nose                leaf                lean
##          0.00000000          0.00000000          0.00000000
##             leather                leav               lemon
##          0.00000000          0.00000000          0.00000000
##              lemoni                lend              length
##          0.00000000          0.00000000          0.00000000
##                 les             licoric                lift
##          0.00000000          0.00000000          0.00000000
##               light                like                lime
##          0.00000000          0.00000000          0.00000000
##              linger        linger_finish               littl
##          0.00000000          0.00000000          0.00000000
##                live                load                loir
##          0.00000000          0.00000000          0.00000000
##         loir_valley                long          long_finish
##          0.00000000          0.10835301          0.00000000
##                 lot                love                lush
##          0.00000000          0.00000000          0.00000000
##                made                make              malbec
##          0.00000000          0.00000000          0.00000000
##               mango                mani                mark
##          0.00000000          0.00000000          0.00000000
##               matur                meat              medium
##          0.00000000          0.00000000          0.00000000
##          mediumbodi               melon             mendoza
##          0.00000000          0.00000000          0.00000000
##      mendoza_provinc              merlot      merlot_cabernet
##          0.00000000          0.05111151          0.08115250
```

```
##              midpal                mild               miner
##          0.00000000          0.00000000          0.00000000
##                mint                 mix               mocha
##          0.00000000          0.00000000          0.00000000
##               moder           montalcino montalcino_sangioves
##          0.00000000          0.00000000          0.00000000
##               month            mountain               mouth
##          0.00000000          0.00000000          0.00000000
##            mouthfeel                much                napa
##          0.05932371          0.00000000          0.00000000
##        napa_cabernet          napa_valley               natur
##          0.00000000          0.00000000          0.00000000
##            nebbiolo            nectarin                need
##          0.00000000          0.00000000          0.00000000
##                 new            new_york          new_zealand
##          0.04915548          0.07032841          0.00000000
##                next                nice                noir
##          0.00000000          0.00000000          0.00000000
##               north         northeastern   northeastern_itali
##          0.07979934          0.00000000          0.00000000
##            northern       northern_spain                nose
##          0.00000000          0.00000000          0.03652355
##          nose_palat                note                 now
##          0.00000000          0.03116226          0.00000000
##           now_franc              now_us               nuanc
##          0.00000000          0.00000000          0.00000000
##                 nut                 oak                oaki
##          0.00000000          0.00000000          0.00000000
##               offer                 old            old_vine
##          0.03822461          0.00000000          0.00000000
##                oliv                 one                open
##          0.00000000          0.00000000          0.00000000
##          open_aroma                opul               orang
##          0.00000000          0.00000000          0.00000000
##              oregon     oregon_willamett               overal
##          0.00000000          0.00000000          0.00000000
##                pack                pair               palat
##          0.00000000          0.00000000          0.02204908
##         palat_deliv         palat_offer          palat_show
##          0.00000000          0.00000000          0.00000000
##                paso           paso_robl               peach
##          0.00000000          0.00000000          0.00000000
##                pear                peel              pepper
##          0.00000000          0.00000000          0.00000000
##             pepperi             perfect               perfum
##          0.00000000          0.00000000          0.00000000
##             persist               petit          petit_sirah
##          0.00000000          0.00000000          0.00000000
##         petit_verdot                 pie            piedmont
##          0.00000000          0.00000000          0.00000000
##      piedmont_barolo            pineappl                pink
##          0.00000000          0.00000000          0.00000000
##                pinot         pinot_grigio           pinot_gris
##          0.00000000          0.00000000          0.00000000
```

```
##           pinot_noir                 play              pleasant
##           0.00000000           0.00000000           0.07376257
##               plenti                 plum                 plump
##           0.00000000           0.00000000           0.00000000
##                plush               polish                portug
##           0.00000000           0.00000000           0.00000000
##             portugues         portugues_red              potenti
##           0.00000000           0.00000000           0.00000000
##                power              premier          premier_cru
##           0.00000000           0.00000000           0.00000000
##               pretti                price                produc
##           0.06690534           0.00000000           0.00000000
##              provenc               provid               provinc
##           0.00000000           0.00000000           0.00000000
##       provinc_mendoza                prune                 pure
##           0.00000000           0.00000000           0.00000000
##              qualiti                 quit                 raci
##           0.00000000           0.00000000           0.00000000
##                raisin             raspberri      raspberri_cherri
##           0.00000000           0.00000000           0.00000000
##                rather                  raw                readi
##           0.00000000           0.00000000           0.00000000
##           readi_drink                  red            red_berri
##           0.00000000           0.02432509           0.06928811
##            red_blend            red_cherri          red_currant
##           0.00000000           0.00000000           0.00000000
##            red_fruit              refresh               region
##           0.00000000           0.00000000           0.00000000
##                remain               reserv     reserv_california
##           0.00000000           0.00000000           0.00000000
##              reserva               reveal                 rhne
##           0.00000000           0.00000000           0.00000000
##             rhnestyl          rhnestyl_red                 rich
##           0.00000000           0.00000000           0.00000000
##                riesl                right                rioja
##           0.00000000           0.00000000           0.00000000
##                 ripe            ripe_fruit              riserva
##           0.00000000           0.00000000           0.00000000
##                river          river_valley                roast
##           0.00000000           0.00000000           0.00000000
##                 robl          robl_central                  ros
##           0.00000000           0.00000000           0.11377897
##                 rose                round              russian
##           0.00000000           0.00000000           0.00000000
##         russian_river                rustic                 sage
##           0.00000000           0.00000000           0.00000000
##            sangioves                santa              sardinia
##           0.00000000           0.00000000           0.00000000
##             sauvignon       sauvignon_blanc                savori
##           0.00000000           0.00000000           0.00000000
##                scent                 seem               select
##           0.00000000           0.00000000           0.00000000
##                 sens                  set                sharp
##           0.00000000           0.00000000           0.00000000
```

```
##                show               sicili       sicili_sardinia
##          0.00000000           0.00000000            0.00000000
##                side               sierra        sierra_foothil
##          0.00000000           0.00000000            0.00000000
##               silki                simpl                   sip
##          0.00000000           0.00000000            0.00000000
##               sirah                skin                 slight
##          0.00000000           0.00000000            0.00000000
##               smell                smoke                 smoki
##          0.00000000           0.00000000            0.00000000
##              smooth                soft                 soften
##          0.00000000           0.00000000            0.00000000
##               soil               solid               somewhat
##          0.00000000           0.00000000            0.00000000
##              sonoma    sonoma_chardonnay          sonoma_coast
##          0.00000000           0.00000000            0.00000000
##       sonoma_counti         sonoma_pinot                  soon
##          0.00000000           0.00000000            0.00000000
##               sour                sourc                 south
##          0.00000000           0.00000000            0.00000000
##        south_africa      south_australia               southern
##          0.00000000           0.00000000            0.00000000
##           southwest      southwest_franc                 spain
##          0.00000000           0.00000000            0.00000000
##        spain_rioja               sparkl          sparkl_blend
##          0.00000000           0.00000000            0.00000000
##               spice         spice_flavor                 spici
##          0.00000000           0.00000000            0.00000000
##               start                still                 stone
##          0.00000000           0.00000000            0.00000000
##         stone_fruit      straightforward            strawberri
##          0.00000000           0.00000000            0.00000000
##              streak               strong              structur
##          0.00000000           0.00000000            0.00000000
##               style                subtl                 sugar
##          0.00000000           0.00000000            0.00000000
##             suggest              superior                 suppl
##          0.00000000           0.00000000            0.00000000
##             support                sweet                 syrah
##          0.00000000           0.00000000            0.00000000
##                take              tangerin                 tangi
##          0.00000000           0.00000000            0.00000000
##              tannic               tannin          tannin_drink
##          0.00000000           0.00000000            0.00000000
##                tart                tast                   tea
##          0.00000000           0.00000000            0.00000000
##          tempranillo               textur                  that
##          0.00000000           0.00000000            0.00000000
##               there               thick                though
##          0.00000000           0.00000000            0.00000000
##               tight                time                 toast
##          0.00000000           0.00000000            0.00000000
##              toasti              tobacco                togeth
##          0.00000000           0.00000000            0.00000000
```

