Text mining project

mindu931

16 January 2019

# Abstract

Wine sensory audition and assessment is never easy, even with a wine specialist. However, the description which is printed on each bottle can bring us some helpful information. Base on that ideal, I decided to analyze the data about wine review (take form Kaggle) and help the consumers make the decision on two different aspects. The first one is which is a good and excellent wine. The second aspect is which is the valueble bottle (high value with a acceptable amount of money). I use different tecniques to process the text data, then compare Naive Bayes, SVM linear and SVM RBF kernel when building the best model for predicting. The final results are quite impressive with accuracy 78.24% for the first issue and 77.09% for the second one. Base on these model, some suggestions are provided for the consumer to choose a satisfied bottle.

# Introduction

## Motivation

I always confuse when standing in a wine cellar. How can I choose an excellent wine for my family? What is the best wine I can choose with my budget? The problem can be solved easier if I had a good sense. But, sadly, I don’t, like many people in this world. 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 these advices for my event. Unluckily, sometimes my relatives don’t think that it is an excellent wine. ## Aim I guess that the information from the provider (description, country, region, designation…) should bring some important information. So, I try to solve the problems base on text mining ideal: Take the wine data which is already reviewed and marked by some specialists, then process these 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. ## Content of this report In this report, Firstly, I will provide information about the data as well as the way I define the class for each bottle. Then, I briefly summary relevant theory. The method I used will be presented in details. It includes the steps for preprocessing, techiques for comparing and choosing the best model. Finally is some suggestions for choosing wine, some discussions and my conclusion.

# Data

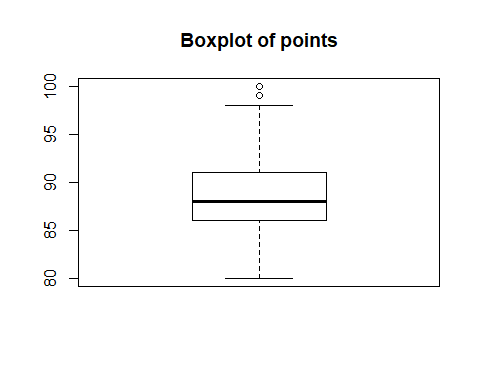
Wine review data is taken form Kaagle [link](https://www.kaggle.com/zynicide/wine-reviews). I choose the second .csv version of this dataset. In this version, duplicated data is removed. The data include 120975 samples about wine. Each sample contains the following information: country, description, 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 isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.  
## 2 This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry 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 pineapple pokes through, with crisp acidity underscoring the flavors. The wine was 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 slightly astringent, semidry finish.  
## 5 Much like the regular bottling from 2012, this comes across as rather rough and tannic, with rustic, earthy, herbal characteristics. Nonetheless, 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 herbs and, in this case, horseradish. In the mouth, this is fairly full bodied, with tomatoey acidity. Spicy, herbal flavors complement dark plum fruit, while 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  Bianco (Etna)  
## 2 Quinta dos Avidagos 2011 Avidagos Red (Douro)  
## 3 Rainstorm 2013 Pinot Gris (Willamette Valley)  
## 4 St. Julian 2013 Reserve Late Harvest Riesling (Lake Michigan Shore)  
## 5 Sweet Cheeks 2012 Vintner's Reserve Wild Child Block Pinot Noir (Willamette Valley)  
## 6 Tandem 2011 Ars In Vitro Tempranillo-Merlot (Navarra)  
## variety winery  
## 1 White Blend Nicosia  
## 2 Portuguese Red Quinta dos Avidagos  
## 3 Pinot Gris Rainstorm  
## 4 Riesling St. Julian  
## 5 Pinot Noir Sweet Cheeks  
## 6 Tempranillo-Merlot Tandem

Points 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 point (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")



## Theory

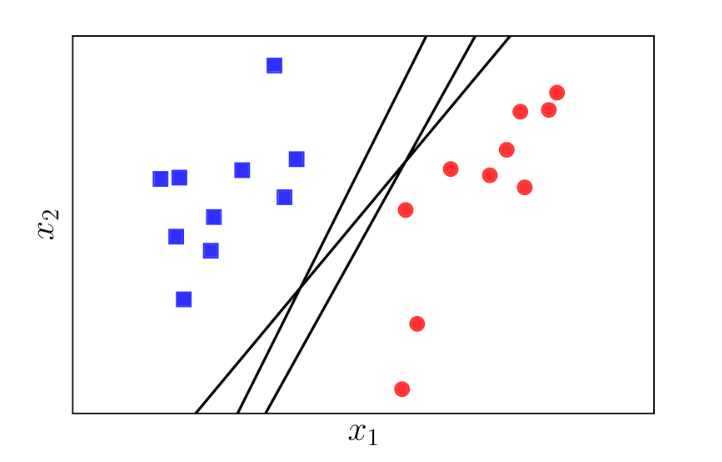
In this section, I will briefly talk about some theory in nature language processing(NLP) and machine learning(ML) in a simple way of reviewing. For some terms, only the aspect that relevant to this project is presented. I assume that the reader already has some basic knowledge about NLP and ML before(this 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 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 word in the bag of words is a corpus.
* Document term matrix(Dtm) is a matrix, which has column is all words of the corpus and column is the document ID. 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:

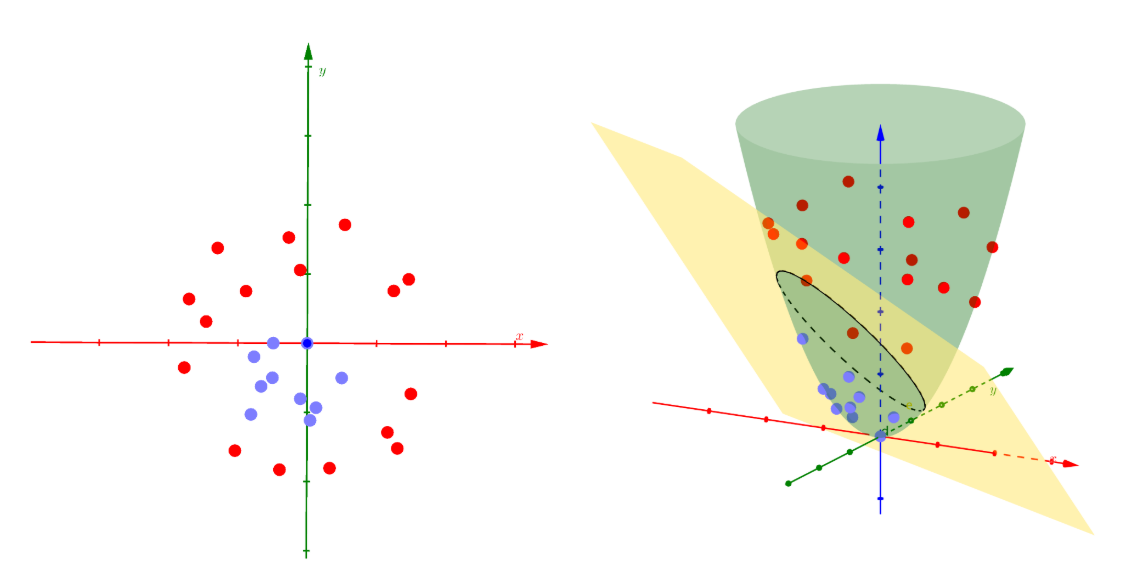
number of documents where the term t appears. If this value equal 0, it will be adjust to 1  
D: total number of documents in the corpus

To be specific, If a word is represented many times in almost documents, it is not so importance, and vice versa.

