732A92 Text mining project

Evaluation wine based on descriptions

Student: Duc Duong (mindu931)

Instructor: prof. Marco Kuhlmann

Linköping, 16th January 2019

Contents

Abstract

Wine sensory examination and assessment is not easy, even with a wine specialist. However, the description printed on each bottle can bring us some helpful information. Based on that ideal, I decided to analyze the data about wine review (taken form Kaggle) and help the consumers make the decision on two different aspects. The first one is distinguish good and excellent wine. The second aspect is investigating the finance beneficial bottle (high value with an acceptable amount of money). Different natural language processing techniques are applied to process the text data. Then I compare Naive Bayes, support vector machine (SVM) Linear and SVM RBF kernels when building the best model for predicting. After the best kernel is selected, n-gram and Part-Of-Speech Tagger are used to improve the model. The final results are quite impressive with accuracy 78.69% for the first issue and 77.09% for the second one. Based on these models, some suggestions are provided for the consumers to choose satisfied bottles.

1. Introduction

1.1 Motivation

While standing in a wine cellar, I was always confusing with a lot of questions: How can I choose an excellent wine for my family? What is the best wine fitted with my budget? The problems could be solved easier if I had a good sense. But, sadly, like many people in the world, I don't. The problems still continue when I listen to the advice of the salesman and even taste some wine. So, I usually choose randomly a bottle based on this advice for my events. Unluckily, sometimes my relatives didn't think that it was an excellent wine.

1.2 Aim

I guess that the information from the provider (description, country, region, designation...) should bring some important clues. So, I try to solve the issues based on text mining idea: Take the wine data which is already reviewed and marked by some specialists, then process the data together with the points 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.

1.3 Outline

In this report, Firstly, I will provide information about the data. Then, I briefly summary relevant theories. The method I used will be presented in details. It includes the steps for preprocessing, techniques for comparing and choosing the best model. Finally, there are some suggestions for choosing wine, some discussions and my conclusion.

2. Data

Wine review data is taken form Kaggle (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 data:

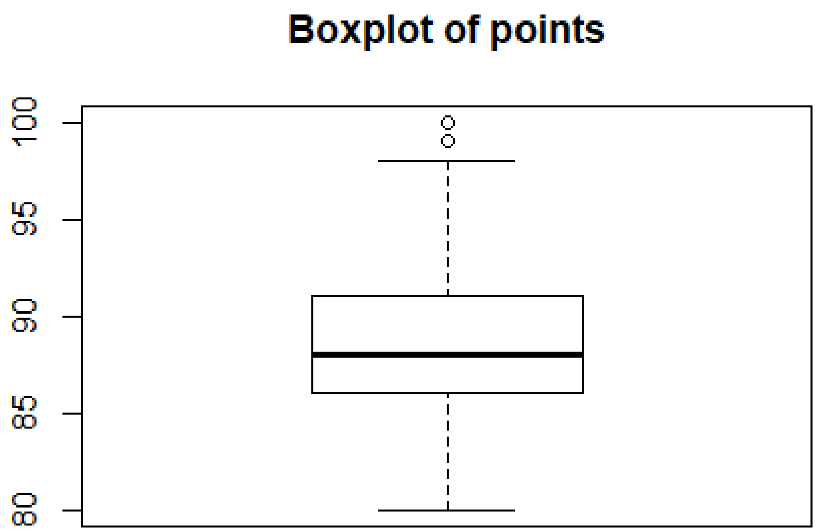
```
head(wine)

##   X  country
## 1 0    Italy
## 2 1 Portugal
## 3 2       US
## 4 3       US
## 5 4       US
## 6 5    Spain
##
description
## 1
Aromas include tropical fruit, broom, brimstone and dried herb. The palate is
n't overly expressive, offering unripened apple, citrus and dried sage alongs
ide brisk acidity.
## 2                                   This is ripe and fruity, a wine that i
s smooth while still structured. Firm tannins are filled out with juicy red b
erry fruits and freshened with acidity. It's  already drinkable, although it
will certainly be better from 2016.
## 3
Tart and snappy, the flavors of lime flesh and rind dominate. Some green pine
apple pokes through, with crisp acidity underscoring the flavors. The wine wa
s all stainless-steel fermented.
## 4                                                               Pineapple
rind, lemon pith and orange blossom start off the aromas. The palate is a bit
more opulent, with notes of honey-drizzled guava and mango giving way to a sl
ightly astringent, semidry finish.
## 5             Much like the regular bottling from 2012, this comes across
as rather rough and tannic, with rustic, earthy, herbal characteristics. None
theless, if you think of it as a pleasantly unfussy country wine, it's a good
companion to a hearty winter stew.
## 6 Blackberry and raspberry aromas show a typical Navarran whiff of green h
erbs and, in this case, horseradish. In the mouth, this is fairly full bodied
, with tomatoey acidity. Spicy, herbal flavors complement dark plum fruit, wh
ile the finish is fresh but grabby.
##                          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  Bia
nco  (Etna)
## 2                                       Quinta dos Avidagos 2011 Avidagos
Red (Douro)
## 3                                       Rainstorm 2013 Pinot Gris (Willame
tte Valley)
## 4                 St. Julian 2013 Reserve Late Harvest Riesling (Lake Mich
igan Shore)
## 5 Sweet Cheeks 2012 Vintner's Reserve Wild Child Block Pinot Noir (Willame
tte Valley)
## 6                                   Tandem 2011 Ars In Vitro Tempranillo-Merlo
t (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 are important information in this data. Points are marked by wine specialists, which make it reliable. The owner only public data which have at least 80 points (out of 100) which means the data contain only good and excellent wine. Here is the distribution of the points:

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



3. Theory

In this section, I will briefly talk about some theory in natural language processing (NLP) and machine learning (ML) in a simple way of reviewing. For some terms, only the aspect that relevant to this project is presented. I assume that the reader already has some basic knowledge 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 written in English.

- Terms in text preprocessing:
  - Stop words: This is the term for useless words in the text. To be specific, these words usually have no (or very low) meaning but represent a lot. For example: a, an, the… When processing text data. We usually remove these words.

  - Steaming: This is the process of reducing a word to its root form. For example: processes, processing and processed are 3 different words, but it just 3 different representatives of the word "process". Another example is evaluated and evaluation.

  - Tokenization: the process of splitting text into smaller parts. Each part can consider as a feature when training in machine learning kernels. 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 training. For example, all words in the bag of words counts as a corpus. Each element in a corpus can be considered as a feature.

- Document term matrix (DTM) is a matrix, which has columns are all features of the corpus and rows is the document ID. The cross between row and column is the point for that word in that document. Points can be calculated 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 formula:

$$tfidf(t, d, D) = tf(t, d) * idf(t, D)$$

$$tf(t, d) = \frac{f_{t,d}}{\sum f_{t\prime,d(t\prime \in d)}}$$

$$idf(t, D) = log \frac{|D|}{|d \in D: t \in d|}$$

$|d \in D: t \in d|$ number of documents where the term t appears. If this value equal 0, it will be adjust to 1.
D: total number of documents.

- Machine learning kernels: a kernel can understand as a core algorithm used when training a model in ML. Here are the three kernels that I use in this report:

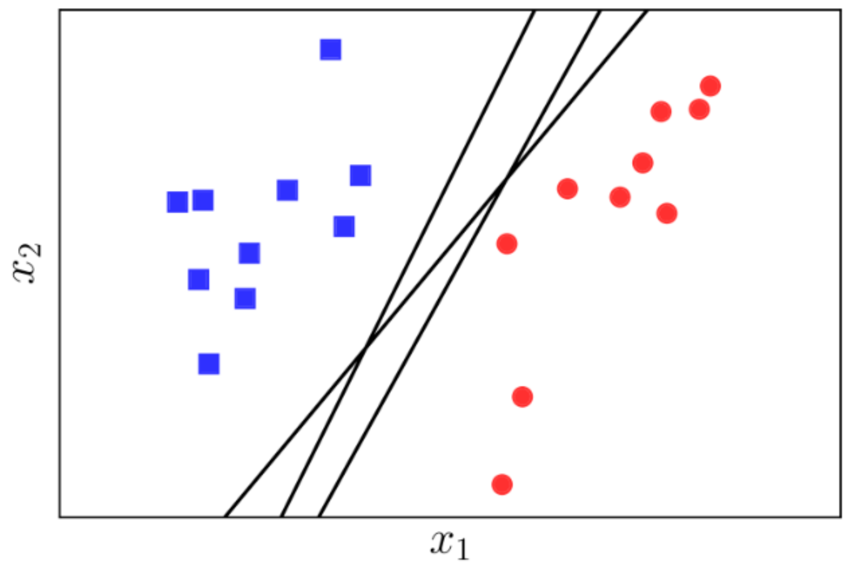
  - Naive Bayes (NB): Let initial with we have a vector X of a document, which built in document-term matrix. The probability of assign a class label y to that vector is:

$$P(X = (x_i, \ldots x_P)|Y = y) = \prod_{i=1}^{p} P\,(X_i = x_i|Y = y)$$

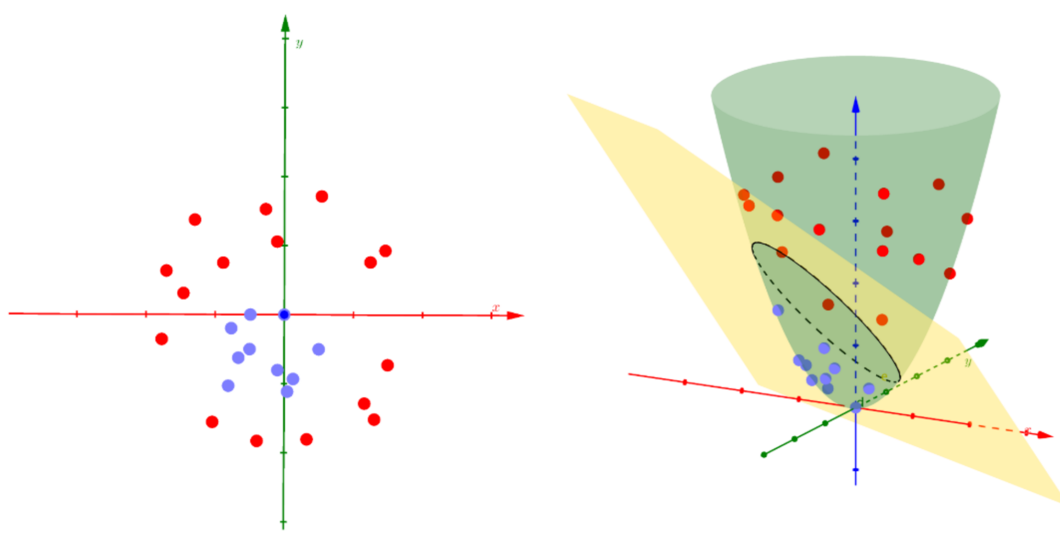
  Naive-Bayes theory make an assumption that all features are independent. The assumption is not realistic. Somehow, it works well.

  - SVM: Stands for support vector machine. In the previous kernel, we have vector X for each document. Image that each document is represent by vector X in a multi-dimension space. (The number of dimension is the number of features). The aim is finding a hyperplane to seperate the space into two classes.

  - SVM linear and RBF kenels: SVM Linear will decide the hyperplane as a flat. SVM RBF will use an other dimension to solve the problem. Let's see an example of how linear and RBF kerels work in some pictures:



SVM linear kernel [5]



SVM RBF kernel [6]

- Ngram: n-gram is n continue sequence words. In this project, n-gram model is used for making corpus, which means features are sequence of n words. An example is n=2 continue words will be features. So, the sentence: "this is a sentence" will have "this_is", "is_a", "a_sentence" as features.
- Part-Of-Speech Tagger (POS Tagger): This is the method to process text and decided each word is nouns, ajdtive, adverb…

4. Method

In this part, I would like to discuss about the method for the first problem of this project: Which is the “good” and “excellent” wine. The second problem about wine that has “high” and “medium” value is solved using a similar method but will be discussed in details in a different part.

- Define the class: The first step is to define the class for each document. Base on the distribution of the point, I will take the point 88 as the boundary because it is the median value, which will make the data become balance. So, the wine that has higher than 88 will be the excellent wine, while the rest is the good wine.
- Preprocess data: I use 80% of the data as the training set and 20% as the testing set. The training data is processed as follow:
  - Make the text ready: I decided that the information about the country, designation, province, region, and wine variety is important. So, I decided to merge that information into the description as a paragraph. That paragraph is the one that I will process.
  - Language convert[3]: In the data set. There is some variety that is not unified (same type but written in the different language). So I convert it to the most famous words 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 on the list. Then, I add three words “the”, “and” and “wine” to the stop word list because, in this data set, it doesn’t have any meaning. Here is the final stop word list:

```
> 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 the package SnowballC.
- Tokenizing and making the bag of words as the corpus. Then I only keep 99% spare term, which mean that only the words that appear in at least 1% of documents are kept. It makes sense because there are some rare words which only appear in one or a small number of 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 steps is similar, except the DTM. DTM of testing data will be create based on the corpus of training data (which means some words that don't appear in the training data will be drop)

- Compare kernels: At this step, I'll use the ready DTM for training model, then test with the testing DTM. Three kernel Naive-Bayes, SVM linear and SVM RBF are used to compare. Because of the limitation in my resource (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 samples for testing) to compute. Different parameters used for training is also reported. Notice that the time for training NB and SVM linear is quite fast (a few minutes) but It takes a long time for training SVM RBF (~14 hours, for 505 terms in DTM, include time for tunning model)

- Improve the best model: After comparing, the best model is selected. Then I continue to improve the model by using two methods:
  - Using n-gram model.
  - Adjust the weight for some terms using Part-Of-Speech Tagger (POS Tagger).