```
##               tomato                 tone                  top
##           0.00000000           0.00000000           0.00000000
##              toscana                touch               tropic
##           0.00000000           0.00000000           0.00000000
##         tropic_fruit                 turn              tuscani
##           0.00000000           0.00000000           0.00000000
##     tuscani_brunello       tuscani_chianti                  two
##           0.00000000           0.00000000           0.00000000
##           underbrush         us_california              us_estat
##           0.00000000           0.00000000           0.00000000
##            us_oregon            us_reserv         us_washington
##           0.00000000           0.00000000           0.00000000
##                valley       valley_cabernet        valley_central
##           0.00000000           0.00000000           0.00000000
##    valley_chardonnay           valley_napa          valley_pinot
##           0.00000000           0.00000000           0.00000000
##           valley_red         valley_sonoma          valley_syrah
##           0.00000000           0.00000000           0.00000000
##            valley_wa      valley_willamett               vanilla
##           0.00000000           0.00000000           0.00000000
##                variet               varieti              veltlin
##           0.00000000           0.00000000           0.00000000
##               velveti                veneto               verdot
##           0.00000000           0.00000000           0.00000000
##                  veri               vibrant                 vine
##           0.00000000           0.00000000           0.00000000
##              vineyard   vineyard_california   vineyard_washington
##           0.00000000           0.00000000           0.00000000
##                vintag              viognier               violet
##           0.00000000           0.00000000           0.00000000
##          wa_columbia                 walla          walla_valley
##           0.00000000           0.00000000           0.00000000
##          walla_walla                 warm            washington
##           0.00000000           0.00000000           0.00000000
##   washington_columbia                  way               weight
##           0.00000000           0.00000000           0.00000000
##                 well                  wet                whiff
##           0.00000000           0.00000000           0.00000000
##                white           white_blend          white_peach
##           0.00000000           0.00000000           0.00000000
##          white_pepper                 wild                 will
##           0.00000000           0.00000000           0.00000000
##            willamett      willamett_valley                 wine
##           0.00000000           0.00000000           0.00000000
##               winemak                wineri              without
##           0.00000000           0.00000000           0.00000000
##                 wood              wood_age                 wrap
##           0.00000000           0.00000000           0.00000000
##                 year                 yellow                  yet
##           0.00000000           0.00000000           0.00000000
##                 york           york_finger                young
##           0.07024294           0.00000000           0.00000000
##              zealand                 zest                zesti
##           0.00000000           0.00000000           0.00000000
```

```
##             zinfandel
##            0.00000000
```

And here is the result:

```
train_svmLinear_model <- train(x= wine_train_set, y=train$quality , method = 'svmLinear3')
train_svmLinear_model
```

```
## L2 Regularized Support Vector Machine (dual) with Linear Kernel
##
## 20795 samples
##   664 predictor
##     2 classes: 'excellent', 'good'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 20795, 20795, 20795, 20795, 20795, 20795, ...
## Resampling results across tuning parameters:
##
##   cost  Loss  Accuracy   Kappa
##   0.25  L1    0.7711318  0.5403063
##   0.25  L2    0.7809449  0.5602228
##   0.50  L1    0.7763530  0.5511321
##   0.50  L2    0.7811762  0.5608099
##   1.00  L1    0.7793415  0.5572261
##   1.00  L2    0.7808706  0.5602739
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were cost = 0.5 and Loss = L2.
```

```
model_svmLinear_result <- predict(train_svmLinear_model, newdata = wine_test_set)

conf_svmLinear_train <- table(model_svmLinear_result, wine_testing_result)
names(dimnames(conf_svmLinear_train)) <- c("Predicted class", "Actual class")
confusionMatrix(conf_svmLinear_train)
```

```
## Confusion Matrix and Statistics
##
##                Actual class
## Predicted class excellent good
##       excellent      1695  440
##       good            745 2319
##
##                Accuracy : 0.7721
##                  95% CI : (0.7604, 0.7834)
##     No Information Rate : 0.5307
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5391
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.6947
##             Specificity : 0.8405
```

```
##          Pos Pred Value : 0.7939
##          Neg Pred Value : 0.7569
##              Prevalence : 0.4693
##          Detection Rate : 0.3260
##    Detection Prevalence : 0.4107
##       Balanced Accuracy : 0.7676
##
##        'Positive' Class : excellent
##
```

The accurency is is slightly decreses. Good new is the False negative is slightly better. It's like the trade off between models. Consider that and the dimension of Dtm is large (which lead to high trainning time), while there are too much features with 0 point. I conclude that 1-2 gram is not so useful.

- An other attempt to improve is adjust the weight for Dtm. As I obserse in the data. Description usually talk about ingerdiens: lemon, cherry or region: Veneto... So, the first idea is double the point for nouns and see how it works. On the other hand, Adjactive used for drescribe the flavor of wine could make an important role. So, the second ideal is double the point for adj only. To do these. I implement postaging and see which word is nouns, adj, verb... Here is the list nouns that I extracted:'

```
nouns
```

```
##   [1] "accent"       "acid"         "age"          "alcohol"
##   [5] "almond"       "alreadi"      "alsac"        "approach"
##   [9] "apricot"      "argentina"    "aroma"        "austria"
##  [13] "bake"         "balanc"       "barolo"       "barrel"
##  [17] "bean"         "beauti"       "berri"        "bit"
##  [21] "blanc"        "blend"        "bordeaux"     "bordeauxstyl"
##  [25] "bottl"        "bouquet"      "brunello"     "brut"
##  [29] "burgundi"     "butter"       "cab"          "cabernet"
##  [33] "california"   "caramel"      "carri"        "cassi"
##  [37] "cedar"        "cellar"       "champagn"     "char"
##  [41] "charact"      "chardonnay"   "cherri"       "chewi"
##  [45] "chocol"       "chunki"       "cinnamon"     "citrus"
##  [49] "classico"     "clove"        "coast"        "cocoa"
##  [53] "coffe"        "cola"         "color"        "columbia"
##  [57] "combin"       "concentr"     "core"         "counti"
##  [61] "cranberri"    "cream"        "creami"       "creek"
##  [65] "crisp"        "cru"          "crush"        "cut"
##  [69] "delic"        "depth"        "despit"       "domin"
##  [73] "dri"          "drink"        "earth"        "earthi"
##  [77] "eleg"         "element"      "end"          "espresso"
##  [81] "estat"        "express"      "ferment"      "fill"
##  [85] "finger"       "firm"         "flavor"       "flower"
##  [89] "focus"        "food"         "foothil"      "forest"
##  [93] "frame"        "franc"        "french"       "front"
##  [97] "fruit"        "fruiti"       "gentl"        "glass"
## [101] "grape"        "grapefruit"   "grenach"      "grigio"
## [105] "grill"        "grip"         "herb"         "hill"
## [109] "hint"         "honey"        "integr"       "interest"
## [113] "jam"          "jammi"        "lack"         "lake"
## [117] "layer"        "lead"         "leaf"         "leather"
## [121] "leav"         "lemon"        "length"       "lift"
```

```
## [125] "light"        "lime"         "linger"       "load"
## [129] "mango"        "mark"         "matur"        "meat"
## [133] "medium"       "melon"        "mendoza"      "merlot"
## [137] "midpal"       "miner"        "mint"         "mix"
## [141] "mocha"        "moder"        "month"        "mountain"
## [145] "mouth"        "mouthfeel"    "napa"         "natur"
## [149] "nebbiolo"     "nectarin"     "noir"         "nose"
## [153] "nut"          "oak"          "oaki"         "one"
## [157] "opul"         "oregon"       "pack"         "pair"
## [161] "palat"        "paso"         "peach"        "pear"
## [165] "peel"         "pepper"       "pepperi"      "perfum"
## [169] "pie"          "piedmont"     "pineappl"     "pinot"
## [173] "play"         "plum"         "polish"       "portug"
## [177] "power"        "premier"      "price"        "produc"
## [181] "provenc"      "provid"       "provinc"      "prune"
## [185] "qualiti"      "raspberri"    "readi"        "region"
## [189] "reserva"      "rhnestyl"     "rioja"        "riserva"
## [193] "river"        "roast"        "sage"         "sardinia"
## [197] "sauvignon"    "scent"        "show"         "side"
## [201] "sierra"       "silki"        "sip"          "skin"
## [205] "smell"        "smoke"        "soil"         "sonoma"
## [209] "sourc"        "southwest"    "spain"        "sparkl"
## [213] "spice"        "stone"        "streak"       "structur"
## [217] "style"        "sugar"        "support"      "syrah"
## [221] "tangerin"     "tannin"       "tart"         "tast"
## [225] "tea"          "textur"       "time"         "toast"
## [229] "toasti"       "tobacco"      "tone"         "touch"
## [233] "tropic"       "turn"         "valley"       "vanilla"
## [237] "varieti"      "veltlin"      "velveti"      "veneto"
## [241] "verdot"       "veri"         "vine"         "vineyard"
## [245] "vintag"       "viognier"     "violet"       "walla"
## [249] "way"          "weight"       "will"         "wine"
## [253] "winemak"      "wineri"       "wood"         "wrap"
## [257] "year"         "yellow"       "zealand"      "zest"
## [261] "zesti"        "zinfandel"
```

Here is the adj list:

```
adj
```

```
##   [1] "aftertast"      "ampl"           "appl"
##   [4] "australia"      "big"            "bitter"
##   [7] "black"          "blackberri"     "blue"
##  [10] "bodi"           "bold"           "bright"
##  [13] "brisk"          "central"        "classic"
##  [16] "clean"          "close"          "complex"
##  [19] "cool"           "currant"        "cuve"
##  [22] "dark"           "deep"           "doesnt"
##  [25] "edg"            "enough"         "exot"
##  [28] "extra"          "fair"           "fine"
##  [31] "floral"         "fragrant"       "fresh"
##  [34] "full"           "generous"       "good"
##  [37] "grand"          "great"          "green"
```