* 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: Let initial with we have a vector X of a document, which builds in document-term matrix. The probability of assign a class label y to that vector is:
* Naive-Bayes theory make an asump 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. immage 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



SVM RBF kernel

* Ngram: n-gram is n continue sequence words. In this project, n-gram is used for making corpus. An example is n=2 continue words wil be features. To be more specific, 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…

## Method

In this part, I’ll talk about the method for the first problem of this project: Which is the “good” and “excellent” wine. The second problem about wine that 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:
* Split train and test: I use 80% of the data as the training set and 20% as the testing set. Then, 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: In the data set. there is some name of variety that is not unified (same type but written in the different language). So I convert it to the most famous word as follow:
* ++ Replace German names with English names: “weissburgunder” is replaced as “chardonnay”. “spatburgunder” is replaced as “pinot noir”. “grauburgunder” is replaced as “pinot gris”.  
  ++ Replace the Spanish “garnacha”" with the french “grenache”.  
  ++ Replace the Italian “pinot nero”" with the french “pinot noir”.  
  ++ Replace the Portugues “alvarinho”" with the spanish “albarino”.
* Remove non-ASCII characters.
* Remove punction .
* Make words to lower from.
* Remove number.
* Remove stop words: I use the default stop words lists in the tm library and then adjust it. Firstly, the lists have the word “very”. In my opinion, very is a valuable word in this data. Because in the descriptions, for example, “sweet” and “very sweet” are different levels of flavor. So, I remove that word 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:
* Steamming: I steam the word by using SnowballC steam.
* Tonkenzine 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 few documents. That words are not helpful for training because maybe we never meet it again in the testing set.
* Create the document terms matrix base on tf-idf.
* Now the data is ready for training. For testing data, the preprocess is similar, except the Dtm. Dtm of testing data will be create 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. In this step, because of the limit 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 when using n-gram and adjust the weight for some terms using Part-Of-Speech Tagger (POS Tagger).
* Final result: Final model is selected and run again with 100% data. The final result is reported based on this model.

## Result and explain

* Firstly, Let’s takes a look at 5 original paragraphs and the paragraphs after preprocessing to see how it work:

train$description[1:5]

## [1] "Smoky and meaty, this medium-bodied Cab from rocky Atlas Peak offers sensible acidity wrapped around spicy, dusty tannins and a savory aftertaste of tobacco. US Haystack Peak Vineyards California Atlas Peak Napa Cabernet Sauvignon"   
## [2] "Red cherry and plum aromas are citric in style, while the palate is quite citrusy, too, with flavors of plum skin and pie cherry. This finishes tannic and grabby. Argentina Mendoza Province Mendoza Cabernet Sauvignon-Malbec"   
## [3] "Scorched earth, menthol, cooking spice and mature plum aromas take shape in the glass. The forward palate offers dried black cherry, coffee and licorice alongside smooth, polished tannins. Drink soon to capture the remaining fruit flavor. Italy Tuscany Chianti Classico Red Blend"   
## [4] "The wine is dry, and firm with tannins still masking the fruit. With its dark coffee flavor it has been heavily extracted and this needs to soften before the wine will be ready. Look for the berry fruits and acidity as the wine ages. Drink from 2017. Portugal Casa de Cambres Reserva Douro Portuguese Red"  
## [5] "Spring wild flower and green apple aromas lead the nose. The apple note follows over to the racy palate, joining white peach and citrus peel. Brisk acidity and a vivacious perlage offset the fruity flavors. Italy Brut Veneto Glera"

wine\_train\_set[1:5]

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

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

wine\_train\_set[1,]

## accent acid add africa   
## 0.00000000 0.06589887 0.00000000 0.00000000   
## aftertast age alcohol almond   
## 0.21144858 0.00000000 0.00000000 0.00000000   
## almost along alongsid alreadi   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## alsac also although ampl   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## anis anoth appeal appel   
## 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 0.22373411 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   
## 0.00000000 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 broad brunello   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## brut burgundi burst butter   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## cab cabernet california can   
## 0.22520822 0.10774135 0.06468103 0.00000000   
## candi caramel carri cassi   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## catalonia cedar cellar central   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## champagn char charact chardonnay   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## cherri chewi chianti chile   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## chocol chunki cinnamon citrus   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## citrusi classic classico clean   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## close clove coast cocoa   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## coffe cola color columbia   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## combin come complex concentr   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## cool core counti cranberri   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## cream 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 douro dri   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## drink dusti earth earthi   
## 0.00000000 0.20366764 0.00000000 0.00000000   
## easi edg eleg element   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## end enjoy enough espresso   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## estat even excel exot   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## express extra extract fair   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## famili featur feel ferment   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## fill find fine finegrain   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## finger finish firm first   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## flavor fleshi floral flower   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## focus follow food foothil   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## forest forward fragrant frame   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## franc french fresh front   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## fruit fruiti full fullbodi   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## generous gentl germani get   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## give glass good grand   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## grape grapefruit graphit great   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## green grenach grill grip   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## gris grown hard heat   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## heavi herb herbal high   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## highlight hill hint hold   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## honey honeysuckl impress includ   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## integr intens interest intrigu   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## itali jam jammi juic   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## juici just keep lack   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## lake last layer lead   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## leaf lean least leather   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## leav lemon lend length   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## les licoric lift light   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## like lime linger littl   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## live 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 0.20891038   
## 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 mouth mouthfeel much   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## napa natur nebbiolo nectarin   
## 0.15066910 0.00000000 0.00000000 0.00000000   
## need new next nice   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## noir north northeastern northern   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## nose note now nuanc   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## oak oaki offdri offer   
## 0.00000000 0.00000000 0.00000000 0.11163263   
## old oliv one open   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## opul orang oregon overal   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pack pair palat part   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## paso peach pear peel   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pepper pepperi perfect perfum   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## persist petit pie piedmont   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pineappl pink pinot play   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pleasant plenti plum plump   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## plush polish pomegran portug   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## portugues potenti power premier   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pretti price produc provid   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## provinc prune pure qualiti   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## quit raci raisin ranch   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## raspberri rather raw readi   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## red refresh region remain   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## reserv reserva rhnestyl rich   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## riesl right rioja ripe   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## riserva river roast robl   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## robust ros rose round   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## russian sage sangioves santa   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## sardinia sauvignon savori scent   
## 0.00000000 0.09764457 0.17173581 0.00000000   
## seem select sens set   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## sharp short show sicili   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## side sierra silki simpl   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## sip sirah skin slight   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## smell smoke smoki smooth   
## 0.00000000 0.00000000 0.17861632 0.00000000   
## soft soften soil solid   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## somewhat sonoma soon sour   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## south southern southwest spain   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## sparkl spice spici start   
## 0.00000000 0.00000000 0.15747439 0.00000000   
## still stone straightforward strawberri   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## streak strong structur style   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## subtl sugar suggest superior   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## suppl support sweet syrah   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## take tangerin tangi tannic   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## tannin tart tast tea   
## 0.07522370 0.00000000 0.00000000 0.00000000   
## tempranillo textur that there   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## thick though tight time   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## toast toasti tobacco togeth   
## 0.00000000 0.00000000 0.17242706 0.00000000   
## tomato tone top touch   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## tropic turn tuscani two   
## 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 0.00000000   
## vibrant vine vineyard vintag   
## 0.00000000 0.00000000 0.10226734 0.00000000   
## viognier violet walla warm   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## washington way weight well   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## wet whiff white wild   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## will willamett wine winemak   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## wineri without wonder wood   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## wrap year yellow yet   
## 0.24120640 0.00000000 0.00000000 0.00000000   
## york young youth zealand   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## zest zesti zinfandel   
## 0.00000000 0.00000000 0.00000000