5. Result and explain

5.1 Preprocess text

Firstly, let's take a look at the first original paragraph and the paragraphs after preprocessing to see how it works:

```
train$description[1]

## [1] "This smells mostly of oak and barrel spice, with barely any fruit see
ping through all the wood grain that's on display. Plump on the palate, this o
veroaked Chardonnay tastes almost entirely of resin and spice. Throw in clove
and vanilla flavors on the finish, and you get the picture. Chile Duette Casa
blanca Valley   Chardonnay"

wine_train_set[[1]]$content

## [1] "smell most oak barrel spice bare fruit seep wood grain that display p
lump palat overoak chardonnay tast almost entir resin spice throw clove vanil
la flavor finish get pictur chile duett casablanca valley chardonnay"
```

Then, Here is the DTM for the first document of training set:

```
wine_train_set[1,]

##         accent         acid       across          add
##     0.00000000   0.00000000   0.00000000   0.00000000
##         africa    aftertast          age      alcohol
##     0.00000000   0.00000000   0.00000000   0.00000000
##         almond       almost        along     alongsid
##     0.00000000   0.15504157   0.00000000   0.00000000
##        alreadi        alsac         also     although
##     0.00000000   0.00000000   0.00000000   0.00000000
##           ampl         anis        anoth       appeal
##     0.00000000   0.00000000   0.00000000   0.00000000
##           appl     approach      apricot    argentina
##     0.00000000   0.00000000   0.00000000   0.00000000
##          aroma       aromat       around      astring
##     0.00000000   0.00000000   0.00000000   0.00000000
##        attract    australia      austria         back
##     0.00000000   0.00000000   0.00000000   0.00000000
```

```
##         bake        balanc       barbara        barolo
##   0.00000000    0.00000000    0.00000000    0.00000000
##       barrel        beauti         berri          best
##   0.15305057    0.00000000    0.00000000    0.00000000
##       better           big           bit        bitter
##   0.00000000    0.00000000    0.00000000    0.00000000
##        black    blackberri         blanc         blend
##   0.00000000    0.00000000    0.00000000    0.00000000
##      blossom          blue     blueberri          bodi
##   0.00000000    0.00000000    0.00000000    0.00000000
##         bold      bordeaux  bordeauxstyl         bottl
##   0.00000000    0.00000000    0.00000000    0.00000000
##      bouquet   boysenberri        brambl        bright
##   0.00000000    0.00000000    0.00000000    0.00000000
##        bring         brisk          brut      burgundi
##   0.00000000    0.00000000    0.00000000    0.00000000
##       butter           cab      cabernet    california
##   0.00000000    0.00000000    0.00000000    0.00000000
##          can         candi       caramel       carnero
##   0.00000000    0.00000000    0.00000000    0.00000000
##        carri         cassi     catalonia         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    0.19943433    0.00000000    0.00000000
##        chile        chocol        chunki      cinnamon
##   0.14711002    0.00000000    0.00000000    0.00000000
##       citrus       citrusi       classic      classico
##   0.00000000    0.00000000    0.00000000    0.00000000
##        clean         close         clove         coast
##   0.00000000    0.00000000    0.15208806    0.00000000
##        cocoa         coffe          cola         color
##   0.00000000    0.00000000    0.00000000    0.00000000
##     columbia        combin          come       complex
##   0.00000000    0.00000000    0.00000000    0.00000000
##     concentr          cool          core        counti
##   0.00000000    0.00000000    0.00000000    0.00000000
##    cranberri        creami         creek         crisp
##   0.00000000    0.00000000    0.00000000    0.00000000
##          cru         crush          ctes       currant
##   0.00000000    0.00000000    0.00000000    0.00000000
##          cut          cuve          dark          deep
##   0.00000000    0.00000000    0.00000000    0.00000000
##          del         delic        delici         deliv
##   0.00000000    0.00000000    0.00000000    0.00000000
##         dens         depth        despit       develop
##   0.00000000    0.00000000    0.00000000    0.00000000
##       doesnt         domin           dri         drink
##   0.00000000    0.00000000    0.00000000    0.00000000
```

```
##        dusti       earth      earthi        easi
##   0.00000000  0.00000000  0.00000000  0.00000000
##          edg        eleg     element         end
##   0.00000000  0.00000000  0.00000000  0.00000000
##        enjoy      enough    espresso       estat
##   0.00000000  0.00000000  0.00000000  0.00000000
##         even       excel        exot     express
##   0.00000000  0.00000000  0.00000000  0.00000000
##        extra     extract        fair      famili
##   0.00000000  0.00000000  0.00000000  0.00000000
##       featur        feel     ferment        fill
##   0.00000000  0.00000000  0.00000000  0.00000000
##         find        fine      finger      finish
##   0.00000000  0.00000000  0.00000000  0.04989804
##         firm       first      flavor      fleshi
##   0.00000000  0.00000000  0.02956396  0.00000000
##       floral      flower       focus      follow
##   0.00000000  0.00000000  0.00000000  0.00000000
##         food     foothil      forest     forward
##   0.00000000  0.00000000  0.00000000  0.00000000
##     fragrant       frame       franc      french
##   0.00000000  0.00000000  0.00000000  0.00000000
##        fresh       front       fruit      fruiti
##   0.00000000  0.00000000  0.04106600  0.00000000
##         full    fullbodi    generous       gentl
##   0.00000000  0.00000000  0.00000000  0.00000000
##      germani         get        give       glass
##   0.00000000  0.18456806  0.00000000  0.00000000
##         good       grand       grape  grapefruit
##   0.00000000  0.00000000  0.00000000  0.00000000
##      graphit       great       green     grenach
##   0.00000000  0.00000000  0.00000000  0.00000000
##        grill        grip        gris       grown
##   0.00000000  0.00000000  0.00000000  0.00000000
##         hard       heavi        herb      herbal
##   0.00000000  0.00000000  0.00000000  0.00000000
##         high   highlight        hill        hint
##   0.00000000  0.00000000  0.00000000  0.00000000
##         hold       honey  honeysuckl         hot
##   0.00000000  0.00000000  0.00000000  0.00000000
##      impress      includ      integr      intens
##   0.00000000  0.00000000  0.00000000  0.00000000
##     interest     intrigu       invit       itali
##   0.00000000  0.00000000  0.00000000  0.00000000
##          jam       jammi        juic       juici
##   0.00000000  0.00000000  0.00000000  0.00000000
##         just        keep        lack        lake
##   0.00000000  0.00000000  0.00000000  0.00000000
##         last       layer        lead        leaf
##   0.00000000  0.00000000  0.00000000  0.00000000
```

```
##          lean         least       leather          leav 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##         lemon          lend        length           les 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##       licoric          lift         light          like 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##          lime        linger         littl          live 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##          load          loir          long           lot 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##          love          lush          made          make 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##        malbec         mango          mani          mark 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##         matur          meat        medium    mediumbodi 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##         melon       mendoza        merlot        midpal 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##          mild         miner          mint           mix 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##         mocha         moder    montalcino         month 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##      mountain       mourvdr         mouth     mouthfeel 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##          much          napa    napasonoma         natur 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##      nebbiolo      nectarin          need           new 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##          next          nice          noir         north 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##  northeastern      northern          nose          note 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##           now         nuanc           oak          oaki 
##    0.00000000    0.00000000    0.08861248    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    0.05340088 
##          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         pinot          play      pleasant 
##    0.00000000    0.00000000    0.00000000    0.00000000 
##        plenti          plum         plump        polish 
##    0.00000000    0.00000000    0.19247727    0.00000000 
```

```
##       pomegran         portug      portugues        potenti 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##          power        present         pretti          price 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##         produc         provid        provinc          prune 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##           pure          purpl        qualiti           quit 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##           raci         raisin          ranch      raspberri 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##         rather          readi            red        refresh 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##         region         remain         reserv        reserva 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##         reveal       rhnestyl           rich          riesl 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##          right          rioja           ripe        riserva 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##          river          roast           robl            ros 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##           rose          round        russian         rustic 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##           sage      sangioves          santa       sardinia 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##      sauvignon         savori          scent           seem 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##         select           sens            set          sharp 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##           show         sicili           side         sierra 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##          silki          simpl            sip          sirah 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##           skin         slight          smell          smoke 
##     0.00000000     0.00000000     0.17657168     0.00000000 
##          smoki         smooth           soft         soften 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##           soil          solid       somewhat         sonoma 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##           soon           sour          sourc          south 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##       southern      southwest          spain         sparkl 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##          spice          spici          start          still 
##     0.15574770     0.00000000     0.00000000     0.00000000 
##          stone straightforward     strawberri         streak 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##         strong       structur          style          subtl 
##     0.00000000     0.00000000     0.00000000     0.00000000 
##          sugar        suggest       superior          suppl 
##     0.00000000     0.00000000     0.00000000     0.00000000 
```

```
##        support       sweet       syrah        take
##     0.00000000  0.00000000  0.00000000  0.00000000
##       tangerin       tangi      tannic      tannin
##     0.00000000  0.00000000  0.00000000  0.00000000
##           tart        tast         tea tempranillo
##     0.00000000  0.12918828  0.00000000  0.00000000
##         textur        that       there       thick
##     0.00000000  0.15647508  0.00000000  0.00000000
##         though       tight        time       toast
##     0.00000000  0.00000000  0.00000000  0.00000000
##         toasti     tobacco      togeth      tomato
##     0.00000000  0.00000000  0.00000000  0.00000000
##           tone     toscana       touch      tropic
##     0.00000000  0.00000000  0.00000000  0.00000000
##           turn     tuscani         two  underbrush
##     0.00000000  0.00000000  0.00000000  0.00000000
##         valley     vanilla      variet     varieti
##     0.05228140  0.11012388  0.00000000  0.00000000
##        velveti      veneto      verdot        veri
##     0.00000000  0.00000000  0.00000000  0.00000000
##        vibrant        vine    vineyard      vintag
##     0.00000000  0.00000000  0.00000000  0.00000000
##       viognier      violet        warm  washington
##     0.00000000  0.00000000  0.00000000  0.00000000
##            way      weight        well         wet
##     0.00000000  0.00000000  0.00000000  0.00000000
##          whiff       white        wild        will
##     0.00000000  0.00000000  0.00000000  0.00000000
##      willamett        wine     winemak      wineri
##     0.00000000  0.00000000  0.00000000  0.00000000
##        without      wonder        wood        wrap
##     0.00000000  0.00000000  0.14000828  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
```