```
##  [40] "hard"            "heavi"           "high"
##  [43] "juic"            "last"            "lean"
##  [46] "licoric"         "littl"           "live"
##  [49] "lush"            "malbec"          "mild"
##  [52] "new"             "next"            "nice"
##  [55] "northeastern"    "northern"        "nuanc"
##  [58] "old"             "oliv"            "open"
##  [61] "overal"          "perfect"         "petit"
##  [64] "pink"            "pleasant"        "plush"
##  [67] "pure"            "raw"             "red"
##  [70] "refresh"         "rhne"            "rich"
##  [73] "right"           "ripe"            "robl"
##  [76] "russian"         "rustic"          "select"
##  [79] "sharp"           "sicili"          "simpl"
##  [82] "slight"          "smooth"          "soft"
##  [85] "solid"           "sour"            "south"
##  [88] "southern"        "straightforward" "strawberri"
##  [91] "strong"          "superior"        "suppl"
##  [94] "sweet"           "tannic"          "tempranillo"
##  [97] "thick"           "tight"           "top"
## [100] "variet"          "vibrant"         "warm"
## [103] "wet"             "white"           "wild"
## [106] "young"
```

```
column_id <- c()

#multify for noun
for (i in 1:dim(wine_train_set)[2]) {
  check <- colnames(wine_train_set)[i] %in% nouns
  if(check)
    {
      column_id <- c(column_id, i)
    }
}

wine_train_set[,column_id] <- wine_train_set[,column_id]*2
wine_test_set[,column_id] <- wine_test_set[,column_id]*2


wine_train_set[1,]
```

```
##               accent                 acid                  add
##           0.00000000           0.00000000           0.00000000
##               africa            aftertast                  age
##           0.00000000           0.00000000           0.00000000
##            age_drink              alcohol              almond
##           0.00000000           0.00000000           0.00000000
##               almost                along             alongsid
##           0.00000000           0.05780700           0.00000000
##              alreadi                 alsac          alsac_alsac
##           0.00000000           0.00000000           0.00000000
##                 also             although                 ampl
##           0.00000000           0.00000000           0.00000000
##                anis               appeal                 appl
```

```
##           0.00000000           0.00000000           0.00000000
##            appl_pear             approach              apricot
##           0.00000000           0.00000000           0.00000000
##            argentina                aroma          aroma_flavor
##           0.00000000           0.00000000           0.00000000
##           aroma_lead               aromat               around
##           0.00000000           0.00000000           0.00000000
##              astring              attract            australia
##           0.00000000           0.00000000           0.00000000
##              austria                 back                 bake
##           0.00000000           0.00000000           0.00000000
##           bake_spice               balanc               barolo
##           0.00000000           0.00000000           0.00000000
##      barolo_nebbiolo               barrel                 bean
##           0.00000000           0.00000000           0.00000000
##               beauti                berri           berri_aroma
##           0.00000000           0.07412781           0.00000000
##          berri_flavor           berri_fruit                 best
##           0.00000000           0.00000000           0.00000000
##               better                  big                  bit
##           0.00000000           0.00000000           0.00000000
##               bitter                black          black_cherri
##           0.00000000           0.00000000           0.00000000
##        black_currant           black_fruit          black_pepper
##           0.00000000           0.00000000           0.00000000
##           black_plum            blackberri     blackberri_cherri
##           0.00000000           0.00000000           0.00000000
##                blanc                blend        blend_cabernet
##           0.00000000           0.05101848           0.00000000
##              blossom                 blue             blueberri
##           0.00000000           0.00000000           0.00000000
##                boast                 bodi                 bold
##           0.00000000           0.00000000           0.00000000
##             bordeaux     bordeaux_bordeaux          bordeauxstyl
##           0.00000000           0.00000000           0.00000000
##      bordeauxstyl_red                bottl              bouquet
##           0.00000000           0.10178298           0.00000000
##               bright           bright_acid                bring
##           0.00000000           0.00000000           0.00000000
##                brisk              brunello           brunello_di
##           0.00000000           0.00000000           0.00000000
##                 brut              burgundi               butter
##           0.00000000           0.00000000           0.00000000
##                  cab              cabernet        cabernet_franc
##           0.00000000           0.07375056           0.06528019
##    cabernet_sauvignon           california california_california
##           0.00000000           0.00000000           0.00000000
##       california_napa       california_paso    california_russian
##           0.00000000           0.00000000           0.00000000
##      california_santa     california_sonoma                  can
##           0.00000000           0.00000000           0.00000000
##                candi              caramel                carri
##           0.00000000           0.00000000           0.00000000
##                cassi                cedar               cellar
```

```
##           0.00000000           0.00000000           0.00000000
##              central        central_coast       central_valley
##           0.00000000           0.00000000           0.00000000
##              champagn       champagn_blend    champagn_champagn
##           0.00000000           0.00000000           0.00000000
##                 char              charact           chardonnay
##           0.00000000           0.00000000           0.08252304
##              cherri        cherri_flavor         cherri_fruit
##           0.00000000           0.00000000           0.00000000
##     cherri_raspberri                chewi              chianti
##           0.00000000           0.00000000           0.00000000
##                chile               chocol                chunki
##           0.00000000           0.00000000           0.00000000
##              cinnamon               citrus              citrusi
##           0.00000000           0.00000000           0.00000000
##              classic              classico                clean
##           0.00000000           0.00000000           0.00000000
##                close                clove                coast
##           0.00000000           0.00000000           0.00000000
##     coast_chardonnay          coast_pinot         coast_sonoma
##           0.00000000           0.00000000           0.00000000
##                cocoa                coffe                cola
##           0.00000000           0.00000000           0.00000000
##                color              columbia      columbia_valley
##           0.00000000           0.00000000           0.00000000
##                combin                come              complex
##           0.00000000           0.00000000           0.00000000
##              concentr                cool                core
##           0.00000000           0.00000000           0.00000000
##              counti       counti_central             cranberri
##           0.00000000           0.00000000           0.00000000
##                cream               creami               creek
##           0.00000000           0.00000000           0.00000000
##                crisp           crisp_acid                cru
##           0.00000000           0.00000000           0.00000000
##         cru_burgundi                crush                ctes
##           0.00000000           0.00000000           0.00000000
##              ctes_de              currant                 cut
##           0.00000000           0.00000000           0.00000000
##                cuve                 dark          dark_chocol
##           0.00000000           0.00000000           0.00000000
##                deep                  del                delic
##           0.00000000           0.00000000           0.00000000
##              delici                deliv                 dens
##           0.00000000           0.00000000           0.00000000
##                depth               despit              develop
##           0.00000000           0.00000000           0.00000000
##        di_montalcino               doesnt                domin
##           0.00000000           0.00000000           0.00000000
##                 dri              dri_herb                drink
##           0.00000000           0.00000000           0.00000000
##          drink_franc          drink_itali           drink_now
##           0.00000000           0.00000000           0.00000000
##         drink_portug                dusti                earth
```

```
##          0.00000000          0.00000000          0.00000000
##              earthi                easi                 edg
##          0.00000000          0.00000000          0.00000000
##               eleg             element                 end
##          0.00000000          0.00000000          0.00000000
##               enjoy              enough            espresso
##          0.00000000          0.00000000          0.00000000
##               estat     estat_california                even
##          0.00000000          0.00000000          0.00000000
##               excel                exot             express
##          0.00000000          0.00000000          0.00000000
##               extra             extract                fair
##          0.00000000          0.00000000          0.00000000
##              famili              featur                feel
##          0.00000000          0.00000000          0.00000000
##             ferment                fill                find
##          0.00000000          0.00000000          0.00000000
##                fine              finger         finger_lake
##          0.00000000          0.00000000          0.00000000
##              finish         finish_drink         finish_itali
##          0.00000000          0.00000000          0.00000000
##           finish_us                firm          firm_tannin
##          0.00000000          0.00000000          0.00000000
##               first              flavor        flavor_finish
##          0.00000000          0.00000000          0.00000000
##           flavor_us              fleshi               floral
##          0.00000000          0.00000000          0.06381930
##              flower               focus              follow
##          0.00000000          0.00000000          0.00000000
##                food             foothil              forest
##          0.00000000          0.00000000          0.00000000
##             forward            fragrant               frame
##          0.00000000          0.00000000          0.00000000
##               franc       franc_bordeaux           franc_les
##          0.05821335          0.00000000          0.00000000
##              french           french_oak               fresh
##          0.00000000          0.00000000          0.03314167
##            fresh_acid               front               fruit
##          0.00000000          0.00000000          0.03348005
##         fruit_flavor              fruiti                full
##          0.00000000          0.08738315          0.00000000
##            full_bodi            fullbodi            generous
##          0.00000000          0.00000000          0.00000000
##               gentl             germani                 get
##          0.00000000          0.00000000          0.00000000
##                give               glass                good
##          0.00000000          0.00000000          0.00000000
##               grand               grape          grapefruit
##          0.00000000          0.00000000          0.00000000
##             graphit               great               green
##          0.00000000          0.00000000          0.04841613
##          green_appl             grenach              grigio
##          0.00000000          0.00000000          0.00000000
##               grill                grip                gris
```