And let see the first line of Dtm for testing set see make sure that the prepared data is correct. As we can see, the term remains. Just the points are different.

wine\_test\_set[1,]

## accent acid add africa   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## aftertast age alcohol almond   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## almost along alongsid alreadi   
## 0.00000000 0.17528513 0.00000000 0.00000000   
## alsac also although ampl   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## anis anoth appeal appel   
## 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 0.24650114 0.00000000   
## barrel beauti berri best   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## better big bit bitter   
## 0.00000000 0.00000000 0.16564107 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 0.25115043 0.00000000 0.00000000   
## bring brisk broad brunello   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## brut burgundi burst butter   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## cab cabernet california can   
## 0.00000000 0.00000000 0.06513681 0.00000000   
## candi caramel carri cassi   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## catalonia cedar cellar central   
## 0.00000000 0.00000000 0.00000000 0.23026935   
## champagn char charact chardonnay   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## cherri chewi chianti chile   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## chocol chunki cinnamon citrus   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## citrusi classic classico clean   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## close clove coast cocoa   
## 0.00000000 0.00000000 0.24045480 0.00000000   
## coffe cola color columbia   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## combin come complex concentr   
## 0.20946410 0.00000000 0.00000000 0.00000000   
## cool core counti cranberri   
## 0.00000000 0.00000000 0.00000000 0.19935230   
## cream 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 0.26325872 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 douro dri   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## drink dusti earth earthi   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## easi edg eleg element   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## end enjoy enough espresso   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## estat even excel exot   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## express extra extract fair   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## famili featur feel ferment   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## fill find fine finegrain   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## finger finish firm first   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## flavor fleshi floral flower   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## focus follow food foothil   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## forest forward fragrant frame   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## franc french fresh front   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## fruit fruiti full fullbodi   
## 0.05123295 0.00000000 0.00000000 0.00000000   
## generous gentl germani get   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## give glass good grand   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## grape grapefruit graphit great   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## green grenach grill grip   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## gris grown hard heat   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## heavi herb herbal high   
## 0.00000000 0.13511245 0.00000000 0.00000000   
## highlight hill hint hold   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## honey honeysuckl impress includ   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## integr intens interest intrigu   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## itali jam jammi juic   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## juici just keep lack   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## lake last layer lead   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## leaf lean least leather   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## leav lemon lend length   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## les licoric lift light   
## 0.00000000 0.00000000 0.21116053 0.00000000   
## like lime linger littl   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## live loir long lot   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## love lush made make   
## 0.00000000 0.00000000 0.15541108 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 mouth mouthfeel much   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## napa natur nebbiolo nectarin   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## need new next nice   
## 0.00000000 0.00000000 0.00000000 0.16815013   
## noir north northeastern northern   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## nose note now nuanc   
## 0.10661294 0.00000000 0.00000000 0.00000000   
## oak oaki offdri offer   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## old oliv one open   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## opul orang oregon overal   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pack pair palat part   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## paso peach pear peel   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pepper pepperi perfect perfum   
## 0.00000000 0.20467584 0.00000000 0.00000000   
## persist petit pie piedmont   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pineappl pink pinot play   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pleasant plenti plum plump   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## plush polish pomegran portug   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## portugues potenti power premier   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pretti price produc provid   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## provinc prune pure qualiti   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## quit raci raisin ranch   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## raspberri rather raw readi   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## red refresh region remain   
## 0.07526885 0.00000000 0.00000000 0.00000000   
## reserv reserva rhnestyl rich   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## riesl right rioja ripe   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## riserva river roast robl   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## robust ros rose round   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## russian sage sangioves santa   
## 0.00000000 0.00000000 0.00000000 0.17617324   
## sardinia sauvignon savori scent   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## seem select sens set   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## sharp short show sicili   
## 0.00000000 0.00000000 0.10868265 0.00000000   
## side sierra silki simpl   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## sip sirah skin slight   
## 0.22868163 0.00000000 0.00000000 0.00000000   
## smell smoke smoki smooth   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## soft soften soil solid   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## somewhat sonoma soon sour   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## south southern southwest spain   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## sparkl spice spici start   
## 0.00000000 0.09650675 0.00000000 0.00000000   
## still stone straightforward strawberri   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## streak strong structur style   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## subtl sugar suggest superior   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## suppl support sweet syrah   
## 0.00000000 0.00000000 0.11827294 0.14984209   
## take tangerin tangi tannic   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## tannin tart tast tea   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## tempranillo textur that there   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## thick though tight time   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## toast toasti tobacco togeth   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## tomato tone top touch   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## tropic turn tuscani two   
## 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 0.00000000   
## vibrant vine vineyard vintag   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## viognier violet walla warm   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## washington way weight well   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## wet whiff white wild   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## will willamett wine winemak   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## wineri without wonder wood   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## wrap year yellow yet   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## york young youth zealand   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## zest zesti zinfandel   
## 0.00000000 0.00000000 0.00000000

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

train\_nb\_model

## Naive Bayes   
##   
## 19356 samples  
## 511 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 0.7492364 0.4985562  
## TRUE 0.5908953 0.2162288  
##   
## 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)

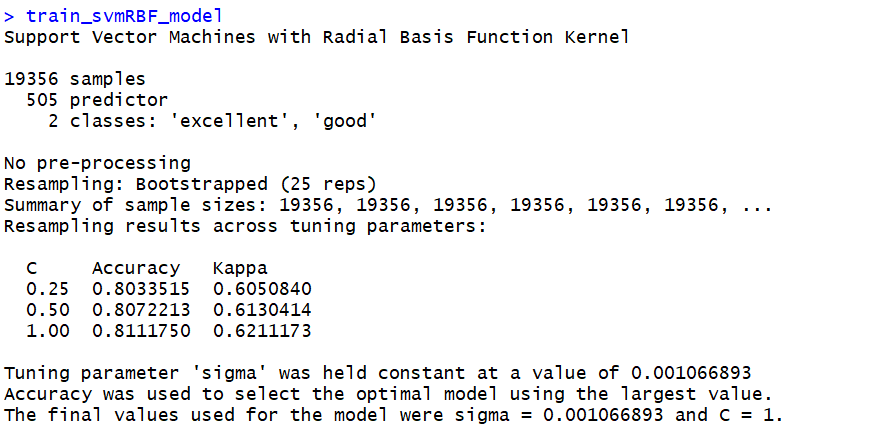
## Confusion Matrix and Statistics  
##   
## Actual class  
## Predicted class excellent good  
## excellent 1575 508  
## good 719 2037  
##   
## Accuracy : 0.7464   
## 95% CI : (0.7339, 0.7586)  
## No Information Rate : 0.5259   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.4892   
## Mcnemar's Test P-Value : 2.034e-09   
##   
## Sensitivity : 0.6866   
## Specificity : 0.8004   
## Pos Pred Value : 0.7561   
## Neg Pred Value : 0.7391   
## Prevalence : 0.4741   
## Detection Rate : 0.3255   
## Detection Prevalence : 0.4305   
## Balanced Accuracy : 0.7435   
##   
## 'Positive' Class : excellent   
##