And let see the first document of DTM for testing set to make sure that the prepared data is correct. As we can see, the terms still remain. Just the points are different.

```
wine_test_set[1,]
```

```
##         accent        acid      across         add
##     0.00000000  0.00000000  0.00000000  0.00000000
##         africa  aftertast         age     alcohol
##     0.00000000  0.00000000  0.00000000  0.00000000
##         almond      almost       along    alongsid
##     0.00000000  0.00000000  0.16027711  0.00000000
```

```
##        alreadi         alsac          also      although
##     0.00000000    0.00000000    0.00000000    0.00000000
##           ampl          anis         anoth        appeal
##     0.00000000    0.00000000    0.00000000    0.00000000
##           appl      approach       apricot     argentina
##     0.00000000    0.00000000    0.00000000    0.00000000
##          aroma        aromat        around       astring
##     0.05810157    0.00000000    0.00000000    0.00000000
##        attract     australia       austria          back
##     0.00000000    0.00000000    0.00000000    0.00000000
##           bake        balanc       barbara        barolo
##     0.00000000    0.00000000    0.00000000    0.00000000
##         barrel        beauti         berri          best
##     0.00000000    0.00000000    0.00000000    0.00000000
##         better           big           bit        bitter
##     0.00000000    0.00000000    0.00000000    0.00000000
##          black    blackberri         blanc         blend
##     0.08717125    0.00000000    0.00000000    0.00000000
##        blossom          blue     blueberri          bodi
##     0.00000000    0.00000000    0.00000000    0.00000000
##           bold      bordeaux  bordeauxstyl         bottl
##     0.00000000    0.00000000    0.00000000    0.14176587
##        bouquet   boysenberri        brambl        bright
##     0.00000000    0.00000000    0.00000000    0.00000000
##          bring         brisk          brut      burgundi
##     0.00000000    0.00000000    0.00000000    0.00000000
##         butter           cab      cabernet    california
##     0.00000000    0.00000000    0.00000000    0.05886771
##            can         candi       caramel       carnero
##     0.00000000    0.00000000    0.00000000    0.00000000
##          carri         cassi     catalonia         cedar
##     0.00000000    0.00000000    0.00000000    0.00000000
##         cellar       central      champagn          char
##     0.00000000    0.10262244    0.00000000    0.00000000
##        charact    chardonnay        cherri         chewi
##     0.00000000    0.00000000    0.00000000    0.00000000
##          chile        chocol        chunki      cinnamon
##     0.00000000    0.00000000    0.00000000    0.00000000
##         citrus       citrusi       classic      classico
##     0.00000000    0.00000000    0.00000000    0.00000000
##          clean         close         clove         coast
##     0.00000000    0.00000000    0.00000000    0.10710576
##          cocoa         coffe          cola         color
##     0.00000000    0.00000000    0.00000000    0.16347060
##       columbia        combin          come       complex
##     0.00000000    0.00000000    0.00000000    0.00000000
##       concentr          cool          core        counti
##     0.00000000    0.00000000    0.00000000    0.00000000
##      cranberri        creami         creek         crisp
##     0.18309503    0.00000000    0.00000000    0.00000000
```

```
##        cru      crush       ctes    currant
## 0.00000000 0.19399253 0.00000000 0.00000000
##        cut       cuve       dark       deep
## 0.00000000 0.00000000 0.00000000 0.00000000
##        del      delic     delici      deliv
## 0.00000000 0.00000000 0.00000000 0.00000000
##       dens      depth     despit    develop
## 0.00000000 0.00000000 0.00000000 0.00000000
##     doesnt      domin        dri      drink
## 0.00000000 0.00000000 0.00000000 0.00000000
##      dusti      earth     earthi       easi
## 0.00000000 0.00000000 0.00000000 0.00000000
##        edg       eleg    element        end
## 0.00000000 0.00000000 0.00000000 0.00000000
##      enjoy     enough   espresso      estat
## 0.00000000 0.00000000 0.00000000 0.00000000
##       even      excel       exot    express
## 0.00000000 0.00000000 0.00000000 0.00000000
##      extra    extract       fair     famili
## 0.00000000 0.00000000 0.00000000 0.00000000
##     featur       feel    ferment       fill
## 0.00000000 0.00000000 0.00000000 0.00000000
##       find       fine     finger     finish
## 0.00000000 0.00000000 0.00000000 0.00000000
##       firm      first     flavor     fleshi
## 0.00000000 0.00000000 0.03361649 0.00000000
##     floral     flower      focus     follow
## 0.00000000 0.00000000 0.00000000 0.00000000
##       food    foothil     forest    forward
## 0.00000000 0.00000000 0.00000000 0.00000000
##   fragrant      frame      franc     french
## 0.00000000 0.00000000 0.00000000 0.00000000
##      fresh      front      fruit     fruiti
## 0.00000000 0.00000000 0.00000000 0.00000000
##       full   fullbodi   generous      gentl
## 0.00000000 0.00000000 0.00000000 0.00000000
##    germani        get       give      glass
## 0.00000000 0.00000000 0.00000000 0.00000000
##       good      grand      grape grapefruit
## 0.00000000 0.00000000 0.15286684 0.00000000
##    graphit      great      green    grenach
## 0.00000000 0.00000000 0.00000000 0.00000000
##      grill       grip       gris      grown
## 0.00000000 0.00000000 0.00000000 0.00000000
##       hard      heavi       herb     herbal
## 0.00000000 0.00000000 0.00000000 0.00000000
##       high  highlight       hill       hint
## 0.00000000 0.00000000 0.00000000 0.00000000
##       hold      honey honeysuckl        hot
## 0.00000000 0.00000000 0.00000000 0.00000000
```

```
##         impress        includ        integr        intens
##      0.00000000    0.00000000    0.00000000    0.00000000
##        interest       intrigu         invit         itali
##      0.00000000    0.00000000    0.00000000    0.00000000
##             jam         jammi          juic         juici
##      0.00000000    0.00000000    0.00000000    0.00000000
##            just          keep          lack          lake
##      0.00000000    0.00000000    0.00000000    0.00000000
##            last         layer          lead          leaf
##      0.00000000    0.00000000    0.00000000    0.00000000
##            lean         least       leather          leav
##      0.00000000    0.00000000    0.00000000    0.00000000
##           lemon          lend        length           les
##      0.00000000    0.00000000    0.00000000    0.00000000
##         licoric          lift         light          like
##      0.00000000    0.00000000    0.21708525    0.00000000
##            lime        linger         littl          live
##      0.00000000    0.00000000    0.00000000    0.00000000
##            load          loir          long           lot
##      0.00000000    0.00000000    0.00000000    0.00000000
##            love          lush          made          make
##      0.00000000    0.00000000    0.00000000    0.00000000
##          malbec         mango          mani          mark
##      0.00000000    0.00000000    0.00000000    0.00000000
##           matur          meat        medium    mediumbodi
##      0.00000000    0.00000000    0.00000000    0.00000000
##           melon       mendoza        merlot        midpal
##      0.00000000    0.00000000    0.00000000    0.00000000
##            mild         miner          mint           mix
##      0.00000000    0.00000000    0.00000000    0.00000000
##           mocha         moder    montalcino         month
##      0.00000000    0.00000000    0.00000000    0.00000000
##        mountain       mourvdr         mouth     mouthfeel
##      0.00000000    0.00000000    0.00000000    0.00000000
##            much          napa    napasonoma         natur
##      0.00000000    0.00000000    0.00000000    0.00000000
##        nebbiolo       nectarin          need           new
##      0.00000000    0.00000000    0.00000000    0.00000000
##            next          nice          noir         north
##      0.00000000    0.00000000    0.00000000    0.00000000
##    northeastern      northern          nose          note
##      0.00000000    0.00000000    0.00000000    0.00000000
##             now         nuanc           oak          oaki
##      0.00000000    0.00000000    0.00000000    0.00000000
##           offer           old          oliv           one
##      0.10675293    0.20729912    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    0.06026340
```

```
##           paso        peach        pear        peel
##     0.00000000   0.00000000   0.00000000   0.00000000
##         pepper      pepperi      perfect      perfum
##     0.12844115   0.00000000   0.00000000   0.00000000
##        persist        petit          pie    piedmont
##     0.00000000   0.00000000   0.00000000   0.00000000
##       pineappl        pinot         play    pleasant
##     0.00000000   0.00000000   0.00000000   0.00000000
##         plenti         plum        plump      polish
##     0.00000000   0.00000000   0.00000000   0.00000000
##       pomegran       portug    portugues     potenti
##     0.00000000   0.00000000   0.00000000   0.00000000
##          power      present       pretti       price
##     0.00000000   0.00000000   0.00000000   0.00000000
##         produc       provid      provinc       prune
##     0.00000000   0.00000000   0.00000000   0.00000000
##           pure        purpl      qualiti        quit
##     0.00000000   0.00000000   0.00000000   0.00000000
##           raci       raisin        ranch   raspberri
##     0.00000000   0.00000000   0.00000000   0.00000000
##         rather        readi          red     refresh
##     0.00000000   0.00000000   0.00000000   0.00000000
##         region       remain       reserv     reserva
##     0.00000000   0.00000000   0.00000000   0.00000000
##         reveal     rhnestyl         rich       riesl
##     0.00000000   0.00000000   0.00000000   0.00000000
##          right        rioja         ripe     riserva
##     0.00000000   0.00000000   0.00000000   0.00000000
##          river        roast         robl         ros
##     0.00000000   0.00000000   0.00000000   0.00000000
##           rose        round      russian      rustic
##     0.18517684   0.00000000   0.00000000   0.00000000
##           sage    sangioves        santa    sardinia
##     0.00000000   0.00000000   0.00000000   0.00000000
##      sauvignon       savori        scent        seem
##     0.00000000   0.00000000   0.00000000   0.00000000
##         select         sens          set       sharp
##     0.00000000   0.00000000   0.00000000   0.00000000
##           show       sicili         side      sierra
##     0.09596939   0.00000000   0.00000000   0.00000000
##          silki        simpl          sip       sirah
##     0.00000000   0.00000000   0.00000000   0.00000000
##           skin      slight        smell       smoke
##     0.00000000   0.00000000   0.00000000   0.00000000
##          smoki       smooth         soft      soften
##     0.16658973   0.00000000   0.00000000   0.00000000
##           soil        solid     somewhat      sonoma
##     0.00000000   0.00000000   0.00000000   0.00000000
##           soon         sour        sourc       south
##     0.00000000   0.00000000   0.00000000   0.00000000
```

```
##        southern       southwest           spain          sparkl
##      0.00000000      0.00000000      0.00000000      0.00000000
##           spice           spici           start           still
##      0.00000000      0.00000000      0.00000000      0.00000000
##           stone straightforward      strawberri          streak
##      0.00000000      0.00000000      0.00000000      0.00000000
##          strong        structur           style           subtl
##      0.00000000      0.00000000      0.00000000      0.00000000
##           sugar         suggest        superior           suppl
##      0.00000000      0.00000000      0.00000000      0.00000000
##         support           sweet           syrah            take
##      0.00000000      0.00000000      0.27086967      0.00000000
##        tangerin           tangi          tannic          tannin
##      0.00000000      0.00000000      0.00000000      0.00000000
##            tart            tast             tea     tempranillo
##      0.00000000      0.00000000      0.20864375      0.00000000
##          textur            that           there           thick
##      0.10817950      0.00000000      0.00000000      0.00000000
##          though           tight            time           toast
##      0.00000000      0.00000000      0.00000000      0.00000000
##          toasti         tobacco          togeth          tomato
##      0.00000000      0.00000000      0.00000000      0.00000000
##            tone         toscana           touch          tropic
##      0.00000000      0.00000000      0.00000000      0.00000000
##            turn         tuscani             two      underbrush
##      0.00000000      0.00000000      0.00000000      0.00000000
##          valley         vanilla          variet         varieti
##      0.00000000      0.00000000      0.00000000      0.00000000
##         velveti         veneto          verdot            veri
##      0.00000000      0.00000000      0.00000000      0.12898784
##         vibrant            vine        vineyard          vintag
##      0.00000000      0.00000000      0.09436784      0.00000000
##        viognier          violet            warm      washington
##      0.00000000      0.19008872      0.00000000      0.00000000
##             way          weight            well             wet
##      0.00000000      0.00000000      0.00000000      0.00000000
##           whiff           white            wild            will
##      0.00000000      0.11183840      0.00000000      0.00000000
##       willamett            wine         winemak          wineri
##      0.00000000      0.00000000      0.00000000      0.00000000
##         without          wonder            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
```

5.2 Compare kernels

Now, the data is ready. I train that dataset with three different kernels as mentioned. And here are the results:

```
train_nb_model
```

```
## Naive Bayes
##
## 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:
##
##   usekernel  Accuracy   Kappa
##   FALSE      0.7513089  0.5033206
##    TRUE      0.5757021  0.1881251
##
## 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.
```

```
confusionMatrix(conf_nb_train)
```

```
## Confusion Matrix and Statistics
##
##                Actual class
## Predicted class excellent good
##       excellent      1525  487
##       good            738 2089
##
##                Accuracy : 0.7468
##                  95% CI : (0.7343, 0.7591)
##     No Information Rate : 0.5323
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4881
##  Mcnemar's Test P-Value : 9.141e-13
##
##             Sensitivity : 0.6739
##             Specificity : 0.8109
##          Pos Pred Value : 0.7580
##          Neg Pred Value : 0.7389
```

```
##               Prevalence : 0.4677
##           Detection Rate : 0.3151
##     Detection Prevalence : 0.4158
##        Balanced Accuracy : 0.7424
##
##         'Positive' Class : excellent
##
```

```
train_svmLinear_model
```

```
## L2 Regularized Support Vector Machine (dual) with Linear 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:
##
##   cost  Loss  Accuracy   Kappa
##   0.25  L1    0.7843732  0.5673696
##   0.25  L2    0.7861278  0.5709545
##   0.50  L1    0.7854490  0.5696725
##   0.50  L2    0.7854986  0.5697548
##   1.00  L1    0.7853391  0.5694828
##   1.00  L2    0.7847845  0.5683523
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were cost = 0.25 and Loss = L2.
```

```
confusionMatrix(conf_svmLinear_train)
```

```
## Confusion Matrix and Statistics
##
##                  Actual class
## Predicted class excellent good
##       excellent      1676  449
##       good            587 2127
##
##                Accuracy : 0.7859
##                  95% CI : (0.7741, 0.7974)
##     No Information Rate : 0.5323
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5684
##  Mcnemar's Test P-Value : 2.078e-05
##
##             Sensitivity : 0.7406
##             Specificity : 0.8257
```

```
##          Pos Pred Value : 0.7887
##          Neg Pred Value : 0.7837
##              Prevalence : 0.4677
##          Detection Rate : 0.3464
##    Detection Prevalence : 0.4391
##       Balanced Accuracy : 0.7832
##
##        'Positive' Class : excellent
##
```

```
> 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.
```

SVMRBF kenel

```
> 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
```

As we can see in the result. The accuracy when training data for NB, SVM Linear and SVM RBF is 75.1%, 78.6%, and 81.1%. As a result, I expected that the accuracy when testing with test data will be similar. But they are 74.68% for NB, 78.59% for SVM Linear and 71.36%. SVM Linear is the kernel which has the highest value of accuracy for testing. Noticed that No information rate (NIR) 0.5323 mean that a class takes 53.23% ("good" class), which mean the data is balanced. We can judge that the models are actually worked.