```
##           0.00000000           0.00000000           0.00000000
##                grner        grner_veltlin                grown
##           0.00000000           0.00000000           0.00000000
##                 hard                heavi                herb
##           0.00000000           0.00000000           0.08983105
##               herbal                here                high
##           0.00000000           0.00000000           0.00000000
##                 hill                hint                hold
##           0.00000000           0.00000000           0.00000000
##                honey             impress              includ
##           0.00000000           0.00000000           0.00000000
##               integr              intens            interest
##           0.00000000           0.05463712           0.00000000
##              intrigu               itali         itali_tuscani
##           0.00000000           0.00000000           0.00000000
##                  jam                jammi                juic
##           0.00000000           0.00000000           0.00000000
##                juici                just                keep
##           0.00000000           0.00000000           0.00000000
##                 lack                lake           lake_finger
##           0.00000000           0.00000000           0.00000000
##                 last               layer                lead
##           0.00000000           0.00000000           0.00000000
##            lead_nose                leaf                lean
##           0.00000000           0.00000000           0.00000000
##              leather                leav               lemon
##           0.00000000           0.00000000           0.00000000
##               lemoni                lend              length
##           0.00000000           0.00000000           0.00000000
##                  les              licoric                lift
##           0.00000000           0.00000000           0.00000000
##                light                like                lime
##           0.00000000           0.00000000           0.00000000
##               linger        linger_finish               littl
##           0.00000000           0.00000000           0.00000000
##                 live                load                loir
##           0.00000000           0.00000000           0.00000000
##          loir_valley                long          long_finish
##           0.00000000           0.10835301           0.00000000
##                  lot                love                lush
##           0.00000000           0.00000000           0.00000000
##                 made                make              malbec
##           0.00000000           0.00000000           0.00000000
##                mango                mani                mark
##           0.00000000           0.00000000           0.00000000
##                matur                meat              medium
##           0.00000000           0.00000000           0.00000000
##           mediumbodi               melon             mendoza
##           0.00000000           0.00000000           0.00000000
##       mendoza_provinc              merlot       merlot_cabernet
##           0.00000000           0.10222301           0.08115250
##               midpal                mild               miner
##           0.00000000           0.00000000           0.00000000
##                 mint                 mix               mocha
```

```
##           0.00000000           0.00000000           0.00000000
##                moder             montalcino montalcino_sangioves
##           0.00000000           0.00000000           0.00000000
##                month              mountain                mouth
##           0.00000000           0.00000000           0.00000000
##            mouthfeel                 much                napa
##           0.11864741           0.00000000           0.00000000
##         napa_cabernet          napa_valley                natur
##           0.00000000           0.00000000           0.00000000
##              nebbiolo              nectarin                need
##           0.00000000           0.00000000           0.00000000
##                  new             new_york          new_zealand
##           0.04915548           0.07032841           0.00000000
##                next                 nice                noir
##           0.00000000           0.00000000           0.00000000
##                north          northeastern    northeastern_itali
##           0.07979934           0.00000000           0.00000000
##              northern        northern_spain                nose
##           0.00000000           0.00000000           0.07304711
##           nose_palat                 note                 now
##           0.00000000           0.03116226           0.00000000
##            now_franc               now_us                nuanc
##           0.00000000           0.00000000           0.00000000
##                  nut                  oak                oaki
##           0.00000000           0.00000000           0.00000000
##                offer                  old            old_vine
##           0.03822461           0.00000000           0.00000000
##                 oliv                  one                open
##           0.00000000           0.00000000           0.00000000
##           open_aroma                 opul                orang
##           0.00000000           0.00000000           0.00000000
##               oregon      oregon_willamett               overal
##           0.00000000           0.00000000           0.00000000
##                 pack                 pair                palat
##           0.00000000           0.00000000           0.04409815
##          palat_deliv          palat_offer           palat_show
##           0.00000000           0.00000000           0.00000000
##                 paso            paso_robl                peach
##           0.00000000           0.00000000           0.00000000
##                 pear                 peel               pepper
##           0.00000000           0.00000000           0.00000000
##              pepperi              perfect               perfum
##           0.00000000           0.00000000           0.00000000
##              persist                petit           petit_sirah
##           0.00000000           0.00000000           0.00000000
##         petit_verdot                  pie             piedmont
##           0.00000000           0.00000000           0.00000000
##       piedmont_barolo             pineappl                 pink
##           0.00000000           0.00000000           0.00000000
##                pinot          pinot_grigio           pinot_gris
##           0.00000000           0.00000000           0.00000000
##           pinot_noir                 play             pleasant
##           0.00000000           0.00000000           0.07376257
##               plenti                 plum                plump
```

```
##           0.00000000           0.00000000           0.00000000
##                plush               polish                portug
##           0.00000000           0.00000000           0.00000000
##            portugues        portugues_red              potenti
##           0.00000000           0.00000000           0.00000000
##                power              premier          premier_cru
##           0.00000000           0.00000000           0.00000000
##               pretti                price               produc
##           0.06690534           0.00000000           0.00000000
##              provenc               provid              provinc
##           0.00000000           0.00000000           0.00000000
##      provinc_mendoza                prune                 pure
##           0.00000000           0.00000000           0.00000000
##              qualiti                 quit                 raci
##           0.00000000           0.00000000           0.00000000
##               raisin             raspberri     raspberri_cherri
##           0.00000000           0.00000000           0.00000000
##               rather                  raw                readi
##           0.00000000           0.00000000           0.00000000
##          readi_drink                  red            red_berri
##           0.00000000           0.02432509           0.06928811
##            red_blend           red_cherri          red_currant
##           0.00000000           0.00000000           0.00000000
##            red_fruit              refresh               region
##           0.00000000           0.00000000           0.00000000
##               remain               reserv    reserv_california
##           0.00000000           0.00000000           0.00000000
##              reserva               reveal                rhne
##           0.00000000           0.00000000           0.00000000
##             rhnestyl         rhnestyl_red                rich
##           0.00000000           0.00000000           0.00000000
##                riesl                right                rioja
##           0.00000000           0.00000000           0.00000000
##                ripe           ripe_fruit              riserva
##           0.00000000           0.00000000           0.00000000
##                river          river_valley                roast
##           0.00000000           0.00000000           0.00000000
##                robl          robl_central                  ros
##           0.00000000           0.00000000           0.11377897
##                rose                round              russian
##           0.00000000           0.00000000           0.00000000
##         russian_river               rustic                sage
##           0.00000000           0.00000000           0.00000000
##            sangioves                santa             sardinia
##           0.00000000           0.00000000           0.00000000
##            sauvignon       sauvignon_blanc               savori
##           0.00000000           0.00000000           0.00000000
##                scent                 seem               select
##           0.00000000           0.00000000           0.00000000
##                sens                  set                sharp
##           0.00000000           0.00000000           0.00000000
##                show               sicili       sicili_sardinia
##           0.00000000           0.00000000           0.00000000
##                side               sierra        sierra_foothil
```

```
##          0.00000000          0.00000000          0.00000000
##               silki               simpl                 sip
##          0.00000000          0.00000000          0.00000000
##               sirah                skin               slight
##          0.00000000          0.00000000          0.00000000
##               smell               smoke                smoki
##          0.00000000          0.00000000          0.00000000
##              smooth                soft               soften
##          0.00000000          0.00000000          0.00000000
##                soil               solid             somewhat
##          0.00000000          0.00000000          0.00000000
##              sonoma    sonoma_chardonnay        sonoma_coast
##          0.00000000          0.00000000          0.00000000
##       sonoma_counti         sonoma_pinot                soon
##          0.00000000          0.00000000          0.00000000
##                sour               sourc               south
##          0.00000000          0.00000000          0.00000000
##        south_africa      south_australia             southern
##          0.00000000          0.00000000          0.00000000
##           southwest      southwest_franc               spain
##          0.00000000          0.00000000          0.00000000
##         spain_rioja               sparkl        sparkl_blend
##          0.00000000          0.00000000          0.00000000
##               spice         spice_flavor               spici
##          0.00000000          0.00000000          0.00000000
##               start               still               stone
##          0.00000000          0.00000000          0.00000000
##        stone_fruit      straightforward          strawberri
##          0.00000000          0.00000000          0.00000000
##              streak              strong            structur
##          0.00000000          0.00000000          0.00000000
##               style               subtl               sugar
##          0.00000000          0.00000000          0.00000000
##             suggest            superior               suppl
##          0.00000000          0.00000000          0.00000000
##             support               sweet               syrah
##          0.00000000          0.00000000          0.00000000
##                take             tangerin               tangi
##          0.00000000          0.00000000          0.00000000
##              tannic              tannin        tannin_drink
##          0.00000000          0.00000000          0.00000000
##                tart                tast                 tea
##          0.00000000          0.00000000          0.00000000
##         tempranillo              textur                that
##          0.00000000          0.00000000          0.00000000
##               there               thick              though
##          0.00000000          0.00000000          0.00000000
##               tight                time               toast
##          0.00000000          0.00000000          0.00000000
##              toasti             tobacco              togeth
##          0.00000000          0.00000000          0.00000000
##              tomato                tone                 top
##          0.00000000          0.00000000          0.00000000
##             toscana               touch              tropic
```