train\_svmLinear\_model

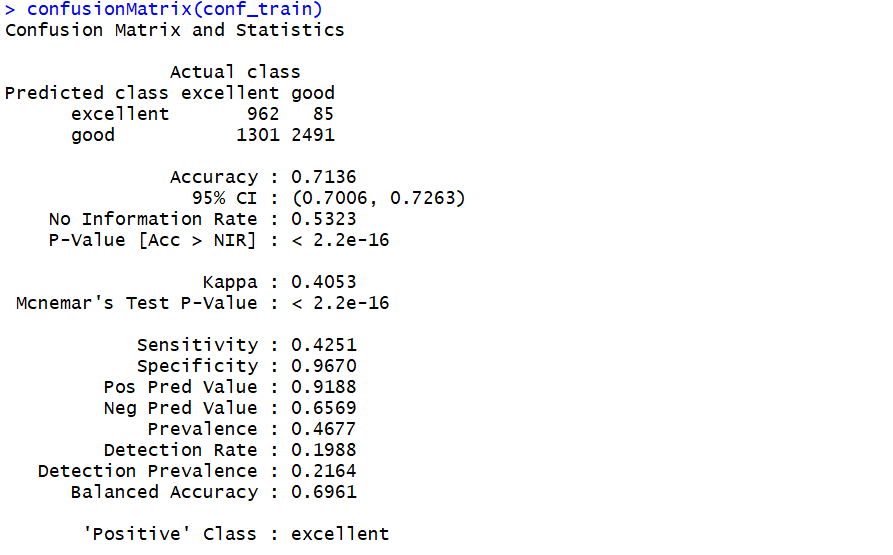
## L2 Regularized Support Vector Machine (dual) with Linear Kernel   
##   
## 19356 samples  
## 511 predictor  
## 2 classes: 'excellent', 'good'   
##   
## No pre-processing  
## Resampling: Bootstrapped (25 reps)   
## Summary of sample sizes: 19356, 19356, 19356, 19356, 19356, 19356, ...   
## Resampling results across tuning parameters:  
##   
## cost Loss Accuracy Kappa   
## 0.25 L1 0.7806989 0.5599345  
## 0.25 L2 0.7836121 0.5657780  
## 0.50 L1 0.7815625 0.5618363  
## 0.50 L2 0.7832133 0.5650346  
## 1.00 L1 0.7810136 0.5607921  
## 1.00 L2 0.7825600 0.5637394  
##   
## 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 1682 432  
## good 612 2113  
##   
## Accuracy : 0.7843   
## 95% CI : (0.7724, 0.7958)  
## No Information Rate : 0.5259   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5657   
## Mcnemar's Test P-Value : 3.026e-08   
##   
## Sensitivity : 0.7332   
## Specificity : 0.8303   
## Pos Pred Value : 0.7956   
## Neg Pred Value : 0.7754   
## Prevalence : 0.4741   
## Detection Rate : 0.3476   
## Detection Prevalence : 0.4369   
## Balanced Accuracy : 0.7817   
##   
## 'Positive' Class : excellent   
##



SVMRBF kenel



SVMRBF result

As we can see in the result, The accuracy when training data for NB, SVM Linear and SVM RBF is 75.01%, 78.64%, and 81.11%. 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. 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 model is actually work.

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 are just lucky and nothing happened. (similar with 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 confusion matrix and the training time for choosing the best model. In my opinion, It’s SVM Linear. (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)

* 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,]

## age\_drink alsac\_alsac appl\_pear   
## 0.0000000 0.0000000 0.0000000   
## aroma\_flavor bake\_spice barolo\_nebbiolo   
## 0.0000000 0.0000000 0.0000000   
## berri\_aroma berri\_flavor berri\_fruit   
## 0.0000000 0.0000000 0.0000000   
## black\_cherri black\_currant black\_fruit   
## 0.0000000 0.0000000 0.0000000   
## black\_pepper black\_plum blackberri\_cherri   
## 0.0000000 0.0000000 0.0000000   
## blend\_cabernet bordeaux\_bordeaux bordeauxstyl\_red   
## 0.0000000 0.0000000 0.0000000   
## bright\_acid brunello\_di cabernet\_franc   
## 0.0000000 0.0000000 0.0000000   
## cabernet\_sauvignon california\_california california\_napa   
## 0.1182729 0.0000000 0.0000000   
## california\_paso california\_russian california\_santa   
## 0.0000000 0.0000000 0.0000000   
## california\_sonoma central\_coast central\_valley   
## 0.0000000 0.0000000 0.0000000   
## champagn\_blend champagn\_champagn cherri\_flavor   
## 0.0000000 0.0000000 0.0000000   
## cherri\_fruit cherri\_raspberri coast\_chardonnay   
## 0.0000000 0.0000000 0.0000000   
## coast\_pinot coast\_sonoma columbia\_valley   
## 0.0000000 0.0000000 0.0000000   
## counti\_central crisp\_acid ctes\_de   
## 0.0000000 0.0000000 0.0000000   
## dark\_chocol dark\_fruit di\_montalcino   
## 0.0000000 0.0000000 0.0000000   
## dri\_herb drink\_franc drink\_itali   
## 0.0000000 0.0000000 0.0000000   
## drink\_now drink\_portug estat\_california   
## 0.0000000 0.0000000 0.0000000   
## estat\_grown finger\_lake finish\_drink   
## 0.0000000 0.0000000 0.0000000   
## finish\_itali finish\_us firm\_tannin   
## 0.0000000 0.0000000 0.0000000   
## flavor\_finish flavor\_us franc\_bordeaux   
## 0.0000000 0.0000000 0.0000000   
## franc\_les french\_oak fresh\_acid   
## 0.0000000 0.0000000 0.0000000   
## fruit\_flavor full\_bodi green\_appl   
## 0.0000000 0.0000000 0.0000000   
## itali\_tuscani lake\_finger lead\_nose   
## 0.0000000 0.0000000 0.0000000   
## linger\_finish loir\_valley long\_finish   
## 0.0000000 0.0000000 0.0000000   
## medium\_bodi mendoza\_provinc montalcino\_sangioves   
## 0.0000000 0.0000000 0.0000000   
## napa\_cabernet napa\_valley new\_york   
## 0.1916446 0.0000000 0.0000000   
## new\_zealand northeastern\_itali northern\_spain   
## 0.0000000 0.0000000 0.0000000   
## nose\_palat now\_franc now\_us   
## 0.0000000 0.0000000 0.0000000   
## old\_vine open\_aroma oregon\_willamett   
## 0.0000000 0.0000000 0.0000000   
## palat\_deliv palat\_offer palat\_show   
## 0.0000000 0.0000000 0.0000000   
## paso\_robl petit\_sirah petit\_verdot   
## 0.0000000 0.0000000 0.0000000   
## piedmont\_barolo pinot\_gris pinot\_noir   
## 0.0000000 0.0000000 0.0000000   
## portugues\_red provinc\_mendoza raspberri\_cherri   
## 0.0000000 0.0000000 0.0000000   
## readi\_drink red\_berri red\_blend   
## 0.0000000 0.0000000 0.0000000   
## red\_cherri red\_currant red\_fruit   
## 0.0000000 0.0000000 0.0000000   
## reserv\_california rhnestyl\_red ripe\_fruit   
## 0.0000000 0.0000000 0.0000000   
## river\_valley robl\_central russian\_river   
## 0.0000000 0.0000000 0.0000000   
## sauvignon\_blanc sicili\_sardinia sierra\_foothil   
## 0.0000000 0.0000000 0.0000000   
## sonoma\_chardonnay sonoma\_coast sonoma\_counti   
## 0.0000000 0.0000000 0.0000000   
## sonoma\_pinot south\_africa south\_australia   
## 0.0000000 0.0000000 0.0000000   
## southern\_itali southwest\_franc spain\_rioja   
## 0.0000000 0.0000000 0.0000000   
## sparkl\_blend spice\_flavor stone\_fruit   
## 0.0000000 0.0000000 0.0000000   
## tannin\_drink tropic\_fruit tuscani\_brunello   
## 0.0000000 0.0000000 0.0000000   
## tuscani\_chianti 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 0.0000000   
## 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 0.1498877 0.0000000   
## wa\_columbia walla\_valley walla\_walla   
## 0.0000000 0.0000000 0.0000000   
## washington\_columbia white\_blend white\_peach   
## 0.0000000 0.0000000 0.0000000   
## white\_pepper wild\_berri willamett\_valley   
## 0.0000000 0.0000000 0.0000000   
## wood\_age year\_us york\_finger   
## 0.0000000 0.0000000 0.0000000