But, accuracy is just one side of the story. Let see about the classification: We have two classes "good" and "excellent". If the class is predicted exactly, 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 accept (similar to false negative-FN). In contrast, If you pretend to buy good wine but have excellent wine. You were just lucky and nothing happened. (similar to false positive-FP)

SVM RBF kernel 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.

After all, I consider the accuracy, the number of each class in confusion matrix and the training time for choosing the best kernel. In my opinion, **SVM Linear is the best kernel**. (SVM RBF is interesting but it's hard when I try to improve the model with that high training time, consider the scope of this project)

5.3 Improve the model

I'll try to improve the SVM linear model with the following methods:

- Improve with n-gram: Firstly, I try to train the model with 2-gram, (the preprocess still remain) here is the first line of DTM:

```
wine_train_set[1,]
```

```
##                acid_us              age_drink             alsac_alsac
##              0.0000000              0.0000000               0.0000000
##              appl_pear           aroma_flavor              aroma_lead
##              0.0000000              0.0000000               0.0000000
##             bake_spice       barolo_nebbiolo             berri_aroma
##              0.0000000              0.0000000               0.0000000
##            berri_flavor            berri_fruit            black_cherri
##              0.0000000              0.0000000               0.0000000
##          black_currant            black_fruit            black_pepper
##              0.0000000              0.0000000               0.0000000
##             black_plum     blackberri_cherri          blend_cabernet
##              0.0000000              0.0000000               0.0000000
##      bordeaux_bordeaux       bordeauxstyl_red             bright_acid
##              0.0000000              0.0000000               0.0000000
##         cabernet_franc     cabernet_sauvignon california_california
##              0.0000000              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           counti_sonoma
##              0.0000000              0.0000000               0.0000000
##             crisp_acid                ctes_de             dark_chocol
##              0.0000000              0.0000000               0.0000000
##             dark_fruit          di_montalcino                dri_herb
##              0.0000000              0.0000000               0.0000000
##            drink_franc            drink_itali               drink_now
##              0.0000000              0.0000000               0.0000000
##           drink_portug       estat_california             estat_grown
##              0.0000000              0.0000000               0.0000000
##            finger_lake           finish_drink            finish_itali
##              0.0000000              0.0000000               0.0000000
##              finish_us            firm_tannin       flavor_blackberri
##              0.0000000              0.0000000               0.0000000
##          flavor_finish              flavor_us          franc_bordeaux
##              0.1711132              0.0000000               0.0000000
##             french_oak             fresh_acid            fruit_flavor
##              0.0000000              0.0000000               0.0000000
##              full_bodi             green_appl          itali_tuscani
##              0.0000000              0.0000000               0.0000000
```

```
##          lake_finger           lead_nose       linger_finish
##            0.0000000           0.0000000           0.0000000
##          loir_valley         long_finish         medium_bodi
##            0.0000000           0.0000000           0.0000000
##      mendoza_provinc montalcino_sangioves        napa_cabernet
##            0.0000000           0.0000000           0.0000000
##          napa_valley            new_york         new_zealand
##            0.0000000           0.0000000           0.0000000
##   northeastern_itali       northern_spain          nose_palat
##            0.0000000           0.0000000           0.0000000
##            now_franc              now_us            old_vine
##            0.0000000           0.0000000           0.0000000
##           open_aroma     oregon_willamett         palat_deliv
##            0.0000000           0.0000000           0.0000000
##          palat_offer          palat_show           paso_robl
##            0.0000000           0.0000000           0.0000000
##          petit_sirah        petit_verdot     piedmont_barolo
##            0.0000000           0.0000000           0.0000000
##           pinot_gris          pinot_noir        portugues_red
##            0.0000000           0.0000000           0.0000000
##      provinc_mendoza     raspberri_cherri         readi_drink
##            0.0000000           0.0000000           0.0000000
##            red_berri           red_blend          red_cherri
##            0.0000000           0.0000000           0.0000000
##          red_currant           red_fruit   reserv_california
##            0.0000000           0.0000000           0.0000000
##         rhnestyl_red          ripe_fruit        river_valley
##            0.0000000           0.0000000           0.0000000
##         robl_central        russian_river        santa_barbara
##            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        us_california            us_estat
##            0.0000000           0.0000000           0.0000000
##            us_oregon            us_reserv        us_washington
##            0.0000000           0.0000000           0.0000000
##      valley_cabernet       valley_central   valley_chardonnay
##            0.0000000           0.0000000           0.1922822
##          valley_napa         valley_pinot          valley_red
##            0.0000000           0.0000000           0.0000000
##        valley_sonoma         valley_syrah           valley_wa
##            0.0000000           0.0000000           0.0000000
```

```
##      valley_willamett   vineyard_california   vineyard_washington
##             0.0000000             0.0000000             0.0000000
##           wa_columbia   washington_columbia           white_blend
##             0.0000000             0.0000000             0.0000000
##           white_peach          white_pepper      willamett_valley
##             0.0000000             0.0000000             0.0000000
##              wood_age               year_us           york_finger
##             0.0000000             0.0000000             0.0000000
```

Here are the results:

```
train_svmLinear_model

## L2 Regularized Support Vector Machine (dual) with Linear Kernel
##
## 19356 samples
##   156 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:
##
##   cost  Loss  Accuracy   Kappa
##   0.25  L1    0.6436272  0.2823927
##   0.25  L2    0.6494504  0.2968135
##   0.50  L1    0.6455226  0.2896008
##   0.50  L2    0.6498153  0.2976152
##   1.00  L1    0.6466283  0.2928574
##   1.00  L2    0.6499836  0.2979705
##
## 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.

confusionMatrix(conf_svmLinear_train)

## Confusion Matrix and Statistics
##
##                Actual class
## Predicted class excellent good
##       excellent      1200  668
##       good           1063 1908
##
##                Accuracy : 0.6423
##                  95% CI : (0.6286, 0.6558)
##     No Information Rate : 0.5323
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.2739
##  Mcnemar's Test P-Value : < 2.2e-16
```

```
##
##             Sensitivity : 0.5303
##             Specificity : 0.7407
##          Pos Pred Value : 0.6424
##          Neg Pred Value : 0.6422
##              Prevalence : 0.4677
##          Detection Rate : 0.2480
##    Detection Prevalence : 0.3860
##       Balanced Accuracy : 0.6355
##
##        'Positive' Class : excellent
##
```

The result is worse than the previous in all indicators, which means that 2-gram is not useful. Based on my observation. After preprocess and steaming, the descriptions are quite discrete. I mean the words are likely not connected to each other. It leads to the 2-gram model does not work.