```
##           0.00000000           0.00000000           0.00000000
##         tropic_fruit                 turn              tuscani
##           0.00000000           0.00000000           0.00000000
##     tuscani_brunello      tuscani_chianti                  two
##           0.00000000           0.00000000           0.00000000
##           underbrush       us_california            us_estat
##           0.00000000           0.00000000           0.00000000
##            us_oregon            us_reserv       us_washington
##           0.00000000           0.00000000           0.00000000
##               valley      valley_cabernet       valley_central
##           0.00000000           0.00000000           0.00000000
##    valley_chardonnay          valley_napa         valley_pinot
##           0.00000000           0.00000000           0.00000000
##           valley_red        valley_sonoma         valley_syrah
##           0.00000000           0.00000000           0.00000000
##            valley_wa     valley_willamett              vanilla
##           0.00000000           0.00000000           0.00000000
##               variet              varieti              veltlin
##           0.00000000           0.00000000           0.00000000
##              velveti               veneto               verdot
##           0.00000000           0.00000000           0.00000000
##                 veri              vibrant                 vine
##           0.00000000           0.00000000           0.00000000
##             vineyard   vineyard_california   vineyard_washington
##           0.00000000           0.00000000           0.00000000
##               vintag             viognier               violet
##           0.00000000           0.00000000           0.00000000
##          wa_columbia                walla         walla_valley
##           0.00000000           0.00000000           0.00000000
##          walla_walla                 warm           washington
##           0.00000000           0.00000000           0.00000000
##   washington_columbia                  way               weight
##           0.00000000           0.00000000           0.00000000
##                 well                  wet                whiff
##           0.00000000           0.00000000           0.00000000
##                white           white_blend          white_peach
##           0.00000000           0.00000000           0.00000000
##         white_pepper                 wild                 will
##           0.00000000           0.00000000           0.00000000
##            willamett      willamett_valley                 wine
##           0.00000000           0.00000000           0.00000000
##              winemak              wineri               without
##           0.00000000           0.00000000           0.00000000
##                 wood             wood_age                 wrap
##           0.00000000           0.00000000           0.00000000
##                 year               yellow                  yet
##           0.00000000           0.00000000           0.00000000
##                 york          york_finger                young
##           0.07024294           0.00000000           0.00000000
##             zealand                 zest                zesti
##           0.00000000           0.00000000           0.00000000
##            zinfandel
##           0.00000000
```

here is the resut:

```
train_svmLinear_model <- train(x= wine_train_set, y=train$quality , method = 'svmLinear3')
train_svmLinear_model
```

```
## L2 Regularized Support Vector Machine (dual) with Linear Kernel
##
## 20795 samples
##   664 predictor
##     2 classes: 'excellent', 'good'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 20795, 20795, 20795, 20795, 20795, 20795, ...
## Resampling results across tuning parameters:
##
##   cost  Loss  Accuracy   Kappa
##   0.25  L1    0.7703370  0.5387221
##   0.25  L2    0.7789838  0.5563933
##   0.50  L1    0.7750863  0.5486402
##   0.50  L2    0.7797305  0.5580057
##   1.00  L1    0.7775631  0.5537664
##   1.00  L2    0.7796604  0.5579795
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were cost = 0.5 and Loss = L2.
```

```
model_svmLinear_result <- predict(train_svmLinear_model, newdata = wine_test_set)

conf_svmLinear_train <- table(model_svmLinear_result, wine_testing_result)
names(dimnames(conf_svmLinear_train)) <- c("Predicted class", "Actual class")
confusionMatrix(conf_svmLinear_train)
```

```
## Confusion Matrix and Statistics
##
##                Actual class
## Predicted class excellent good
##       excellent      1574  365
##       good            866 2394
##
##                Accuracy : 0.7632
##                  95% CI : (0.7514, 0.7747)
##     No Information Rate : 0.5307
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5189
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.6451
##             Specificity : 0.8677
##          Pos Pred Value : 0.8118
##          Neg Pred Value : 0.7344
##              Prevalence : 0.4693
```

```
##          Detection Rate : 0.3028
##    Detection Prevalence : 0.3730
##       Balanced Accuracy : 0.7564
##
##        'Positive' Class : excellent
##
```

Or duoble the point for adj:

```
#remove double point for nouns
wine_train_set[,column_id] <- wine_train_set[,column_id]/2
wine_test_set[,column_id] <- wine_test_set[,column_id]/2

#multify for adj
for (i in 1:dim(wine_train_set)[2]) {
  check <- colnames(wine_train_set)[i] %in% adj
  if(check)
    {
      column_id <- c(column_id, i)
    }
}

wine_train_set[,column_id] <- wine_train_set[,column_id]*2
wine_test_set[,column_id] <- wine_test_set[,column_id]*2

wine_train_set[1,]
```

```
##               accent                 acid                  add
##           0.00000000           0.00000000           0.00000000
##               africa             aftertast                  age
##           0.00000000           0.00000000           0.00000000
##            age_drink              alcohol                almond
##           0.00000000           0.00000000           0.00000000
##               almost                along              alongsid
##           0.00000000           0.05780700           0.00000000
##              alreadi                alsac           alsac_alsac
##           0.00000000           0.00000000           0.00000000
##                 also              although                 ampl
##           0.00000000           0.00000000           0.00000000
##                 anis               appeal                 appl
##           0.00000000           0.00000000           0.00000000
##            appl_pear              approach              apricot
##           0.00000000           0.00000000           0.00000000
##            argentina                aroma          aroma_flavor
##           0.00000000           0.00000000           0.00000000
##           aroma_lead               aromat               around
##           0.00000000           0.00000000           0.00000000
##              astring              attract            australia
##           0.00000000           0.00000000           0.00000000
##              austria                 back                 bake
##           0.00000000           0.00000000           0.00000000
##           bake_spice               balanc               barolo
##           0.00000000           0.00000000           0.00000000
```

```
##       barolo_nebbiolo               barrel                 bean
##            0.00000000           0.00000000           0.00000000
##                beauti                berri           berri_aroma
##            0.00000000           0.07412781           0.00000000
##          berri_flavor          berri_fruit                 best
##            0.00000000           0.00000000           0.00000000
##                better                  big                  bit
##            0.00000000           0.00000000           0.00000000
##                bitter                black         black_cherri
##            0.00000000           0.00000000           0.00000000
##         black_currant          black_fruit         black_pepper
##            0.00000000           0.00000000           0.00000000
##            black_plum           blackberri    blackberri_cherri
##            0.00000000           0.00000000           0.00000000
##                 blanc                blend       blend_cabernet
##            0.00000000           0.05101848           0.00000000
##               blossom                 blue            blueberri
##            0.00000000           0.00000000           0.00000000
##                 boast                 bodi                 bold
##            0.00000000           0.00000000           0.00000000
##              bordeaux    bordeaux_bordeaux          bordeauxstyl
##            0.00000000           0.00000000           0.00000000
##      bordeauxstyl_red                bottl              bouquet
##            0.00000000           0.10178298           0.00000000
##                bright          bright_acid                bring
##            0.00000000           0.00000000           0.00000000
##                 brisk             brunello          brunello_di
##            0.00000000           0.00000000           0.00000000
##                  brut              burgundi               butter
##            0.00000000           0.00000000           0.00000000
##                   cab              cabernet       cabernet_franc
##            0.00000000           0.07375056           0.06528019
##    cabernet_sauvignon           california california_california
##            0.00000000           0.00000000           0.00000000
##       california_napa       california_paso    california_russian
##            0.00000000           0.00000000           0.00000000
##      california_santa     california_sonoma                  can
##            0.00000000           0.00000000           0.00000000
##                 candi              caramel                carri
##            0.00000000           0.00000000           0.00000000
##                 cassi                cedar               cellar
##            0.00000000           0.00000000           0.00000000
##                central         central_coast        central_valley
##            0.00000000           0.00000000           0.00000000
##              champagn        champagn_blend     champagn_champagn
##            0.00000000           0.00000000           0.00000000
##                   char              charact           chardonnay
##            0.00000000           0.00000000           0.08252304
##                 cherri         cherri_flavor          cherri_fruit
##            0.00000000           0.00000000           0.00000000
##      cherri_raspberri                chewi              chianti
##            0.00000000           0.00000000           0.00000000
##                 chile                chocol                chunki
##            0.00000000           0.00000000           0.00000000
```