Here is the result:

train\_svmLinear\_model

## L2 Regularized Support Vector Machine (dual) with Linear Kernel   
##   
## 19356 samples  
## 159 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 0.6512283 0.2921140  
## 0.25 L2 0.6498888 0.2973455  
## 0.50 L1 0.6504386 0.2953359  
## 0.50 L2 0.6505031 0.2987642  
## 1.00 L1 0.6500846 0.2975160  
## 1.00 L2 0.6508173 0.2994754  
##   
## 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 = L1.

confusionMatrix(conf\_svmLinear\_train)

## Confusion Matrix and Statistics  
##   
## Actual class  
## Predicted class excellent good  
## excellent 1275 658  
## good 1019 1887  
##   
## Accuracy : 0.6534   
## 95% CI : (0.6398, 0.6669)  
## No Information Rate : 0.5259   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.2996   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.5558   
## Specificity : 0.7415   
## Pos Pred Value : 0.6596   
## Neg Pred Value : 0.6493   
## Prevalence : 0.4741   
## Detection Rate : 0.2635   
## Detection Prevalence : 0.3995   
## Balanced Accuracy : 0.6486   
##   
## 'Positive' Class : excellent   
##

The result is quite bad, which means 2-gram is not useful. Based on my observation. After preprocess and steaming, the description is 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[1,]