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

```
wine_train_set_ngram[1,]
```

```
##                 accent                acid              acid_us
##             0.00000000          0.00000000           0.00000000
##                 across                 add               africa
##             0.00000000          0.00000000           0.00000000
##              aftertast                 age            age_drink
##             0.00000000          0.00000000           0.00000000
##                alcohol              almond               almost
##             0.00000000          0.00000000           0.07871341
##                  along            alongsid              alreadi
##             0.00000000          0.00000000           0.00000000
##                  alsac          alsac_alsac                 also
##             0.00000000          0.00000000           0.00000000
##               although                ampl                 anis
##             0.00000000          0.00000000           0.00000000
##                  anoth              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              barbara
##             0.00000000          0.00000000           0.00000000
##                 barolo     barolo_nebbiolo               barrel
```

```
##           0.00000000           0.00000000           0.07770259
##               beauti                berri          berri_aroma
##           0.00000000           0.00000000           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           0.00000000           0.00000000
##              blossom                 blue            blueberri
##           0.00000000           0.00000000           0.00000000
##                 bodi                 bold             bordeaux
##           0.00000000           0.00000000           0.00000000
##    bordeaux_bordeaux         bordeauxstyl     bordeauxstyl_red
##           0.00000000           0.00000000           0.00000000
##                bottl              bouquet          boysenberri
##           0.00000000           0.00000000           0.00000000
##               brambl               bright          bright_acid
##           0.00000000           0.00000000           0.00000000
##                bring                brisk                 brut
##           0.00000000           0.00000000           0.00000000
##             burgundi               butter                  cab
##           0.00000000           0.00000000           0.00000000
##             cabernet      cabernet_franc   cabernet_sauvignon
##           0.00000000           0.00000000           0.00000000
##           california california_california      california_napa
##           0.00000000           0.00000000           0.00000000
##      california_paso   california_russian     california_santa
##           0.00000000           0.00000000           0.00000000
##    california_sonoma                  can                candi
##           0.00000000           0.00000000           0.00000000
##              caramel              carnero                carri
##           0.00000000           0.00000000           0.00000000
##                cassi            catalonia                cedar
##           0.00000000           0.00000000           0.00000000
##               cellar              central        central_coast
##           0.00000000           0.00000000           0.00000000
##       central_valley             champagn       champagn_blend
##           0.00000000           0.00000000           0.00000000
##    champagn_champagn                 char              charact
##           0.00000000           0.00000000           0.00000000
##           chardonnay               cherri        cherri_flavor
##           0.10125128           0.00000000           0.00000000
##          cherri_fruit     cherri_raspberri                chewi
```

```
##          0.00000000          0.00000000          0.00000000
##               chile              chocol              chunki
##          0.07468663          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          0.07721394          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       counti_sonoma
##          0.00000000          0.00000000          0.00000000
##           cranberri              creami               creek
##          0.00000000          0.00000000          0.00000000
##               crisp          crisp_acid                 cru
##          0.00000000          0.00000000          0.00000000
##               crush                ctes             ctes_de
##          0.00000000          0.00000000          0.00000000
##             currant                 cut                cuve
##          0.00000000          0.00000000          0.00000000
##                dark         dark_chocol          dark_fruit
##          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       estat_grown
##          0.00000000        0.00000000        0.00000000
##                even             excel              exot
##          0.00000000        0.00000000        0.00000000
##             express             extra           extract
##          0.00000000        0.00000000        0.00000000
##                fair            famili            featur
##          0.00000000        0.00000000        0.00000000
##                feel           ferment              fill
##          0.00000000        0.00000000        0.00000000
##                find              fine            finger
##          0.00000000        0.00000000        0.00000000
##         finger_lake            finish      finish_drink
##          0.00000000        0.02533285        0.00000000
##        finish_itali         finish_us              firm
##          0.00000000        0.00000000        0.00000000
##         firm_tannin             first            flavor
##          0.00000000        0.00000000        0.01500940
##   flavor_blackberri     flavor_finish         flavor_us
##          0.00000000        0.08424033        0.00000000
##              fleshi            floral            flower
##          0.00000000        0.00000000        0.00000000
##               focus            follow              food
##          0.00000000        0.00000000        0.00000000
##             foothil            forest           forward
##          0.00000000        0.00000000        0.00000000
##            fragrant             frame             franc
##          0.00000000        0.00000000        0.00000000
##       franc_bordeaux            french        french_oak
##          0.00000000        0.00000000        0.00000000
##               fresh        fresh_acid             front
##          0.00000000        0.00000000        0.00000000
##               fruit      fruit_flavor            fruiti
##          0.02084889        0.00000000        0.00000000
##                full         full_bodi          fullbodi
##          0.00000000        0.00000000        0.00000000
##            generous             gentl           germani
##          0.00000000        0.00000000        0.00000000
##                 get              give             glass
##          0.09370379        0.00000000        0.00000000
##                good             grand             grape
##          0.00000000        0.00000000        0.00000000
##          grapefruit           graphit             great
##          0.00000000        0.00000000        0.00000000
##               green        green_appl           grenach
##          0.00000000        0.00000000        0.00000000
##               grill              grip              gris
##          0.00000000        0.00000000        0.00000000
##               grown              hard             heavi
```

```
##            0.00000000            0.00000000            0.00000000
##                  herb                herbal                  high
##            0.00000000            0.00000000            0.00000000
##             highlight                  hill                  hint
##            0.00000000            0.00000000            0.00000000
##                  hold                 honey            honeysuckl
##            0.00000000            0.00000000            0.00000000
##                   hot               impress                includ
##            0.00000000            0.00000000            0.00000000
##                integr                intens              interest
##            0.00000000            0.00000000            0.00000000
##               intrigu                 invit                 itali
##            0.00000000            0.00000000            0.00000000
##         itali_tuscani                   jam                 jammi
##            0.00000000            0.00000000            0.00000000
##                  juic                 juici                  just
##            0.00000000            0.00000000            0.00000000
##                  keep                  lack                  lake
##            0.00000000            0.00000000            0.00000000
##           lake_finger                  last                 layer
##            0.00000000            0.00000000            0.00000000
##                  lead             lead_nose                  leaf
##            0.00000000            0.00000000            0.00000000
##                  lean                 least               leather
##            0.00000000            0.00000000            0.00000000
##                  leav                 lemon                  lend
##            0.00000000            0.00000000            0.00000000
##                length                   les               licoric
##            0.00000000            0.00000000            0.00000000
##                  lift                 light                  like
##            0.00000000            0.00000000            0.00000000
##                  lime                linger         linger_finish
##            0.00000000            0.00000000            0.00000000
##                 littl                  live                  load
##            0.00000000            0.00000000            0.00000000
##                  loir           loir_valley                  long
##            0.00000000            0.00000000            0.00000000
##           long_finish                   lot                  love
##            0.00000000            0.00000000            0.00000000
##                  lush                  made                  make
##            0.00000000            0.00000000            0.00000000
##                malbec                 mango                  mani
##            0.00000000            0.00000000            0.00000000
##                  mark                 matur                  meat
##            0.00000000            0.00000000            0.00000000
##                medium           medium_bodi            mediumbodi
##            0.00000000            0.00000000            0.00000000
##                 melon               mendoza      mendoza_provinc
##            0.00000000            0.00000000            0.00000000
##                merlot                midpal                  mild
```

```
##             0.00000000          0.00000000          0.00000000
##                  miner                mint                 mix
##             0.00000000          0.00000000          0.00000000
##                  mocha               moder          montalcino
##             0.00000000          0.00000000          0.00000000
##   montalcino_sangioves               month            mountain
##             0.00000000          0.00000000          0.00000000
##                mourvdr               mouth           mouthfeel
##             0.00000000          0.00000000          0.00000000
##                   much                napa       napa_cabernet
##             0.00000000          0.00000000          0.00000000
##            napa_valley          napasonoma               natur
##             0.00000000          0.00000000          0.00000000
##               nebbiolo            nectarin                need
##             0.00000000          0.00000000          0.00000000
##                    new            new_york         new_zealand
##             0.00000000          0.00000000          0.00000000
##                   next                nice                noir
##             0.00000000          0.00000000          0.00000000
##                  north        northeastern  northeastern_itali
##             0.00000000          0.00000000          0.00000000
##               northern      northern_spain                nose
##             0.00000000          0.00000000          0.00000000
##             nose_palat                note                 now
##             0.00000000          0.00000000          0.00000000
##              now_franc              now_us               nuanc
##             0.00000000          0.00000000          0.00000000
##                    oak                oaki               offer
##             0.04498787          0.00000000          0.00000000
##                    old            old_vine                oliv
##             0.00000000          0.00000000          0.00000000
##                    one                open          open_aroma
##             0.00000000          0.00000000          0.00000000
##                   opul               orang              oregon
##             0.00000000          0.00000000          0.00000000
##       oregon_willamett              overal                pack
##             0.00000000          0.00000000          0.00000000
##                   pair               palat         palat_deliv
##             0.00000000          0.02711122          0.00000000
##            palat_offer          palat_show                paso
##             0.00000000          0.00000000          0.00000000
##              paso_robl               peach                pear
##             0.00000000          0.00000000          0.00000000
##                   peel              pepper             pepperi
##             0.00000000          0.00000000          0.00000000
##                perfect              perfum             persist
##             0.00000000          0.00000000          0.00000000
##                  petit         petit_sirah        petit_verdot
##             0.00000000          0.00000000          0.00000000
##                    pie            piedmont     piedmont_barolo
```

```
##          0.00000000        0.00000000        0.00000000
##            pineappl             pinot        pinot_gris
##          0.00000000        0.00000000        0.00000000
##          pinot_noir              play          pleasant
##          0.00000000        0.00000000        0.00000000
##              plenti              plum             plump
##          0.00000000        0.00000000        0.09771923
##              polish          pomegran            portug
##          0.00000000        0.00000000        0.00000000
##           portugues     portugues_red           potenti
##          0.00000000        0.00000000        0.00000000
##               power           present            pretti
##          0.00000000        0.00000000        0.00000000
##               price            produc            provid
##          0.00000000        0.00000000        0.00000000
##             provinc   provinc_mendoza             prune
##          0.00000000        0.00000000        0.00000000
##                pure             purpl           qualiti
##          0.00000000        0.00000000        0.00000000
##                quit              raci            raisin
##          0.00000000        0.00000000        0.00000000
##               ranch         raspberri  raspberri_cherri
##          0.00000000        0.00000000        0.00000000
##              rather             readi       readi_drink
##          0.00000000        0.00000000        0.00000000
##                 red         red_berri         red_blend
##          0.00000000        0.00000000        0.00000000
##          red_cherri       red_currant         red_fruit
##          0.00000000        0.00000000        0.00000000
##             refresh            region            remain
##          0.00000000        0.00000000        0.00000000
##              reserv reserv_california           reserva
##          0.00000000        0.00000000        0.00000000
##              reveal          rhnestyl      rhnestyl_red
##          0.00000000        0.00000000        0.00000000
##                rich             riesl             right
##          0.00000000        0.00000000        0.00000000
##               rioja              ripe        ripe_fruit
##          0.00000000        0.00000000        0.00000000
##             riserva             river      river_valley
##          0.00000000        0.00000000        0.00000000
##               roast              robl      robl_central
##          0.00000000        0.00000000        0.00000000
##                 ros              rose             round
##          0.00000000        0.00000000        0.00000000
##             russian     russian_river           rustic
##          0.00000000        0.00000000        0.00000000
##                sage         sangioves             santa
##          0.00000000        0.00000000        0.00000000
##        santa_barbara          sardinia         sauvignon
```

```
##          0.00000000          0.00000000          0.00000000
##     sauvignon_blanc              savori               scent
##          0.00000000          0.00000000          0.00000000
##                seem              select                sens
##          0.00000000          0.00000000          0.00000000
##                 set               sharp                show
##          0.00000000          0.00000000          0.00000000
##              sicili     sicili_sardinia                side
##          0.00000000          0.00000000          0.00000000
##              sierra      sierra_foothil               silki
##          0.00000000          0.00000000          0.00000000
##               simpl                 sip               sirah
##          0.00000000          0.00000000          0.00000000
##                skin              slight               smell
##          0.00000000          0.00000000          0.08964408
##               smoke               smoki              smooth
##          0.00000000          0.00000000          0.00000000
##                soft              soften                soil
##          0.00000000          0.00000000          0.00000000
##               solid            somewhat              sonoma
##          0.00000000          0.00000000          0.00000000
##   sonoma_chardonnay        sonoma_coast       sonoma_counti
##          0.00000000          0.00000000          0.00000000
##        sonoma_pinot                soon                sour
##          0.00000000          0.00000000          0.00000000
##               sourc               south        south_africa
##          0.00000000          0.00000000          0.00000000
##     south_australia            southern           southwest
##          0.00000000          0.00000000          0.00000000
##     southwest_franc               spain         spain_rioja
##          0.00000000          0.00000000          0.00000000
##              sparkl        sparkl_blend               spice
##          0.00000000          0.00000000          0.07907191
##        spice_flavor               spici               start
##          0.00000000          0.00000000          0.00000000
##               still               stone         stone_fruit
##          0.00000000          0.00000000          0.00000000
##     straightforward          strawberri              streak
##          0.00000000          0.00000000          0.00000000
##              strong            structur               style
##          0.00000000          0.00000000          0.00000000
##               subtl               sugar             suggest
##          0.00000000          0.00000000          0.00000000
##            superior               suppl             support
##          0.00000000          0.00000000          0.00000000
##               sweet               syrah                take
##          0.00000000          0.00000000          0.00000000
##            tangerin               tangi              tannic
##          0.00000000          0.00000000          0.00000000
##              tannin        tannin_drink                tart
```

```
##            0.00000000          0.00000000          0.00000000
##                  tast                 tea         tempranillo
##            0.06558790          0.00000000          0.00000000
##                textur                that               there
##            0.00000000          0.07944119          0.00000000
##                 thick              though               tight
##            0.00000000          0.00000000          0.00000000
##                  time               toast              toasti
##            0.00000000          0.00000000          0.00000000
##               tobacco              togeth              tomato
##            0.00000000          0.00000000          0.00000000
##                  tone             toscana               touch
##            0.00000000          0.00000000          0.00000000
##                tropic        tropic_fruit                turn
##            0.00000000          0.00000000          0.00000000
##               tuscani                 two          underbrush
##            0.00000000          0.00000000          0.00000000
##         us_california            us_estat           us_oregon
##            0.00000000          0.00000000          0.00000000
##             us_reserv       us_washington              valley
##            0.00000000          0.00000000          0.02654286
##       valley_cabernet      valley_central   valley_chardonnay
##            0.00000000          0.00000000          0.09466201
##           valley_napa        valley_pinot          valley_red
##            0.00000000          0.00000000          0.00000000
##         valley_sonoma        valley_syrah           valley_wa
##            0.00000000          0.00000000          0.00000000
##      valley_willamett             vanilla              variet
##            0.00000000          0.05590905          0.00000000
##               varieti             velveti              veneto
##            0.00000000          0.00000000          0.00000000
##                verdot                veri             vibrant
##            0.00000000          0.00000000          0.00000000
##                  vine            vineyard vineyard_california
##            0.00000000          0.00000000          0.00000000
##   vineyard_washington              vintag           viognier
##            0.00000000          0.00000000          0.00000000
##                violet         wa_columbia                warm
##            0.00000000          0.00000000          0.00000000
##            washington washington_columbia                 way
##            0.00000000          0.00000000          0.00000000
##                weight                well                 wet
##            0.00000000          0.00000000          0.00000000
##                 whiff               white         white_blend
##            0.00000000          0.00000000          0.00000000
##           white_peach        white_pepper                wild
##            0.00000000          0.00000000          0.00000000
##                  will           willamett    willamett_valley
##            0.00000000          0.00000000          0.00000000
##                  wine             winemak              wineri
```

```
##           0.00000000           0.00000000           0.00000000
##              without               wonder                 wood
##           0.00000000           0.00000000           0.07108113
##             wood_age                 wrap                 year
##           0.00000000           0.00000000           0.00000000
##              year_us               yellow                  yet
##           0.00000000           0.00000000           0.00000000
##                 york          york_finger                young
##           0.00000000           0.00000000           0.00000000
##              zealand                 zest                zesti
##           0.00000000           0.00000000           0.00000000
##            zinfandel
##           0.00000000
```

And here are the results:

```
train_svmLinear_model
```

```
## L2 Regularized Support Vector Machine (dual) with Linear Kernel
##
## 19356 samples
##   661 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:
##
##   cost  Loss  Accuracy   Kappa
##   0.25  L1    0.7804824  0.5592447
##   0.25  L2    0.7903403  0.5793207
##   0.50  L1    0.7862217  0.5711220
##   0.50  L2    0.7911250  0.5810424
##   1.00  L1    0.7893797  0.5776108
##   1.00  L2    0.7902474  0.5793807
##
## 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.
```

```
confusionMatrix(conf_svmLinear_train)
```

```
## Confusion Matrix and Statistics
##
##                  Actual class
## Predicted class excellent good
##       excellent      1603  371
##       good            660 2205
##
##                Accuracy : 0.7869
##                  95% CI : (0.7751, 0.7984)
```

```
##     No Information Rate : 0.5323
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5687
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.7084
##             Specificity : 0.8560
##          Pos Pred Value : 0.8121
##          Neg Pred Value : 0.7696
##              Prevalence : 0.4677
##          Detection Rate : 0.3313
##    Detection Prevalence : 0.4079
##       Balanced Accuracy : 0.7822
##
##        'Positive' Class : excellent
##
```

The accuracy is just slightly increased. The good news is the false negative is really better (371 compare to 448) of course the false positive shows an increase.

To explain, I see that there are some important pairs that impove the model. Considering those results and the dimension of DTM is large (which lead to high training time), while there are too many features with 0 point. I conclude that 1-2 gram is a little bit better.

The second way to improve is adjusted the weight for DTM (with bag of words). As I observed in the data. The description usually talk about ingredients: lemon, cherry or region: California… So, the first idea is double the point for nouns and see how it works. On the other hand, Adjectives used to describe the flavor of wine could act an important role. So, the second idea is double the point for adjectives only. To do these, I implement POS tagger and see which word is nouns, adj, verb…