```
##             cinnamon               citrus             citrusi
##           0.00000000           0.00000000           0.00000000
##              classic              classico                clean
##           0.00000000           0.00000000           0.00000000
##                close                clove                coast
##           0.00000000           0.00000000           0.00000000
##     coast_chardonnay          coast_pinot         coast_sonoma
##           0.00000000           0.00000000           0.00000000
##                cocoa                coffe                cola
##           0.00000000           0.00000000           0.00000000
##                color              columbia       columbia_valley
##           0.00000000           0.00000000           0.00000000
##               combin                 come              complex
##           0.00000000           0.00000000           0.00000000
##             concentr                 cool                core
##           0.00000000           0.00000000           0.00000000
##                counti        counti_central             cranberri
##           0.00000000           0.00000000           0.00000000
##                cream                creami                creek
##           0.00000000           0.00000000           0.00000000
##                crisp           crisp_acid                 cru
##           0.00000000           0.00000000           0.00000000
##          cru_burgundi                crush                ctes
##           0.00000000           0.00000000           0.00000000
##              ctes_de              currant                 cut
##           0.00000000           0.00000000           0.00000000
##                cuve                 dark          dark_chocol
##           0.00000000           0.00000000           0.00000000
##                deep                  del                delic
##           0.00000000           0.00000000           0.00000000
##               delici                deliv                dens
##           0.00000000           0.00000000           0.00000000
##                depth               despit              develop
##           0.00000000           0.00000000           0.00000000
##         di_montalcino               doesnt                domin
##           0.00000000           0.00000000           0.00000000
##                  dri              dri_herb                drink
##           0.00000000           0.00000000           0.00000000
##           drink_franc          drink_itali           drink_now
##           0.00000000           0.00000000           0.00000000
##          drink_portug                dusti                earth
##           0.00000000           0.00000000           0.00000000
##               earthi                 easi                 edg
##           0.00000000           0.00000000           0.00000000
##                 eleg              element                  end
##           0.00000000           0.00000000           0.00000000
##                enjoy               enough             espresso
##           0.00000000           0.00000000           0.00000000
##                estat      estat_california                 even
##           0.00000000           0.00000000           0.00000000
##                excel                 exot              express
##           0.00000000           0.00000000           0.00000000
##                extra              extract                 fair
##           0.00000000           0.00000000           0.00000000
```

```
##              famili              featur                feel
##          0.00000000          0.00000000          0.00000000
##             ferment                fill                find
##          0.00000000          0.00000000          0.00000000
##                fine              finger         finger_lake
##          0.00000000          0.00000000          0.00000000
##              finish         finish_drink         finish_itali
##          0.00000000          0.00000000          0.00000000
##           finish_us               firm          firm_tannin
##          0.00000000          0.00000000          0.00000000
##               first              flavor        flavor_finish
##          0.00000000          0.00000000          0.00000000
##           flavor_us              fleshi               floral
##          0.00000000          0.00000000          0.12763861
##              flower               focus              follow
##          0.00000000          0.00000000          0.00000000
##                food             foothil              forest
##          0.00000000          0.00000000          0.00000000
##             forward            fragrant               frame
##          0.00000000          0.00000000          0.00000000
##                franc       franc_bordeaux           franc_les
##          0.05821335          0.00000000          0.00000000
##              french           french_oak               fresh
##          0.00000000          0.00000000          0.06628334
##           fresh_acid               front               fruit
##          0.00000000          0.00000000          0.03348005
##         fruit_flavor              fruiti                full
##          0.00000000          0.08738315          0.00000000
##            full_bodi            fullbodi             generous
##          0.00000000          0.00000000          0.00000000
##                gentl             germani                 get
##          0.00000000          0.00000000          0.00000000
##                give               glass                good
##          0.00000000          0.00000000          0.00000000
##               grand               grape          grapefruit
##          0.00000000          0.00000000          0.00000000
##             graphit               great               green
##          0.00000000          0.00000000          0.09683225
##           green_appl             grenach              grigio
##          0.00000000          0.00000000          0.00000000
##               grill                grip                gris
##          0.00000000          0.00000000          0.00000000
##               grner         grner_veltlin               grown
##          0.00000000          0.00000000          0.00000000
##                hard               heavi                herb
##          0.00000000          0.00000000          0.08983105
##              herbal                here                high
##          0.00000000          0.00000000          0.00000000
##                hill                hint                hold
##          0.00000000          0.00000000          0.00000000
##               honey             impress              includ
##          0.00000000          0.00000000          0.00000000
##              integr              intens            interest
##          0.00000000          0.05463712          0.00000000
```

```
##              intrigu                itali         itali_tuscani
##           0.00000000           0.00000000           0.00000000
##                  jam                jammi                 juic
##           0.00000000           0.00000000           0.00000000
##                juici                 just                 keep
##           0.00000000           0.00000000           0.00000000
##                 lack                 lake           lake_finger
##           0.00000000           0.00000000           0.00000000
##                 last                layer                 lead
##           0.00000000           0.00000000           0.00000000
##            lead_nose                 leaf                 lean
##           0.00000000           0.00000000           0.00000000
##              leather                 leav                lemon
##           0.00000000           0.00000000           0.00000000
##               lemoni                 lend               length
##           0.00000000           0.00000000           0.00000000
##                  les              licoric                 lift
##           0.00000000           0.00000000           0.00000000
##                light                 like                 lime
##           0.00000000           0.00000000           0.00000000
##               linger         linger_finish                 littl
##           0.00000000           0.00000000           0.00000000
##                 live                 load                 loir
##           0.00000000           0.00000000           0.00000000
##          loir_valley                 long           long_finish
##           0.00000000           0.10835301           0.00000000
##                  lot                 love                 lush
##           0.00000000           0.00000000           0.00000000
##                 made                 make               malbec
##           0.00000000           0.00000000           0.00000000
##                mango                 mani                 mark
##           0.00000000           0.00000000           0.00000000
##                matur                 meat               medium
##           0.00000000           0.00000000           0.00000000
##           mediumbodi                melon              mendoza
##           0.00000000           0.00000000           0.00000000
##       mendoza_provinc               merlot       merlot_cabernet
##           0.00000000           0.10222301           0.08115250
##               midpal                 mild                miner
##           0.00000000           0.00000000           0.00000000
##                 mint                  mix                mocha
##           0.00000000           0.00000000           0.00000000
##                moder            montalcino  montalcino_sangioves
##           0.00000000           0.00000000           0.00000000
##                month             mountain                mouth
##           0.00000000           0.00000000           0.00000000
##            mouthfeel                 much                 napa
##           0.11864741           0.00000000           0.00000000
##         napa_cabernet           napa_valley                natur
##           0.00000000           0.00000000           0.00000000
##              nebbiolo              nectarin                need
##           0.00000000           0.00000000           0.00000000
##                  new             new_york           new_zealand
##           0.09831095           0.07032841           0.00000000
```

```
##                next                nice                noir
##           0.00000000           0.00000000           0.00000000
##               north         northeastern   northeastern_itali
##           0.07979934           0.00000000           0.00000000
##             northern       northern_spain                nose
##           0.00000000           0.00000000           0.07304711
##           nose_palat                note                 now
##           0.00000000           0.03116226           0.00000000
##            now_franc              now_us               nuanc
##           0.00000000           0.00000000           0.00000000
##                  nut                  oak                oaki
##           0.00000000           0.00000000           0.00000000
##                offer                  old            old_vine
##           0.03822461           0.00000000           0.00000000
##                 oliv                  one                open
##           0.00000000           0.00000000           0.00000000
##           open_aroma                 opul               orang
##           0.00000000           0.00000000           0.00000000
##               oregon     oregon_willamett               overal
##           0.00000000           0.00000000           0.00000000
##                 pack                 pair                palat
##           0.00000000           0.00000000           0.04409815
##          palat_deliv          palat_offer          palat_show
##           0.00000000           0.00000000           0.00000000
##                 paso            paso_robl                peach
##           0.00000000           0.00000000           0.00000000
##                 pear                 peel               pepper
##           0.00000000           0.00000000           0.00000000
##              pepperi              perfect               perfum
##           0.00000000           0.00000000           0.00000000
##              persist                petit          petit_sirah
##           0.00000000           0.00000000           0.00000000
##         petit_verdot                  pie            piedmont
##           0.00000000           0.00000000           0.00000000
##      piedmont_barolo             pineappl                 pink
##           0.00000000           0.00000000           0.00000000
##                pinot          pinot_grigio           pinot_gris
##           0.00000000           0.00000000           0.00000000
##           pinot_noir                 play             pleasant
##           0.00000000           0.00000000           0.14752514
##               plenti                 plum                plump
##           0.00000000           0.00000000           0.00000000
##                plush               polish               portug
##           0.00000000           0.00000000           0.00000000
##             portugues        portugues_red              potenti
##           0.00000000           0.00000000           0.00000000
##                power              premier          premier_cru
##           0.00000000           0.00000000           0.00000000
##               pretti                price                produc
##           0.06690534           0.00000000           0.00000000
##              provenc               provid              provinc
##           0.00000000           0.00000000           0.00000000
##      provinc_mendoza                prune                 pure
##           0.00000000           0.00000000           0.00000000
```