## accent acid add   
## 0.00000000 0.03294944 0.00000000   
## africa aftertast age   
## 0.00000000 0.10572429 0.00000000   
## age\_drink alcohol almond   
## 0.00000000 0.00000000 0.00000000   
## almost along alongsid   
## 0.00000000 0.00000000 0.00000000   
## alreadi alsac alsac\_alsac   
## 0.00000000 0.00000000 0.00000000   
## also although ampl   
## 0.00000000 0.00000000 0.00000000   
## anis anoth appeal   
## 0.00000000 0.00000000 0.00000000   
## appel appl appl\_pear   
## 0.00000000 0.00000000 0.00000000   
## approach apricot argentina   
## 0.00000000 0.00000000 0.00000000   
## aroma aroma\_flavor aromat   
## 0.00000000 0.00000000 0.00000000   
## around astring attract   
## 0.11186705 0.00000000 0.00000000   
## australia austria back   
## 0.00000000 0.00000000 0.00000000   
## bake bake\_spice balanc   
## 0.00000000 0.00000000 0.00000000   
## barbara barolo barolo\_nebbiolo   
## 0.00000000 0.00000000 0.00000000   
## barrel beauti berri   
## 0.00000000 0.00000000 0.00000000   
## berri\_aroma berri\_flavor berri\_fruit   
## 0.00000000 0.00000000 0.00000000   
## best better big   
## 0.00000000 0.00000000 0.00000000   
## bit bitter black   
## 0.00000000 0.00000000 0.00000000   
## black\_cherri black\_currant black\_fruit   
## 0.00000000 0.00000000 0.00000000   
## black\_pepper black\_plum blackberri   
## 0.00000000 0.00000000 0.00000000   
## blackberri\_cherri blanc blend   
## 0.00000000 0.00000000 0.00000000   
## blend\_cabernet blossom blue   
## 0.00000000 0.00000000 0.00000000   
## blueberri bodi bold   
## 0.00000000 0.00000000 0.00000000   
## bordeaux bordeaux\_bordeaux bordeauxstyl   
## 0.00000000 0.00000000 0.00000000   
## bordeauxstyl\_red bottl bouquet   
## 0.00000000 0.00000000 0.00000000   
## boysenberri brambl bright   
## 0.00000000 0.00000000 0.00000000   
## bright\_acid bring brisk   
## 0.00000000 0.00000000 0.00000000   
## broad brunello brunello\_di   
## 0.00000000 0.00000000 0.00000000   
## brut burgundi burst   
## 0.00000000 0.00000000 0.00000000   
## butter cab cabernet   
## 0.00000000 0.11260411 0.05387067   
## cabernet\_franc cabernet\_sauvignon california   
## 0.00000000 0.05913647 0.03234051   
## california\_california california\_napa california\_paso   
## 0.00000000 0.00000000 0.00000000   
## california\_russian california\_santa california\_sonoma   
## 0.00000000 0.00000000 0.00000000   
## can candi caramel   
## 0.00000000 0.00000000 0.00000000   
## carri cassi catalonia   
## 0.00000000 0.00000000 0.00000000   
## cedar cellar central   
## 0.00000000 0.00000000 0.00000000   
## central\_coast central\_valley champagn   
## 0.00000000 0.00000000 0.00000000   
## champagn\_blend champagn\_champagn char   
## 0.00000000 0.00000000 0.00000000   
## charact chardonnay cherri   
## 0.00000000 0.00000000 0.00000000   
## cherri\_flavor cherri\_fruit cherri\_raspberri   
## 0.00000000 0.00000000 0.00000000   
## chewi chianti chile   
## 0.00000000 0.00000000 0.00000000   
## chocol chunki cinnamon   
## 0.00000000 0.00000000 0.00000000   
## citrus citrusi classic   
## 0.00000000 0.00000000 0.00000000   
## classico clean close   
## 0.00000000 0.00000000 0.00000000   
## clove coast coast\_chardonnay   
## 0.00000000 0.00000000 0.00000000   
## coast\_pinot coast\_sonoma cocoa   
## 0.00000000 0.00000000 0.00000000   
## coffe cola color   
## 0.00000000 0.00000000 0.00000000   
## columbia columbia\_valley combin   
## 0.00000000 0.00000000 0.00000000   
## come complex concentr   
## 0.00000000 0.00000000 0.00000000   
## cool core counti   
## 0.00000000 0.00000000 0.00000000   
## counti\_central cranberri cream   
## 0.00000000 0.00000000 0.00000000   
## creami creek crisp   
## 0.00000000 0.00000000 0.00000000   
## crisp\_acid cru crush   
## 0.00000000 0.00000000 0.00000000   
## ctes ctes\_de currant   
## 0.00000000 0.00000000 0.00000000   
## cut cuve dark   
## 0.00000000 0.00000000 0.00000000   
## dark\_chocol dark\_fruit deep   
## 0.00000000 0.00000000 0.00000000   
## del delic delici   
## 0.00000000 0.00000000 0.00000000   
## deliv dens depth   
## 0.00000000 0.00000000 0.00000000   
## despit develop di\_montalcino   
## 0.00000000 0.00000000 0.00000000   
## doesnt domin douro   
## 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 0.10183382 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 finegrain   
## 0.00000000 0.00000000 0.00000000   
## finger finger\_lake finish   
## 0.00000000 0.00000000 0.00000000   
## finish\_drink finish\_itali finish\_us   
## 0.00000000 0.00000000 0.00000000   
## firm firm\_tannin first   
## 0.00000000 0.00000000 0.00000000   
## flavor flavor\_finish flavor\_us   
## 0.00000000 0.00000000 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 franc\_les french   
## 0.00000000 0.00000000 0.00000000   
## french\_oak fresh fresh\_acid   
## 0.00000000 0.00000000 0.00000000   
## front fruit fruit\_flavor   
## 0.00000000 0.00000000 0.00000000   
## fruiti full full\_bodi   
## 0.00000000 0.00000000 0.00000000   
## fullbodi generous gentl   
## 0.00000000 0.00000000 0.00000000   
## germani get give   
## 0.00000000 0.00000000 0.00000000   
## glass good grand   
## 0.00000000 0.00000000 0.00000000   
## grape grapefruit graphit   
## 0.00000000 0.00000000 0.00000000   
## great green green\_appl   
## 0.00000000 0.00000000 0.00000000   
## grenach grill grip   
## 0.00000000 0.00000000 0.00000000   
## gris grown hard   
## 0.00000000 0.00000000 0.00000000   
## heat heavi herb   
## 0.00000000 0.00000000 0.00000000   
## herbal high highlight   
## 0.00000000 0.00000000 0.00000000   
## hill hint hold   
## 0.00000000 0.00000000 0.00000000   
## honey honeysuckl impress   
## 0.00000000 0.00000000 0.00000000   
## includ integr intens   
## 0.00000000 0.00000000 0.00000000   
## interest intrigu 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 loir   
## 0.00000000 0.00000000 0.00000000   
## loir\_valley long long\_finish   
## 0.00000000 0.00000000 0.00000000   
## lot love lush   
## 0.00000000 0.00000000 0.00000000   
## made make malbec   
## 0.00000000 0.00000000 0.00000000   
## mango mani mark   
## 0.00000000 0.00000000 0.00000000   
## matur meat medium   
## 0.00000000 0.00000000 0.00000000   
## medium\_bodi mediumbodi melon   
## 0.00000000 0.10445519 0.00000000   
## mendoza mendoza\_provinc merlot   
## 0.00000000 0.00000000 0.00000000   
## midpal mild miner   
## 0.00000000 0.00000000 0.00000000   
## mint mix mocha   
## 0.00000000 0.00000000 0.00000000   
## moder montalcino montalcino\_sangioves   
## 0.00000000 0.00000000 0.00000000   
## month mountain mouth   
## 0.00000000 0.00000000 0.00000000   
## mouthfeel much napa   
## 0.00000000 0.00000000 0.07533455   
## napa\_cabernet napa\_valley natur   
## 0.09582230 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 offdri   
## 0.00000000 0.00000000 0.00000000   
## offer old old\_vine   
## 0.05581632 0.00000000 0.00000000   
## oliv one open   
## 0.00000000 0.00000000 0.00000000   
## open\_aroma opul orang   
## 0.00000000 0.00000000 0.00000000   
## oregon oregon\_willamett overal   
## 0.00000000 0.00000000 0.00000000   
## pack pair palat   
## 0.00000000 0.00000000 0.00000000   
## palat\_deliv palat\_offer palat\_show   
## 0.00000000 0.00000000 0.00000000   
## part paso paso\_robl   
## 0.00000000 0.00000000 0.00000000   
## peach pear peel   
## 0.00000000 0.00000000 0.00000000   
## pepper pepperi perfect   
## 0.00000000 0.00000000 0.00000000   
## perfum persist petit   
## 0.00000000 0.00000000 0.00000000   
## petit\_sirah petit\_verdot pie   
## 0.00000000 0.00000000 0.00000000   
## piedmont piedmont\_barolo pineappl   
## 0.00000000 0.00000000 0.00000000   
## pink 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 0.00000000   
## plush polish pomegran   
## 0.00000000 0.00000000 0.00000000   
## portug portugues portugues\_red   
## 0.00000000 0.00000000 0.00000000   
## potenti power premier   
## 0.00000000 0.00000000 0.00000000   
## pretti price produc   
## 0.00000000 0.00000000 0.00000000   
## provid provinc provinc\_mendoza   
## 0.00000000 0.00000000 0.00000000   
## prune pure 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 raw readi   
## 0.00000000 0.00000000 0.00000000   
## readi\_drink red red\_berri   
## 0.00000000 0.00000000 0.00000000   
## red\_blend red\_cherri red\_currant   
## 0.00000000 0.00000000 0.00000000   
## red\_fruit refresh region   
## 0.00000000 0.00000000 0.00000000   
## remain reserv reserv\_california   
## 0.00000000 0.00000000 0.00000000   
## reserva 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   
## robust ros rose   
## 0.00000000 0.00000000 0.00000000   
## round russian russian\_river   
## 0.00000000 0.00000000 0.00000000   
## sage sangioves santa   
## 0.00000000 0.00000000 0.00000000   
## sardinia sauvignon sauvignon\_blanc   
## 0.00000000 0.04882228 0.00000000   
## savori scent seem   
## 0.08586791 0.00000000 0.00000000   
## select sens set   
## 0.00000000 0.00000000 0.00000000   
## sharp short 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 0.00000000   
## smoke smoki smooth   
## 0.00000000 0.08930816 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   
## south south\_africa south\_australia   
## 0.00000000 0.00000000 0.00000000   
## southern southern\_itali 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 0.00000000   
## spice\_flavor spici start   
## 0.00000000 0.07873720 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   
## 0.03761185 0.00000000 0.00000000   
## tast tea tempranillo   
## 0.00000000 0.00000000 0.00000000   
## textur that there   
## 0.00000000 0.00000000 0.00000000   
## thick though tight   
## 0.00000000 0.00000000 0.00000000   
## time toast toasti   
## 0.00000000 0.00000000 0.00000000   
## tobacco togeth tomato   
## 0.08621353 0.00000000 0.00000000   
## tone top touch   
## 0.00000000 0.00000000 0.00000000   
## tropic tropic\_fruit turn   
## 0.00000000 0.00000000 0.00000000   
## tuscani tuscani\_brunello tuscani\_chianti   
## 0.00000000 0.00000000 0.00000000   
## two us\_california us\_estat   
## 0.00000000 0.00000000 0.00000000   
## us\_oregon us\_reserv us\_washington   
## 0.00000000 0.00000000 0.00000000   
## valley valley\_cabernet valley\_central   
## 0.00000000 0.00000000 0.00000000   
## valley\_chardonnay valley\_napa valley\_pinot   
## 0.00000000 0.00000000 0.00000000   
## valley\_red valley\_sonoma valley\_syrah   
## 0.00000000 0.00000000 0.00000000   
## valley\_wa valley\_willamett vanilla   
## 0.00000000 0.00000000 0.00000000   
## variet varieti velveti   
## 0.00000000 0.00000000 0.00000000   
## veneto verdot veri   
## 0.00000000 0.00000000 0.00000000   
## vibrant vine vineyard   
## 0.00000000 0.00000000 0.05113367   
## vineyard\_california vineyard\_washington vintag   
## 0.07494385 0.00000000 0.00000000   
## viognier violet wa\_columbia   
## 0.00000000 0.00000000 0.00000000   
## walla walla\_valley walla\_walla   
## 0.00000000 0.00000000 0.00000000   
## warm washington washington\_columbia   
## 0.00000000 0.00000000 0.00000000   
## way weight well   
## 0.00000000 0.00000000 0.00000000   
## wet whiff white   
## 0.00000000 0.00000000 0.00000000   
## white\_blend white\_peach white\_pepper   
## 0.00000000 0.00000000 0.00000000   
## wild wild\_berri will   
## 0.00000000 0.00000000 0.00000000   
## willamett willamett\_valley wine   
## 0.00000000 0.00000000 0.00000000   
## winemak wineri without   
## 0.00000000 0.00000000 0.00000000   
## wonder wood wood\_age   
## 0.00000000 0.00000000 0.00000000   
## wrap year year\_us   
## 0.12060320 0.00000000 0.00000000   
## yellow yet york   
## 0.00000000 0.00000000 0.00000000   
## york\_finger young youth   
## 0.00000000 0.00000000 0.00000000   
## zealand zest zesti   
## 0.00000000 0.00000000 0.00000000   
## zinfandel   
## 0.00000000