Here is the noun list that I extracted:

```
nouns
```

```
##   [1] "accent"       "acid"         "age"          "alcohol"
##   [5] "almond"       "alreadi"      "alsac"        "appeal"
##   [9] "approach"     "apricot"      "argentina"    "aroma"
##  [13] "austria"      "bake"         "balanc"       "barbara"
##  [17] "barolo"       "barrel"       "beauti"       "berri"
##  [21] "bit"          "blanc"        "blend"        "bodi"
##  [25] "bordeaux"     "bordeauxstyl" "bottl"        "bouquet"
##  [29] "boysenberri"  "brambl"       "brut"         "butter"
##  [33] "cab"          "cabernet"     "california"   "caramel"
##  [37] "carnero"      "carri"        "cassi"        "catalonia"
##  [41] "cedar"        "cellar"       "champagn"     "char"
##  [45] "charact"      "chardonnay"   "cherri"       "chewi"
##  [49] "chile"        "chocol"       "chunki"       "cinnamon"
##  [53] "citrus"       "classico"     "clove"        "coast"
##  [57] "cocoa"        "coffe"        "cola"         "color"
```

```
##  [61] "columbia"     "combin"       "concentr"     "core"
##  [65] "creek"        "crisp"        "cru"          "crush"
##  [69] "cut"          "delic"        "depth"        "despit"
##  [73] "domin"        "dri"          "drink"        "earth"
##  [77] "earthi"       "eleg"         "element"      "end"
##  [81] "espresso"     "estat"        "express"      "ferment"
##  [85] "fill"         "finger"       "firm"         "flavor"
##  [89] "flower"       "focus"        "food"         "foothil"
##  [93] "forest"       "frame"        "franc"        "french"
##  [97] "front"        "fruit"        "fruiti"       "gentl"
## [101] "glass"        "grape"        "grapefruit"   "grenach"
## [105] "grill"        "grip"         "herb"         "high"
## [109] "highlight"    "hill"         "hint"         "honey"
## [113] "integr"       "interest"     "invit"        "jam"
## [117] "jammi"        "lack"         "lake"         "layer"
## [121] "lead"         "leaf"         "leather"      "leav"
## [125] "lemon"        "length"       "lift"         "light"
## [129] "lime"         "linger"       "load"         "mango"
## [133] "mark"         "matur"        "meat"         "medium"
## [137] "melon"        "mendoza"      "merlot"       "midpal"
## [141] "miner"        "mint"         "mix"          "mocha"
## [145] "moder"        "month"        "mountain"     "mourvdr"
## [149] "mouth"        "mouthfeel"    "napa"         "napasonoma"
## [153] "natur"        "nebbiolo"     "nectarin"     "noir"
## [157] "north"        "nose"         "note"         "oak"
## [161] "oaki"         "one"          "opul"         "oregon"
## [165] "pack"         "pair"         "palat"        "paso"
## [169] "peach"        "pear"         "peel"         "pepper"
## [173] "pepperi"      "perfum"       "pie"          "piedmont"
## [177] "pineappl"     "play"         "plum"         "plump"
## [181] "polish"       "pomegran"     "portug"       "power"
## [185] "present"      "price"        "produc"       "provinc"
## [189] "prune"        "purpl"        "qualiti"      "ranch"
## [193] "raspberri"    "region"       "reserva"      "rhnestyl"
## [197] "rioja"        "riserva"      "river"        "roast"
## [201] "sage"         "sardinia"     "sauvignon"    "scent"
## [205] "show"         "side"         "sierra"       "silki"
## [209] "sip"          "skin"         "smell"        "smoke"
## [213] "soil"         "sonoma"       "sourc"        "southwest"
## [217] "spain"        "sparkl"       "spice"        "stone"
## [221] "streak"       "structur"     "style"        "sugar"
## [225] "support"      "syrah"        "tangerin"     "tannin"
## [229] "tart"         "tast"         "tea"          "textur"
## [233] "time"         "toast"        "toasti"       "tobacco"
## [237] "tone"         "touch"        "tropic"       "turn"
## [241] "valley"       "vanilla"      "varieti"      "velveti"
## [245] "veneto"       "verdot"       "veri"         "vine"
## [249] "vineyard"     "vintag"       "viognier"     "violet"
## [253] "way"          "weight"       "will"         "wine"
## [257] "winemak"      "wineri"       "wonder"       "wood"
```

```
## [261] "wrap"          "year"          "yellow"        "zealand"
## [265] "zest"          "zesti"         "zinfandel"
```

Let's take a look at the Dtm of the first document to make sure that the points for nouns is doubled. As we can see below, the points for the word "california" is 0.139. It's just 0.064 before.

```
wine_train_set[1,]

##         accent          acid        across           add
##     0.00000000    0.00000000    0.00000000    0.00000000
##         africa     aftertast           age       alcohol
##     0.00000000    0.00000000    0.00000000    0.00000000
##         almond        almost         along      alongsid
##     0.00000000    0.15504157    0.00000000    0.00000000
##        alreadi         alsac          also      although
##     0.00000000    0.00000000    0.00000000    0.00000000
##           ampl          anis         anoth        appeal
##     0.00000000    0.00000000    0.00000000    0.00000000
##           appl      approach       apricot     argentina
##     0.00000000    0.00000000    0.00000000    0.00000000
##          aroma        aromat        around       astring
##     0.00000000    0.00000000    0.00000000    0.00000000
##        attract     australia       austria          back
##     0.00000000    0.00000000    0.00000000    0.00000000
##           bake        balanc       barbara        barolo
##     0.00000000    0.00000000    0.00000000    0.00000000
##         barrel        beauti         berri          best
##     0.30610113    0.00000000    0.00000000    0.00000000
##         better           big           bit        bitter
##     0.00000000    0.00000000    0.00000000    0.00000000
##          black    blackberri         blanc         blend
##     0.00000000    0.00000000    0.00000000    0.00000000
##        blossom          blue     blueberri          bodi
##     0.00000000    0.00000000    0.00000000    0.00000000
##           bold      bordeaux  bordeauxstyl         bottl
##     0.00000000    0.00000000    0.00000000    0.00000000
##        bouquet   boysenberri        brambl        bright
##     0.00000000    0.00000000    0.00000000    0.00000000
##          bring         brisk          brut      burgundi
##     0.00000000    0.00000000    0.00000000    0.00000000
##         butter           cab      cabernet    california
##     0.00000000    0.00000000    0.00000000    0.00000000
##            can         candi       caramel       carnero
##     0.00000000    0.00000000    0.00000000    0.00000000
##          carri         cassi     catalonia         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    0.39886867    0.00000000    0.00000000
```

```
##        chile       chocol       chunki     cinnamon
##   0.29422005   0.00000000   0.00000000   0.00000000
##       citrus      citrusi      classic     classico
##   0.00000000   0.00000000   0.00000000   0.00000000
##        clean        close        clove        coast
##   0.00000000   0.00000000   0.30417613   0.00000000
##        cocoa        coffe         cola        color
##   0.00000000   0.00000000   0.00000000   0.00000000
##     columbia       combin         come      complex
##   0.00000000   0.00000000   0.00000000   0.00000000
##     concentr         cool         core       counti
##   0.00000000   0.00000000   0.00000000   0.00000000
##    cranberri       creami        creek        crisp
##   0.00000000   0.00000000   0.00000000   0.00000000
##          cru        crush         ctes      currant
##   0.00000000   0.00000000   0.00000000   0.00000000
##          cut         cuve         dark         deep
##   0.00000000   0.00000000   0.00000000   0.00000000
##          del        delic       delici        deliv
##   0.00000000   0.00000000   0.00000000   0.00000000
##         dens        depth       despit      develop
##   0.00000000   0.00000000   0.00000000   0.00000000
##       doesnt        domin          dri        drink
##   0.00000000   0.00000000   0.00000000   0.00000000
##        dusti        earth       earthi         easi
##   0.00000000   0.00000000   0.00000000   0.00000000
##          edg         eleg      element          end
##   0.00000000   0.00000000   0.00000000   0.00000000
##        enjoy       enough     espresso        estat
##   0.00000000   0.00000000   0.00000000   0.00000000
##         even        excel         exot      express
##   0.00000000   0.00000000   0.00000000   0.00000000
##        extra      extract         fair       famili
##   0.00000000   0.00000000   0.00000000   0.00000000
##       featur         feel      ferment         fill
##   0.00000000   0.00000000   0.00000000   0.00000000
##         find         fine       finger       finish
##   0.00000000   0.00000000   0.00000000   0.04989804
##         firm        first       flavor       fleshi
##   0.00000000   0.00000000   0.05912793   0.00000000
##       floral       flower        focus       follow
##   0.00000000   0.00000000   0.00000000   0.00000000
##         food      foothil       forest      forward
##   0.00000000   0.00000000   0.00000000   0.00000000
##     fragrant        frame        franc       french
##   0.00000000   0.00000000   0.00000000   0.00000000
##        fresh        front        fruit       fruiti
##   0.00000000   0.00000000   0.08213201   0.00000000
##         full     fullbodi     generous        gentl
##   0.00000000   0.00000000   0.00000000   0.00000000
```

```
##        germani         get         give         glass
##     0.00000000  0.18456806   0.00000000    0.00000000
##           good       grand        grape    grapefruit
##     0.00000000  0.00000000   0.00000000    0.00000000
##        graphit       great        green       grenach
##     0.00000000  0.00000000   0.00000000    0.00000000
##          grill        grip         gris         grown
##     0.00000000  0.00000000   0.00000000    0.00000000
##           hard       heavi         herb        herbal
##     0.00000000  0.00000000   0.00000000    0.00000000
##           high   highlight         hill          hint
##     0.00000000  0.00000000   0.00000000    0.00000000
##           hold       honey   honeysuckl           hot
##     0.00000000  0.00000000   0.00000000    0.00000000
##        impress      includ       integr        intens
##     0.00000000  0.00000000   0.00000000    0.00000000
##       interest     intrigu        invit         itali
##     0.00000000  0.00000000   0.00000000    0.00000000
##            jam       jammi         juic         juici
##     0.00000000  0.00000000   0.00000000    0.00000000
##           just        keep         lack          lake
##     0.00000000  0.00000000   0.00000000    0.00000000
##           last       layer         lead          leaf
##     0.00000000  0.00000000   0.00000000    0.00000000
##           lean       least      leather          leav
##     0.00000000  0.00000000   0.00000000    0.00000000
##          lemon        lend       length           les
##     0.00000000  0.00000000   0.00000000    0.00000000
##        licoric        lift        light          like
##     0.00000000  0.00000000   0.00000000    0.00000000
##           lime      linger        littl          live
##     0.00000000  0.00000000   0.00000000    0.00000000
##           load        loir         long           lot
##     0.00000000  0.00000000   0.00000000    0.00000000
##           love        lush         made          make
##     0.00000000  0.00000000   0.00000000    0.00000000
##         malbec       mango         mani          mark
##     0.00000000  0.00000000   0.00000000    0.00000000
##          matur        meat       medium    mediumbodi
##     0.00000000  0.00000000   0.00000000    0.00000000
##          melon     mendoza       merlot        midpal
##     0.00000000  0.00000000   0.00000000    0.00000000
##           mild       miner         mint           mix
##     0.00000000  0.00000000   0.00000000    0.00000000
##          mocha       moder   montalcino         month
##     0.00000000  0.00000000   0.00000000    0.00000000
##       mountain     mourvdr        mouth     mouthfeel
##     0.00000000  0.00000000   0.00000000    0.00000000
##           much        napa   napasonoma         natur
##     0.00000000  0.00000000   0.00000000    0.00000000
```

```
##      nebbiolo     nectarin         need          new
##    0.00000000   0.00000000   0.00000000   0.00000000
##          next         nice         noir        north
##    0.00000000   0.00000000   0.00000000   0.00000000
##  northeastern     northern         nose         note
##    0.00000000   0.00000000   0.00000000   0.00000000
##           now        nuanc          oak         oaki
##    0.00000000   0.00000000   0.17722495   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   0.10680177
##          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        pinot         play     pleasant
##    0.00000000   0.00000000   0.00000000   0.00000000
##        plenti         plum        plump       polish
##    0.00000000   0.00000000   0.38495454   0.00000000
##      pomegran       portug    portugues      potenti
##    0.00000000   0.00000000   0.00000000   0.00000000
##         power      present       pretti        price
##    0.00000000   0.00000000   0.00000000   0.00000000
##        produc       provid      provinc        prune
##    0.00000000   0.00000000   0.00000000   0.00000000
##          pure        purpl      qualiti         quit
##    0.00000000   0.00000000   0.00000000   0.00000000
##          raci       raisin        ranch    raspberri
##    0.00000000   0.00000000   0.00000000   0.00000000
##        rather        readi          red      refresh
##    0.00000000   0.00000000   0.00000000   0.00000000
##        region       remain       reserv      reserva
##    0.00000000   0.00000000   0.00000000   0.00000000
##        reveal     rhnestyl         rich        riesl
##    0.00000000   0.00000000   0.00000000   0.00000000
##         right        rioja         ripe      riserva
##    0.00000000   0.00000000   0.00000000   0.00000000
##         river        roast         robl          ros
##    0.00000000   0.00000000   0.00000000   0.00000000
##          rose        round      russian       rustic
##    0.00000000   0.00000000   0.00000000   0.00000000
##          sage    sangioves        santa     sardinia
##    0.00000000   0.00000000   0.00000000   0.00000000
##     sauvignon       savori        scent         seem
##    0.00000000   0.00000000   0.00000000   0.00000000
```

```
##          select           sens            set          sharp 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##            show         sicili           side         sierra 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##           silki          simpl            sip          sirah 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##            skin        slight          smell          smoke 
##      0.00000000     0.00000000     0.35314335     0.00000000 
##           smoki         smooth           soft         soften 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##            soil          solid       somewhat         sonoma 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##            soon           sour          sourc          south 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##        southern      southwest          spain         sparkl 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##           spice          spici          start          still 
##      0.31149539     0.00000000     0.00000000     0.00000000 
##           stone straightforward     strawberri         streak 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##          strong       structur          style          subtl 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##           sugar        suggest       superior          suppl 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##         support          sweet          syrah           take 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##        tangerin          tangi         tannic         tannin 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##            tart           tast            tea    tempranillo 
##      0.00000000     0.25837656     0.00000000     0.00000000 
##          textur           that          there          thick 
##      0.00000000     0.15647508     0.00000000     0.00000000 
##          though          tight           time          toast 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##          toasti        tobacco         togeth         tomato 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##            tone        toscana          touch         tropic 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##            turn        tuscani            two     underbrush 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##          valley        vanilla         variet        varieti 
##      0.10456280     0.22024776     0.00000000     0.00000000 
##         velveti         veneto         verdot           veri 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##         vibrant           vine       vineyard         vintag 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##       viognier         violet           warm     washington 
##      0.00000000     0.00000000     0.00000000     0.00000000 
##             way         weight           well            wet 
##      0.00000000     0.00000000     0.00000000     0.00000000 
```

```
##           whiff          white           wild           will
##      0.00000000     0.00000000     0.00000000     0.00000000
##       willamett           wine        winemak         wineri
##      0.00000000     0.00000000     0.00000000     0.00000000
##         without         wonder           wood           wrap
##      0.00000000     0.00000000     0.28001657     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
```

Here are the results:

```
train_svmLinear_model
```

```
## L2 Regularized Support Vector Machine (dual) with Linear 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:
## 
##   cost  Loss  Accuracy   Kappa    
##   0.25  L1    0.7835696  0.5658319
##   0.25  L2    0.7845502  0.5679076
##   0.50  L1    0.7843196  0.5674399
##   0.50  L2    0.7840762  0.5670011
##   1.00  L1    0.7842031  0.5672562
##   1.00  L2    0.7835764  0.5660250
## 
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were cost = 0.25 and Loss = L2.
```

```
confusionMatrix(conf_svmLinear_train)
```

```
## Confusion Matrix and Statistics
## 
##                Actual class
## Predicted class excellent good
##       excellent      1651  433
##       good            612 2143
##                                           
##                Accuracy : 0.784           
##                  95% CI : (0.7722, 0.7956)
##     No Information Rate : 0.5323
```

```
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5642
##  Mcnemar's Test P-Value : 3.664e-08
##
##             Sensitivity : 0.7296
##             Specificity : 0.8319
##          Pos Pred Value : 0.7922
##          Neg Pred Value : 0.7779
##              Prevalence : 0.4677
##          Detection Rate : 0.3412
##    Detection Prevalence : 0.4307
##       Balanced Accuracy : 0.7807
##
##        'Positive' Class : excellent
##
```

We can see that The accuracy is decreasing 0.2%. FN is decreasing a few units. I conclude that the model is not better.