```
##             qualiti                quit                raci
##          0.00000000          0.00000000          0.00000000
##              raisin            raspberri     raspberri_cherri
##          0.00000000          0.00000000          0.00000000
##              rather                 raw               readi
##          0.00000000          0.00000000          0.00000000
##         readi_drink                 red           red_berri
##          0.00000000          0.04865018          0.06928811
##           red_blend           red_cherri         red_currant
##          0.00000000          0.00000000          0.00000000
##           red_fruit              refresh              region
##          0.00000000          0.00000000          0.00000000
##              remain               reserv    reserv_california
##          0.00000000          0.00000000          0.00000000
##             reserva               reveal                rhne
##          0.00000000          0.00000000          0.00000000
##            rhnestyl         rhnestyl_red                rich
##          0.00000000          0.00000000          0.00000000
##               riesl               right               rioja
##          0.00000000          0.00000000          0.00000000
##               ripe           ripe_fruit             riserva
##          0.00000000          0.00000000          0.00000000
##              river         river_valley               roast
##          0.00000000          0.00000000          0.00000000
##               robl         robl_central                 ros
##          0.00000000          0.00000000          0.11377897
##               rose               round             russian
##          0.00000000          0.00000000          0.00000000
##        russian_river              rustic                sage
##          0.00000000          0.00000000          0.00000000
##           sangioves               santa             sardinia
##          0.00000000          0.00000000          0.00000000
##           sauvignon      sauvignon_blanc               savori
##          0.00000000          0.00000000          0.00000000
##               scent                seem              select
##          0.00000000          0.00000000          0.00000000
##               sens                 set               sharp
##          0.00000000          0.00000000          0.00000000
##               show              sicili      sicili_sardinia
##          0.00000000          0.00000000          0.00000000
##               side              sierra       sierra_foothil
##          0.00000000          0.00000000          0.00000000
##               silki               simpl                 sip
##          0.00000000          0.00000000          0.00000000
##               sirah                skin              slight
##          0.00000000          0.00000000          0.00000000
##               smell               smoke               smoki
##          0.00000000          0.00000000          0.00000000
##              smooth                soft              soften
##          0.00000000          0.00000000          0.00000000
##               soil               solid            somewhat
##          0.00000000          0.00000000          0.00000000
##              sonoma    sonoma_chardonnay         sonoma_coast
##          0.00000000          0.00000000          0.00000000
```

```
##         sonoma_counti          sonoma_pinot                 soon
##            0.00000000            0.00000000            0.00000000
##                   sour                 sourc                 south
##            0.00000000            0.00000000            0.00000000
##          south_africa       south_australia              southern
##            0.00000000            0.00000000            0.00000000
##             southwest       southwest_franc                 spain
##            0.00000000            0.00000000            0.00000000
##           spain_rioja                sparkl          sparkl_blend
##            0.00000000            0.00000000            0.00000000
##                 spice          spice_flavor                 spici
##            0.00000000            0.00000000            0.00000000
##                 start                 still                 stone
##            0.00000000            0.00000000            0.00000000
##           stone_fruit       straightforward           strawberri
##            0.00000000            0.00000000            0.00000000
##                streak                strong              structur
##            0.00000000            0.00000000            0.00000000
##                 style                 subtl                 sugar
##            0.00000000            0.00000000            0.00000000
##               suggest              superior                suppl
##            0.00000000            0.00000000            0.00000000
##               support                 sweet                 syrah
##            0.00000000            0.00000000            0.00000000
##                  take               tangerin                 tangi
##            0.00000000            0.00000000            0.00000000
##                tannic                tannin          tannin_drink
##            0.00000000            0.00000000            0.00000000
##                 tart                  tast                   tea
##            0.00000000            0.00000000            0.00000000
##           tempranillo                textur                  that
##            0.00000000            0.00000000            0.00000000
##                 there                 thick                though
##            0.00000000            0.00000000            0.00000000
##                 tight                  time                 toast
##            0.00000000            0.00000000            0.00000000
##                toasti                tobacco                togeth
##            0.00000000            0.00000000            0.00000000
##                tomato                  tone                   top
##            0.00000000            0.00000000            0.00000000
##               toscana                 touch                tropic
##            0.00000000            0.00000000            0.00000000
##          tropic_fruit                  turn               tuscani
##            0.00000000            0.00000000            0.00000000
##      tuscani_brunello       tuscani_chianti                   two
##            0.00000000            0.00000000            0.00000000
##             underbrush         us_california              us_estat
##            0.00000000            0.00000000            0.00000000
##              us_oregon             us_reserv         us_washington
##            0.00000000            0.00000000            0.00000000
##                valley       valley_cabernet        valley_central
##            0.00000000            0.00000000            0.00000000
##     valley_chardonnay           valley_napa          valley_pinot
##            0.00000000            0.00000000            0.00000000
```

```
##            valley_red         valley_sonoma          valley_syrah
##            0.00000000            0.00000000            0.00000000
##             valley_wa      valley_willamett               vanilla
##            0.00000000            0.00000000            0.00000000
##                variet               varieti               veltlin
##            0.00000000            0.00000000            0.00000000
##               velveti                veneto                verdot
##            0.00000000            0.00000000            0.00000000
##                  veri               vibrant                  vine
##            0.00000000            0.00000000            0.00000000
##              vineyard   vineyard_california   vineyard_washington
##            0.00000000            0.00000000            0.00000000
##                vintag              viognier                violet
##            0.00000000            0.00000000            0.00000000
##           wa_columbia                  walla          walla_valley
##            0.00000000            0.00000000            0.00000000
##           walla_walla                  warm            washington
##            0.00000000            0.00000000            0.00000000
##   washington_columbia                   way                weight
##            0.00000000            0.00000000            0.00000000
##                  well                   wet                 whiff
##            0.00000000            0.00000000            0.00000000
##                 white           white_blend           white_peach
##            0.00000000            0.00000000            0.00000000
##          white_pepper                  wild                  will
##            0.00000000            0.00000000            0.00000000
##             willamett      willamett_valley                  wine
##            0.00000000            0.00000000            0.00000000
##               winemak                wineri               without
##            0.00000000            0.00000000            0.00000000
##                  wood               wood_age                  wrap
##            0.00000000            0.00000000            0.00000000
##                  year                yellow                   yet
##            0.00000000            0.00000000            0.00000000
##                  york           york_finger                 young
##            0.07024294            0.00000000            0.00000000
##               zealand                  zest                 zesti
##            0.00000000            0.00000000            0.00000000
##             zinfandel
##            0.00000000
```

And here is the result:

```
train_svmLinear_model <- train(x= wine_train_set, y=train$quality , method = 'svmLinear3')
train_svmLinear_model
```

```
## L2 Regularized Support Vector Machine (dual) with Linear Kernel
##
## 20795 samples
##   664 predictor
##     2 classes: 'excellent', 'good'
##
## No pre-processing
```

```
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 20795, 20795, 20795, 20795, 20795, 20795, ...
## Resampling results across tuning parameters:
##
##   cost  Loss  Accuracy   Kappa
##   0.25  L1    0.7743984  0.5470603
##   0.25  L2    0.7805501  0.5595244
##   0.50  L1    0.7768020  0.5520431
##   0.50  L2    0.7807815  0.5601102
##   1.00  L1    0.7782605  0.5551175
##   1.00  L2    0.7806293  0.5598400
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were cost = 0.5 and Loss = L2.
```

```
model_svmLinear_result <- predict(train_svmLinear_model, newdata = wine_test_set)

conf_svmLinear_train <- table(model_svmLinear_result, wine_testing_result)
names(dimnames(conf_svmLinear_train)) <- c("Predicted class", "Actual class")
confusionMatrix(conf_svmLinear_train)
```

```
## Confusion Matrix and Statistics
##
##                Actual class
## Predicted class excellent good
##       excellent      1559  361
##       good            881 2398
##
##                Accuracy : 0.7611
##                  95% CI : (0.7493, 0.7726)
##     No Information Rate : 0.5307
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5144
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.6389
##             Specificity : 0.8692
##          Pos Pred Value : 0.8120
##          Neg Pred Value : 0.7313
##              Prevalence : 0.4693
##          Detection Rate : 0.2999
##    Detection Prevalence : 0.3693
##       Balanced Accuracy : 0.7540
##
##        'Positive' Class : excellent
##
```

Here is the final result if I try with 100% of data:

```
# split train/test
n = dim(wine)[1]
set.seed(12345)
```

```r
id = sample(1:n, floor(n*0.8))
train = wine[id,]
test = wine[-id,]

wine_train_set <- clean(train$description)

train_dtm_tfidf <- DocumentTermMatrix(wine_train_set, control = list(weighting = weightTfIdf))
#train_dtm_tfidf <- DocumentTermMatrix(wine_train_set, control = list( tokenize=NLP_tokenizer))
#train_dtm_tfidf <- DocumentTermMatrix(wine_train_set)
train_dtm_tfidf <- removeSparseTerms(train_dtm_tfidf, 0.99)

#create the test set
wine_test_set <- clean(test$description)
wine_test_set <- DocumentTermMatrix(wine_test_set, control = list(dictionary = Terms(train_dtm_tfidf) ,
#wine_test_set <- DocumentTermMatrix(wine_test_set, control = list(dictionary = Terms(train_dtm_tfidf)

#create matrix for training
wine_train_set <<- as.matrix(train_dtm_tfidf)
wine_test_set <- as.matrix(wine_test_set)
wine_test_set <- wine_test_set[,Terms(train_dtm_tfidf)]
#create the test result
wine_testing_result <- test$quality

train_svmLinear_model <- train(x= wine_train_set, y=train$quality , method = 'svmLinear3')
train_svmLinear_model
```

```
## L2 Regularized Support Vector Machine (dual) with Linear Kernel
##
## 103976 samples
##    504 predictor
##      2 classes: 'excellent', 'good'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 103976, 103976, 103976, 103976, 103976, 103976, ...
## Resampling results across tuning parameters:
##
##   cost  Loss  Accuracy   Kappa
##   0.25  L1    0.7910057  0.5808259
##   0.25  L2    0.7919573  0.5826771
##   0.50  L1    0.7912370  0.5813173
##   0.50  L2    0.7920083  0.5827954
##   1.00  L1    0.7915443  0.5819473
##   1.00  L2    0.7920722  0.5829360
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were cost = 1 and Loss = L2.
```