And here is the result:

train\_svmLinear\_model

## L2 Regularized Support Vector Machine (dual) with Linear Kernel   
##   
## 19356 samples  
## 670 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 0.7782424 0.5546786  
## 0.25 L2 0.7872156 0.5728222  
## 0.50 L1 0.7835761 0.5656332  
## 0.50 L2 0.7877391 0.5739812  
## 1.00 L1 0.7856626 0.5700116  
## 1.00 L2 0.7870588 0.5726723  
##   
## 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 1608 363  
## good 686 2182  
##   
## Accuracy : 0.7832   
## 95% CI : (0.7713, 0.7948)  
## No Information Rate : 0.5259   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5622   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.7010   
## Specificity : 0.8574   
## Pos Pred Value : 0.8158   
## Neg Pred Value : 0.7608   
## Prevalence : 0.4741   
## Detection Rate : 0.3323   
## Detection Prevalence : 0.4073   
## Balanced Accuracy : 0.7792   
##   
## 'Positive' Class : excellent   
##

The accuracy is slightly decreased. Good news is the False negative is slightly better. It’s like the trade-off between models. Consider that and the dimension of Dtm is large (which lead to high training time), while there are too many features with 0 point. I conclude that 1-2 gram is not so useful. Overall, n-gram is not a way to improve the model.

* The other attempt to improve is adjusted the weight for Dtm. As I observed in the data. The description usually talk about ingredients: lemon, cherry or region: Veneto… So, the first idea is double the point for nouns and see how it works. On the other hand, Adjective used to describe the flavor of wine could make an important role. So, the second ideal is double the point for adj only. To do these. I implement POS tagger and see which word is nouns, adj, verb…

Here are the list nouns that I extracted:`

nouns

## [1] "accent" "acid" "age" "alcohol"   
## [5] "almond" "alreadi" "alsac" "appeal"   
## [9] "appel" "appl" "approach" "apricot"   
## [13] "argentina" "aroma" "austria" "bake"   
## [17] "balanc" "barbara" "barolo" "barrel"   
## [21] "beauti" "berri" "bit" "blanc"   
## [25] "blend" "bodi" "bordeaux" "bordeauxstyl"  
## [29] "bottl" "bouquet" "boysenberri" "brambl"   
## [33] "brunello" "brut" "burgundi" "butter"   
## [37] "cab" "cabernet" "california" "caramel"   
## [41] "carri" "cassi" "catalonia" "cedar"   
## [45] "cellar" "champagn" "char" "charact"   
## [49] "chardonnay" "cherri" "chewi" "chocol"   
## [53] "chunki" "cinnamon" "citrus" "classico"   
## [57] "clove" "coast" "cocoa" "coffe"   
## [61] "cola" "color" "columbia" "combin"   
## [65] "concentr" "core" "counti" "cranberri"   
## [69] "cream" "creami" "creek" "crisp"   
## [73] "cru" "crush" "cut" "delic"   
## [77] "depth" "despit" "domin" "douro"   
## [81] "dri" "drink" "earth" "earthi"   
## [85] "eleg" "element" "end" "espresso"   
## [89] "estat" "express" "ferment" "fill"   
## [93] "finegrain" "finger" "firm" "flavor"   
## [97] "flower" "focus" "food" "foothil"   
## [101] "forest" "frame" "franc" "french"   
## [105] "front" "fruit" "fruiti" "gentl"   
## [109] "glass" "grape" "grapefruit" "grenach"   
## [113] "grill" "grip" "heat" "herb"   
## [117] "high" "highlight" "hill" "hint"   
## [121] "honey" "honeysuckl" "integr" "interest"   
## [125] "jam" "jammi" "lack" "lake"   
## [129] "layer" "lead" "leaf" "leather"   
## [133] "leav" "lemon" "length" "lift"   
## [137] "light" "lime" "linger" "loir"   
## [141] "mango" "mark" "matur" "meat"   
## [145] "medium" "melon" "mendoza" "merlot"   
## [149] "midpal" "miner" "mint" "mix"   
## [153] "mocha" "moder" "month" "mountain"   
## [157] "mouth" "mouthfeel" "napa" "natur"   
## [161] "nebbiolo" "nectarin" "noir" "nose"   
## [165] "oak" "oaki" "offdri" "one"   
## [169] "opul" "oregon" "pack" "pair"   
## [173] "palat" "part" "peel" "pepper"   
## [177] "pepperi" "perfum" "pie" "piedmont"   
## [181] "pineappl" "pinot" "play" "plum"   
## [185] "polish" "pomegran" "portug" "power"   
## [189] "premier" "price" "produc" "provinc"   
## [193] "prune" "qualiti" "ranch" "raspberri"   
## [197] "readi" "region" "reserv" "reserva"   
## [201] "rhnestyl" "rioja" "riserva" "river"   
## [205] "roast" "sage" "sardinia" "sauvignon"   
## [209] "scent" "show" "side" "sierra"   
## [213] "silki" "sip" "skin" "smell"   
## [217] "smoke" "soil" "sonoma" "spain"   
## [221] "spice" "stone" "streak" "structur"   
## [225] "style" "sugar" "support" "syrah"   
## [229] "tangerin" "tannin" "tart" "tast"   
## [233] "tea" "textur" "time" "toast"   
## [237] "toasti" "tobacco" "tone" "touch"   
## [241] "tropic" "turn" "valley" "vanilla"   
## [245] "varieti" "velveti" "veneto" "verdot"   
## [249] "veri" "vine" "vineyard" "vintag"   
## [253] "viognier" "violet" "walla" "way"   
## [257] "weight" "will" "wine" "winemak"   
## [261] "wineri" "wonder" "wood" "wrap"   
## [265] "year" "yellow" "york" "youth"   
## [269] "zealand" "zest" "zesti" "zinfandel"