Now double the point only for adjectives: Here is the list adjectives that I extracted:

```
adj
```

```
##   [1] "aftertast"       "ampl"            "appl"
##   [4] "australia"       "big"             "bitter"
##   [7] "black"           "blackberri"      "blue"
##  [10] "blueberri"       "bold"            "bright"
##  [13] "brisk"           "central"         "classic"
##  [16] "clean"           "close"           "complex"
##  [19] "cool"            "creami"          "currant"
##  [22] "cuve"            "dark"            "deep"
##  [25] "doesnt"          "edg"             "enough"
##  [28] "exot"            "extra"           "fair"
##  [31] "fine"            "floral"          "fragrant"
##  [34] "fresh"           "full"            "generous"
##  [37] "good"            "grand"           "great"
##  [40] "green"           "hard"            "heavi"
##  [43] "hot"             "juic"            "last"
##  [46] "lean"            "licoric"         "littl"
##  [49] "live"            "lush"            "malbec"
##  [52] "mild"            "new"             "next"
##  [55] "nice"            "northeastern"    "northern"
##  [58] "nuanc"           "old"             "oliv"
##  [61] "open"            "overal"          "perfect"
##  [64] "petit"           "pleasant"        "pure"
##  [67] "readi"           "red"             "refresh"
##  [70] "rich"            "right"           "ripe"
##  [73] "robl"            "russian"         "rustic"
##  [76] "select"          "sharp"           "sicili"
##  [79] "simpl"           "slight"          "smooth"
```

```
##  [82] "soft"            "solid"           "sour"
##  [85] "south"           "southern"        "straightforward"
##  [88] "strawberri"      "strong"          "superior"
##  [91] "suppl"           "sweet"           "tannic"
##  [94] "tempranillo"     "thick"           "tight"
##  [97] "variet"          "vibrant"         "warm"
## [100] "wet"             "white"           "wild"
## [103] "young"
```

And here are the results after double the points for adjectives:

```
train_svmLinear_model

## L2 Regularized Support Vector Machine (dual) with Linear 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:
##
##   cost  Loss  Accuracy   Kappa
##   0.25  L1    0.7812716  0.5612232
##   0.25  L2    0.7844977  0.5676142
##   0.50  L1    0.7827696  0.5642357
##   0.50  L2    0.7840476  0.5667318
##   1.00  L1    0.7826981  0.5640836
##   1.00  L2    0.7839151  0.5664589
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were cost = 0.25 and Loss = L2.

confusionMatrix(conf_svmLinear_train)

## Confusion Matrix and Statistics
##
##                  Actual class
## Predicted class excellent good
##       excellent      1667  443
##       good            596 2133
##
##                Accuracy : 0.7853
##                  95% CI : (0.7734, 0.7968)
##     No Information Rate : 0.5323
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.567
##  Mcnemar's Test P-Value : 2.41e-06
```

```
##
##             Sensitivity : 0.7366
##             Specificity : 0.8280
##          Pos Pred Value : 0.7900
##          Neg Pred Value : 0.7816
##              Prevalence : 0.4677
##          Detection Rate : 0.3445
##    Detection Prevalence : 0.4360
##       Balanced Accuracy : 0.7823
##
##        'Positive' Class : excellent
##
```

It's clearly that this model is more or less the same with the original one. To explain why double the points is not work. When reviewing the dataset. I see there are three reasons: Firstly, most important nouns is the ingredients, country, region… They can use the same ingredients, plant in the same region. But, the experts may have some unique techniques to make it become excellent wine. So, nouns are not so helpful. I also see that only famous ingredients are mentioned. Maybe the secret one is not published. Secondly, because I steam when processing data. Many original words are steamed to nouns but it is not the nouns in the original paragraph. Which leads to the method is not as useful as expected. Finally, one factor can affect the system is the POS tagger function doesn't work perfectly, especially in the adjective list. For example, the word "blackberri" should be in the noun list but it appears in the adjective list, and vice versa for the words "yellow".

Finally I conclude that the best model for "good" - "excellent" wine classify is SVM linear with 1-2gram.

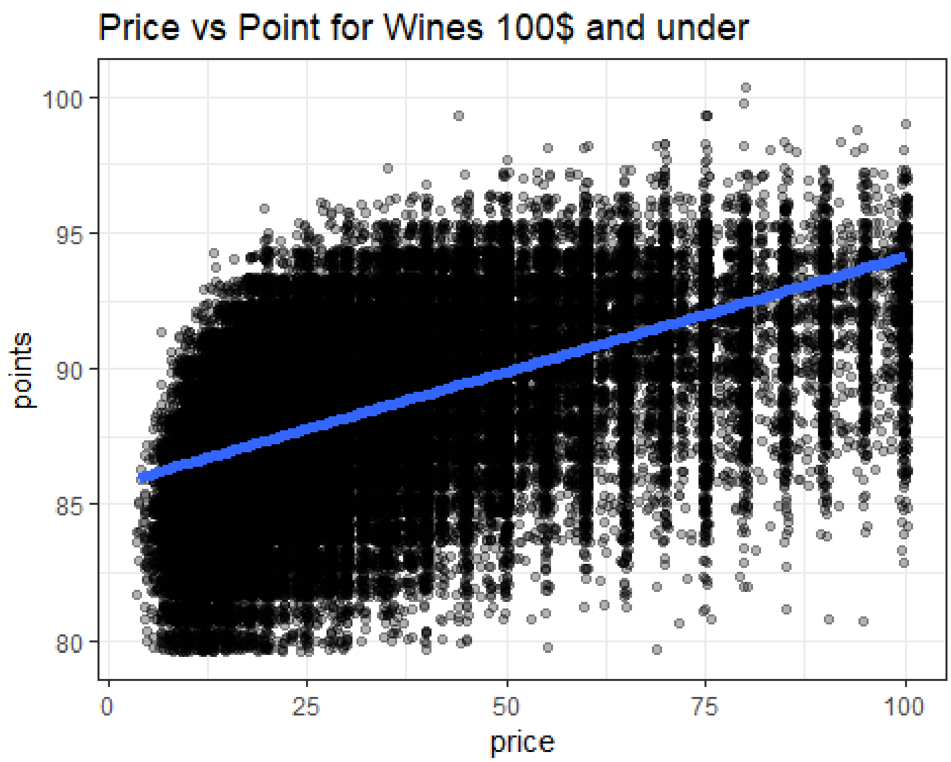
6. Second aspect: High value wine

6.1 Motivation

In most of the times, the price of the bottle can reflect the quality of the wine. For example, when you want to buy 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: Spend money in the right way.

Just want to confirm that the problem is real and solvable. Let see the scatter plot of price vs points for wines 100$ and under:

```
ggplot(subset(wine, price <= 100),
       aes(x = price, y = points)) +
  geom_point(alpha = 0.3,  position = position_jitter()) +
  stat_smooth(method = "lm", size =2) +
  labs(title = 'Price vs Point for Wines 100$ and under') +
  theme_bw()
```



As we can see, there is a positive relationship between the points and the price. But, the points for bottles with the same price can be very different. Very cheap wine(<10$) can have up to 92 points while many 25 dollar wines just have less than 85 points.

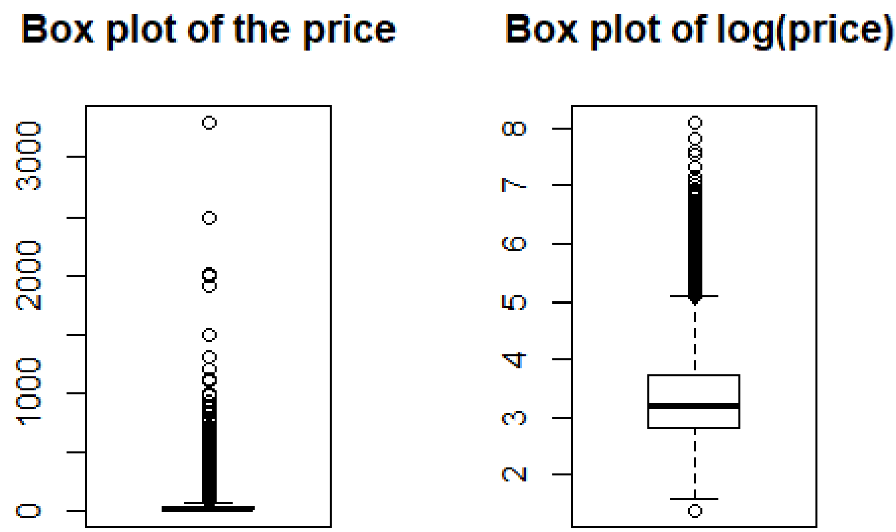
The thing that I see from this graph is: We can have a strategy for choosing excellent wine with a low budget, while there is a high probability of having just a good bottle when choosing randomly even with a quite high price (~50$).

The aim is similar to the first problem: Using NLP and compare ML kernels to solve this problems.

6.2 Method and result

Let's take a look at the price of our data:

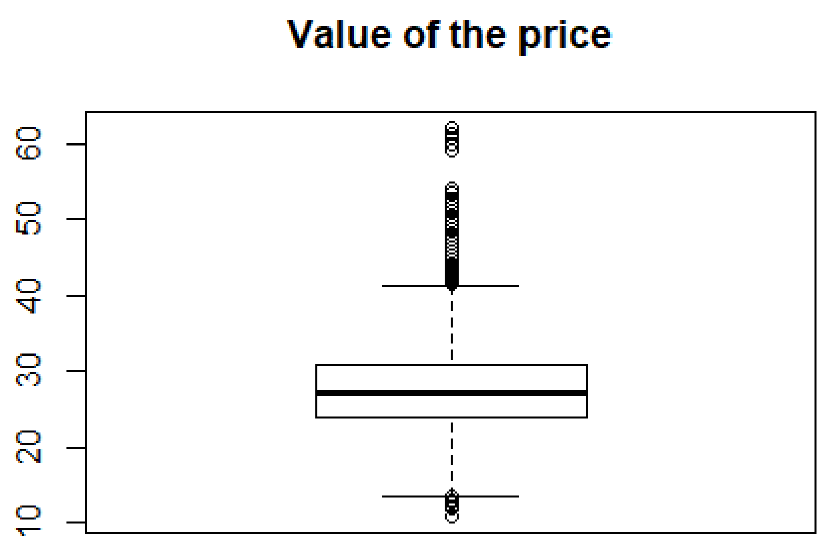
```
par(mfrow=c(1,2))
boxplot(wine$price, main = "Box plot of the price")
boxplot(log(wine$price), main = " Box plot of log(price)")
```



The price has a wide range. It also has a lot of outliers and high value of SD. otherwise, the log of the price looks better and quite similar to the point.

I define the $Value = \frac{points}{log(price)}$ Here is the value:

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



The median number for value is ~27. So the bottle with a higher value than 27 will have a “high” benefit. The rests 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 points. It becomes the bottle with the highest benefit. In contrast. many bottles with the price higher than 1000 dollar. Of course, they are an excellent wine. But is just have a medium benefit because it’s too expensive.