```r
model_svmLinear_result <- predict(train_svmLinear_model, newdata = wine_test_set)

conf_svmLinear_train <- table(model_svmLinear_result, wine_testing_result)
names(dimnames(conf_svmLinear_train)) <- c("Predicted class", "Actual class")
confusionMatrix(conf_svmLinear_train)
```

```
## Confusion Matrix and Statistics
##
##                Actual class
## Predicted class excellent  good
##       excellent      8548  1905
##       good           3753 11789
##
##                Accuracy : 0.7823
##                  95% CI : (0.7773, 0.7873)
##     No Information Rate : 0.5268
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5601
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.6949
##             Specificity : 0.8609
##          Pos Pred Value : 0.8178
##          Neg Pred Value : 0.7585
##              Prevalence : 0.4732
##          Detection Rate : 0.3288
##    Detection Prevalence : 0.4021
##       Balanced Accuracy : 0.7779
##
##        'Positive' Class : excellent
##
```

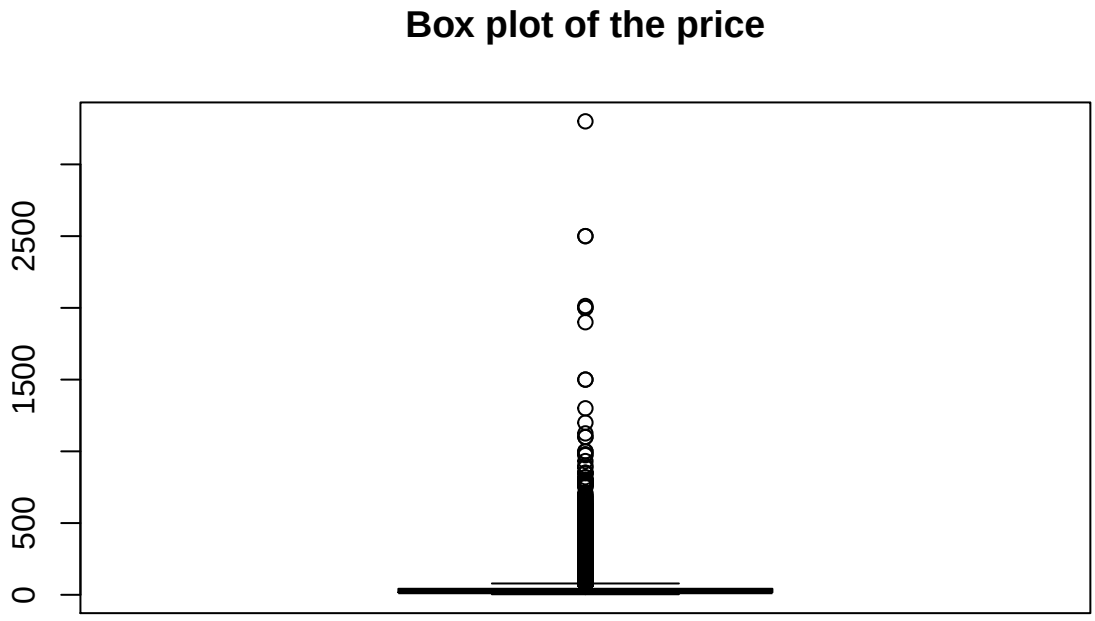
Result....

second aspect: high value wine

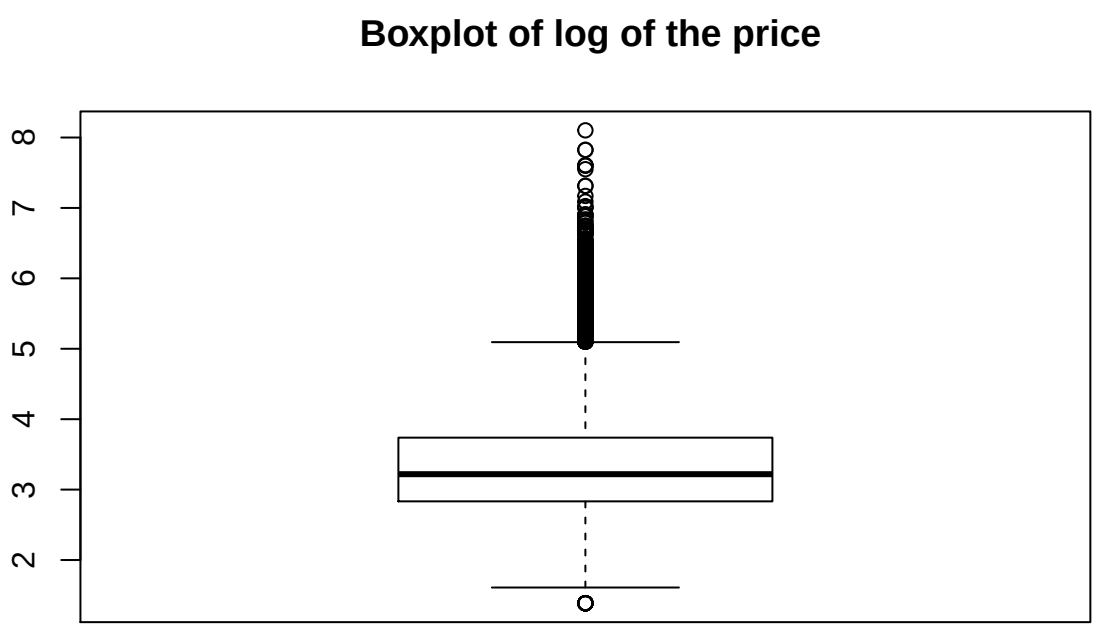
The problem

In most of time, the price of the bottle can reflect the quatity of the wine. For example, If you want to buy an excellent wine, choose the 1000$ bottle. That method should work. But, most of us don't have the financial condition for buying like this. The point is if we have a limit amount of money, or we just don't want to pay too much for a bottle, how can we choose? It's my idea for the second problem. Let's take a look at the price of our data:

```
boxplot(wine$price, main = "Box plot of the price")
```




```
boxplot(log(wine$price), main = " Boxplot of log of the price")
```

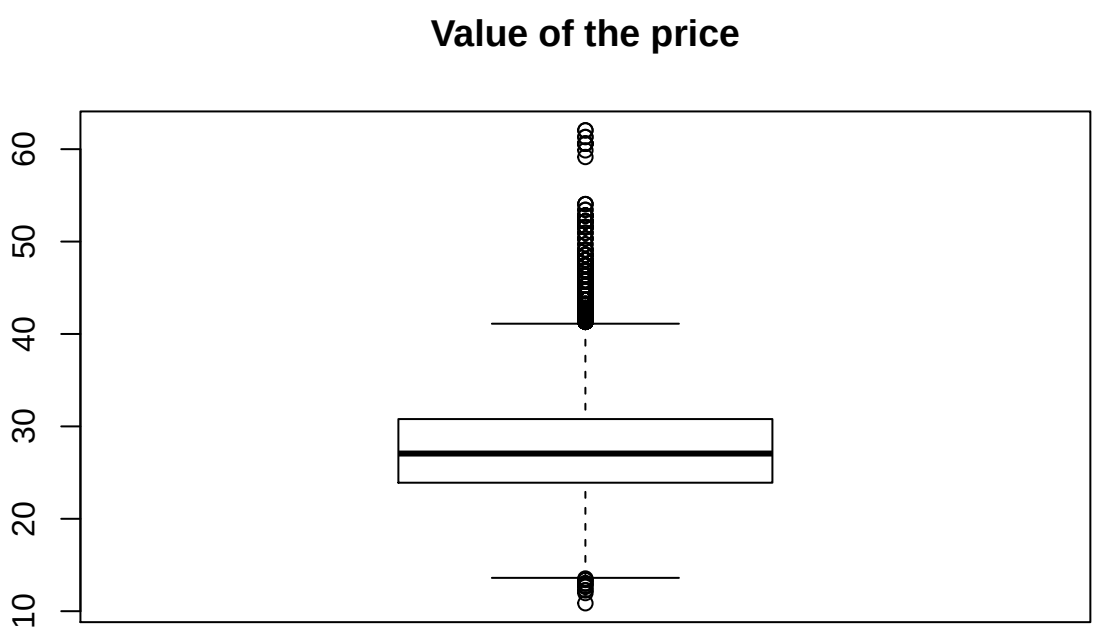


Because the price range quite big, have a lot of outlier and high value of SD. otherwise, the log of the price look better and quite similar with the point.

Aim

I define the value = point/log(price). Here is the value.

```
boxplot(wine$value, main = "Value of the price")
```



The medium number for value is ~27. So the bottle with higher value than 27 will have "high" benefit. The rest have "medium" benefit. This way of define class will make all data become interesting. For example. there is a bottle with just 4$ but have 80 point. it become the bottle with highest value of benefit. In constrast. many bottle with the price higher than 1000 dollar. Of couse they are excellent wine. But is just have medium value because it's too expensive.

I do the same preprocessing for the text. Then try with Nb and SVM linear, here is the result.

Futher discusssion

Suggession for choosing wine

future work: ## Conclusion ## References ## Apendix