I do the same preprocessing for the text. Then continue to compare models with the same 20% sample of the data. Here are the results.

```
Naive Bayes

19356 samples
  505 predictor
    2 classes: 'high', 'medium'

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

  usekernel  Accuracy   Kappa
  FALSE      0.7310714  0.4613442
   TRUE      0.6803452  0.3705769

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.
```

NB benefit

```
Confusion Matrix and Statistics

               Actual class
Predicted class high medium
         high   1912    757
         medium  588   1582

               Accuracy : 0.7221
                 95% CI : (0.7092, 0.7346)
    No Information Rate : 0.5166
    P-Value [Acc > NIR] : < 2.2e-16
```

NB benefit predict

```
L2 Regularized Support Vector Machine (dual) with Linear Kernel

19356 samples
  505 predictor
    2 classes: 'high', 'medium'

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

  cost  Loss  Accuracy   Kappa
  0.25  L1    0.7656750  0.5298202
  0.25  L2    0.7665210  0.5316692
  0.50  L1    0.7669268  0.5324227
  0.50  L2    0.7663409  0.5313085
  1.00  L1    0.7672066  0.5330413
  1.00  L2    0.7659929  0.5306153

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were cost = 1 and Loss = L1.
```

SVM benefit

```
Confusion Matrix and Statistics

               Actual class
Predicted class high medium
         high   2009    629
         medium  491   1710

               Accuracy : 0.7685
                 95% CI : (0.7564, 0.7804)
    No Information Rate : 0.5166
    P-Value [Acc > NIR] : < 2.2e-16
```

SVMLinear_benefit

The indicators for comparing models are the same with the previous problem. As we can see, SVM Linear is better than NB in both accuracy and how the bottle is classified. Now I'll try with POS Tagger

Here are the results when double the point for nouns:

```
Confusion Matrix and Statistics

               Actual class
Predicted class high medium
         high   2014    655
         medium  486   1684

               Accuracy : 0.7642
                 95% CI : (0.752, 0.7761)
    No Information Rate : 0.5166
    P-Value [Acc > NIR] : < 2.2e-16
```

And here are the results when double the point for adjectives:

```
Confusion Matrix and Statistics

               Actual class
Predicted class high medium
         high   2055    686
         medium  445   1653

               Accuracy : 0.7663
                 95% CI : (0.7541, 0.7781)
    No Information Rate : 0.5166
    P-Value [Acc > NIR] : < 2.2e-16
```

In this case, POS Tagger doesn't show any clear effect either with double the point for nouns or adjective. All indicator is more or less the same, or even worse. As I said before, the way of classifying here has mix everything.

And here is with 1-2gram model:

```
L2 Regularized Support Vector Machine (dual) with Linear Kernel

19356 samples
  661 predictor
    2 classes: 'high', 'medium'

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

  cost  Loss  Accuracy   Kappa
  0.25  L1    0.7586720  0.5151721
  0.25  L2    0.7707692  0.5399092
  0.50  L1    0.7665848  0.5312198
  0.50  L2    0.7717997  0.5420806
  1.00  L1    0.7701554  0.5386099
  1.00  L2    0.7715528  0.5416235

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.

Confusion Matrix and Statistics

               Actual class
Predicted class high medium
         high   2058    675
         medium  442   1664

               Accuracy : 0.7692
                 95% CI : (0.757, 0.781)
    No Information Rate : 0.5166
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5363
 Mcnemar's Test P-Value : 3.876e-12
```

The results is more or less the same. In this case, 1-2gram doesn't improve the model.

Finally, I conclude that the final model for "high" - "medium" wine value classify is SVM linear with tf-idf, bag of words.

7. Further discussion

7.1 Suggestions for choosing wine

Because both problems is solve uing SVMLinear model, which use vector based on tf-idf points. As a result, top points of each class should be the most influences of the class. In this section, I'll extract the top influences of each class and draw by word cloud.

Here are the results from the SVM Linear for "good" - "excellent" wine model: The first cloud is for excellent wine and the second one is for good wine. Noticed that the words is steamed.

```
wordcloud(rownames(excellent), excellent[,1], max.words=100, colors=brewer.pa
l(8, "Dark2"), scale=c(4,.5))
```


```
wordcloud(rownames(good), good[,1], max.words=100, colors=brewer.pal(8,
"Dark2"), scale=c(4,.5))
```

Suggestions for choosing wine: For excellent wine, we should pay attention to the word "vineyard", which is a plantation of grape-bearing vines, grown mainly for winemaking[1]. "concentr" (concentrate) "red_blen", "blanc", "itali" (Italy), "fresh" are some words those you should look when searching for an excellent bottle. "red" and "oak" will appear in both classes. "sauvignon", "finish", "fruit" and "franc" (France) are typical words for "good" class.

Here are the top influencer words from the SVM Linear for "high" - "medium" wine model. The first cloud is for high benefit wine and the second one is for medium benefit wine.

```
wordcloud(rownames(high), high[,1], max.words=100, colors=brewer.pal(8, "Dark
2"), scale=c(4,.5))
```


```
wordcloud(rownames(medium), medium[,1], max.words=100, colors=brewer.pal(8, "
Dark2"), scale=c(4,.5))
```

Suggesstions for choosing wine: Suggestions for choosing wine: “franc” (France) and “itali” (Italy) likely have more high-value bottle than medium. “spain” (Spain), “carbenet”, “appl” (apply), “pear”, “cherri” are words for high benefit. “valley”, “black”, “champagn” are considered when looking for a medium benefit wine.

7.2 Future work:

Here are some points that can improve from the project:

- Check again the SVM RBF kernel with a bigger data set and more powerful computer. In my experiments, SVM RBF is better than SVM Linear in most the cases.
- Improve the POS tagger function.
- Check again the result for 1-2gram with a bigger data set. I try with full data but mycomputer is crashed when train the model because DTM for full data is very heavy (we have a lot of terms)
- Take the advice from wine specialist to improve the model. In my opinion, we should fully understand the dataset before thinking about machine learning. In this case. Because my knowledge of wine is limited. I often don’t clearly understand about a bottle after reading the description. It leads to the ideal for improving the model are limited.
- Split red wine and white wine. These types of wine are quite different in both ingredient and technique. Red wine takes 2/3 of data set while ¼ is white wine (1/12 are others). But you must have some knowledge about wine when split the dataset because the type of wine is not provided, we only have variety like “white blend”, “pinot noir”, “portuguese red”, “riesling”..
- Improve the stop word list. Knowledge about wine is required.

8. Conclusion

The issues raised at the beginning are solved with positive results. We figure out that the description and other information printed in a bottle can bring us some idea about how good the bottle is or how high level of benefit it has. We also have some idea about how can we choose a satisfied bottle when reading the descriptions. In the NLP and ML aspect. SVM linear with 1-2gram is the best model for predicting the quality with 78.69% accuracy. For predicting the benefit of a bottle, SVM linear is an acceptable model, with accuracy 77.09%.

References

[1] https://en.wikipedia.org/wiki/Vineyard

[2] https://fderyckel.github.io/2016-12-07-Texts_Classification_in_R/

[3] https://www.kaggle.com/carkar/classifying-wine-type-by-review

[4] https://stackoverflow.com/questions/28764056/could-not-find-function-tagpos

[5] https://machinelearningcoban.com/2017/04/09/smv/

[6] https://machinelearningcoban.com/2017/04/22/kernelsmv/

Appendix

Here is the code used in this project. POS tagger function is modify from [4]. For some models that require a long time for training, I run it independently and attach the pictures of results.

```
library(dplyr)
library(tm)
library(stringr)
library(NLP)
library(openNLP)
library(caret)
library(wordcloud)
library(RColorBrewer)
knitr::opts_chunk$set(echo = TRUE, warning=FALSE, message=FALSE, include=FALS
E)
wine <- read.csv("C:/Users/Duong Minh Duc/Documents/GitHub/Text-Mining-Projec
t/wine.csv")
head(wine)
boxplot(wine$points, main="Boxplot of points")
stopwords <- stopwords("english")
stopwords <- stopwords[!stopwords=="very"]
stopwords <- c("the", "and", "wine", stopwords)

stopwords
# Use for check the not available and duplicate data, but not necessary
wine <- na.omit(wine)
#wine[duplicated(wine),]

#Assign class
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$designation, w
ine$province, wine$region_1, wine$region_2, wine$variety)
wine$description <- as.character(wine$description)

#Language convert
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 non ASCII
wine$description <- iconv(wine$description, from = "UTF-8", to = "ASCII", sub
= "")

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

id2 = sample(1:n, floor(n*0.2))
wine_sample <- 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$description[1]
wine_train_set[[1]]$content

train_dtm_tfidf <- DocumentTermMatrix(wine_train_set, control = list(weightin
g = weightTfIdf))
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) ,weighting = weightTfIdf))

#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,]
wine_test_set[1,]
#train model
train_nb_model <- train(x= wine_train_set, y=train$quality , method = 'naive_
bayes')

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")

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

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")

train_svmLinear_model
confusionMatrix(conf_svmLinear_train)
# #Here is the SVM RBF kenel
# # Because the trainning time is long. I'll not run this code and only attac
h images of the previous run.
# 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)
#
# conf_svmRBF_train <- table(model_svmRBF_result, wine_testing_result)
# names(dimnames(conf_svmRBF_train)) <- c("Predicted class", "Actual class")
# confusionMatrix(conf_svmRBF_train)
NLP_tokenizer <- function(x) {
  unlist(lapply(ngrams(words(x), 2:2), paste, collapse = "_"), use.names = FA
LSE)
```

```
}

wine_train_set <- clean(train$description)

train_dtm_tfidf <- DocumentTermMatrix(wine_train_set, control = list(weightin
g = weightTfIdf, tokenize=NLP_tokenizer))
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) ,weighting = weightTfIdf, tokenize=NLP_tokenizer))


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

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")
train_svmLinear_model
confusionMatrix(conf_svmLinear_train)
NLP_tokenizer <- function(x) {
  unlist(lapply(ngrams(words(x), 1:2), paste, collapse = "_"), use.names = FA
LSE)
}

wine_train_set <- clean(train$description)

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

```r
#create matrix for training
wine_train_set_ngram <<- as.matrix(train_dtm_tfidf)
wine_test_set <- as.matrix(wine_test_set)
wine_test_set_ngram <- wine_test_set[,Terms(train_dtm_tfidf)]
#create the test result
wine_testing_result_ngram <- test$quality

wine_train_set_ngram[1,]
train_svmLinear_model <- train(x= wine_train_set_ngram, y=train$quality , met
hod = 'svmLinear3')

model_svmLinear_result <- predict(train_svmLinear_model, newdata = wine_test_
set_ngram)

conf_svmLinear_train <- table(model_svmLinear_result, wine_testing_result_ngr
am)
names(dimnames(conf_svmLinear_train)) <- c("Predicted class", "Actual class")

train_svmLinear_model
confusionMatrix(conf_svmLinear_train)
#tagPos
tagPOS <-  function(x, ...) {
  s <- as.String(x)
  word_token_annotator <- Maxent_Word_Token_Annotator()
  a2 <- Annotation(1L, "sentence", 1L, nchar(s))
  a2 <- NLP::annotate(s, word_token_annotator, a2)
  a3 <- NLP::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)
}
wine_train_set <- clean(train$description)

train_dtm_tfidf <- DocumentTermMatrix(wine_train_set, control = list(weightin
g = weightTfIdf))
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) ,weighting = weightTfIdf))


#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


#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]

nouns

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

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")

train_svmLinear_model
confusionMatrix(conf_svmLinear_train)
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
column_id <- c()
for (i in 1:dim(wine_train_set)[2]) {
  check <- colnames(wine_train_set)[i] %in% adj
  if(check)
    {
```

```r
      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

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

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")

train_svmLinear_model
confusionMatrix(conf_svmLinear_train)
ggplot(subset(wine, price <= 100),
       aes(x = price, y = points)) +
  geom_point(alpha = 0.3,  position = position_jitter()) +
  stat_smooth(method = "lm", size =2) +
  labs(title = 'Price vs Point for Wines 100$ and under') +
  theme_bw()
par(mfrow=c(1,2))
boxplot(wine$price, main = "Box plot of the price")
boxplot(log(wine$price), main = " Box plot of log(price)")
boxplot(wine$value, main = "Value of the price")
good<- which(train$quality=="good")
good <- wine_train_set_ngram[good,]
good = data.frame(sort(colSums(good), decreasing=TRUE))
excellent<- which(train$quality=="excellent")
excellent <- wine_train_set_ngram[excellent,]
excellent = data.frame(sort(colSums(excellent), decreasing=TRUE))

wordcloud(rownames(excellent), excellent[,1], max.words=100, colors=brewer.pa
l(8, "Dark2"), scale=c(4,.5))

wordcloud(rownames(good), good[,1], max.words=100, colors=brewer.pal(8, "Dark
2"), scale=c(4,.5))
#remove double points for adj
wine_train_set[,column_id] <- wine_train_set[,column_id]/2
wine_test_set[,column_id] <- wine_test_set[,column_id]/2

high<- which(train$benefit=="high")
high <- wine_train_set[high,]
medium<- which(train$benefit=="medium")
medium <- wine_train_set[medium,]

high = data.frame(sort(colSums(high), decreasing=TRUE))
```

```
medium = data.frame(sort(colSums(medium), decreasing=TRUE))

wordcloud(rownames(high), high[,1], max.words=100, colors=brewer.pal(8, "Dark
2"), scale=c(4,.5))

wordcloud(rownames(medium), medium[,1], max.words=100, colors=brewer.pal(8, "
Dark2"), scale=c(4,.5))
```