Here is the adjective list:

adj

## [1] "aftertast" "ampl" "australia"   
## [4] "big" "bitter" "black"   
## [7] "blackberri" "blue" "blueberri"   
## [10] "bold" "bright" "brisk"   
## [13] "broad" "central" "classic"   
## [16] "clean" "close" "complex"   
## [19] "cool" "currant" "cuve"   
## [22] "dark" "deep" "doesnt"   
## [25] "edg" "enough" "exot"   
## [28] "extra" "fair" "fine"   
## [31] "floral" "fragrant" "fresh"   
## [34] "full" "generous" "good"   
## [37] "grand" "great" "green"   
## [40] "hard" "juic" "last"   
## [43] "lean" "licoric" "live"   
## [46] "lush" "malbec" "mild"   
## [49] "new" "next" "nice"   
## [52] "northeastern" "northern" "nuanc"   
## [55] "old" "oliv" "open"   
## [58] "overal" "pear" "perfect"   
## [61] "petit" "pink" "pleasant"   
## [64] "plush" "pure" "raw"   
## [67] "red" "refresh" "rich"   
## [70] "right" "ripe" "robust"   
## [73] "russian" "select" "sharp"   
## [76] "short" "sicili" "simpl"   
## [79] "slight" "smooth" "soft"   
## [82] "solid" "sour" "southern"   
## [85] "sparkl" "straightforward" "strawberri"   
## [88] "strong" "superior" "suppl"   
## [91] "sweet" "tannic" "tempranillo"   
## [94] "thick" "tight" "top"   
## [97] "variet" "vibrant" "warm"   
## [100] "wet" "white" "wild"   
## [103] "young"

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 finger is 0.32…

wine\_train\_set[1,]

## accent acid add africa   
## 0.0000000 0.1317977 0.0000000 0.0000000   
## aftertast age alcohol almond   
## 0.2114486 0.0000000 0.0000000 0.0000000   
## almost along alongsid alreadi   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## alsac also although ampl   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## anis anoth appeal appel   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## appl approach apricot argentina   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## aroma aromat around astring   
## 0.0000000 0.0000000 0.2237341 0.0000000   
## attract australia austria back   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## bake balanc barbara barolo   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## barrel beauti berri best   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## better big bit bitter   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## black blackberri blanc blend   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## blossom blue blueberri bodi   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## bold bordeaux bordeauxstyl bottl   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## bouquet boysenberri brambl bright   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## bring brisk broad brunello   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## brut burgundi burst butter   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## cab cabernet california can   
## 0.4504164 0.2154827 0.1293621 0.0000000   
## candi caramel carri cassi   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## catalonia cedar cellar central   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## champagn char charact chardonnay   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## cherri chewi chianti chile   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## chocol chunki cinnamon citrus   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## citrusi classic classico clean   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## close clove coast cocoa   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## coffe cola color columbia   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## combin come complex concentr   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## cool core counti cranberri   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## cream creami creek crisp   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## cru crush ctes currant   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## cut cuve dark deep   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## del delic delici deliv   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## dens depth despit develop   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## doesnt domin douro dri   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## drink dusti earth earthi   
## 0.0000000 0.2036676 0.0000000 0.0000000   
## easi edg eleg element   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## end enjoy enough espresso   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## estat even excel exot   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## express extra extract fair   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## famili featur feel ferment   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## fill find fine finegrain   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## finger finish firm first   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## flavor fleshi floral flower   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## focus follow food foothil   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## forest forward fragrant frame   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## franc french fresh front   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## fruit fruiti full fullbodi   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## generous gentl germani get   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## give glass good grand   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## grape grapefruit graphit great   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## green grenach grill grip   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## gris grown hard heat   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## heavi herb herbal high   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## highlight hill hint hold   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## honey honeysuckl impress includ   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## integr intens interest intrigu   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## itali jam jammi juic   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## juici just keep lack   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## lake last layer lead   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## leaf lean least leather   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## leav lemon lend length   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## les licoric lift light   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## like lime linger littl   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## live loir long lot   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## love lush made make   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## malbec mango mani mark   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## matur meat medium mediumbodi   
## 0.0000000 0.0000000 0.0000000 0.2089104   
## melon mendoza merlot midpal   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## mild miner mint mix   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## mocha moder montalcino month   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## mountain mouth mouthfeel much   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## napa natur nebbiolo nectarin   
## 0.3013382 0.0000000 0.0000000 0.0000000   
## need new next nice   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## noir north northeastern northern   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## nose note now nuanc   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## oak oaki offdri offer   
## 0.0000000 0.0000000 0.0000000 0.1116326   
## old oliv one open   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## opul orang oregon overal   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## pack pair palat part   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## paso peach pear peel   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## pepper pepperi perfect perfum   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## persist petit pie piedmont   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## pineappl pink pinot play   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## pleasant plenti plum plump   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## plush polish pomegran portug   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## portugues potenti power premier   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## pretti price produc provid   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## provinc prune pure qualiti   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## quit raci raisin ranch   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## raspberri rather raw readi   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## red refresh region remain   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## reserv reserva rhnestyl rich   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## riesl right rioja ripe   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## riserva river roast robl   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## robust ros rose round   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## russian sage sangioves santa   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## sardinia sauvignon savori scent   
## 0.0000000 0.1952891 0.1717358 0.0000000   
## seem select sens set   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## sharp short show sicili   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## side sierra silki simpl   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## sip sirah skin slight   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## smell smoke smoki smooth   
## 0.0000000 0.0000000 0.1786163 0.0000000   
## soft soften soil solid   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## somewhat sonoma soon sour   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## south southern southwest spain   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## sparkl spice spici start   
## 0.0000000 0.0000000 0.1574744 0.0000000   
## still stone straightforward strawberri   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## streak strong structur style   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## subtl sugar suggest superior   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## suppl support sweet syrah   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## take tangerin tangi tannic   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## tannin tart tast tea   
## 0.1504474 0.0000000 0.0000000 0.0000000   
## tempranillo textur that there   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## thick though tight time   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## toast toasti tobacco togeth   
## 0.0000000 0.0000000 0.3448541 0.0000000   
## tomato tone top touch   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## tropic turn tuscani two   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## valley vanilla variet varieti   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## velveti veneto verdot veri   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## vibrant vine vineyard vintag   
## 0.0000000 0.0000000 0.2045347 0.0000000   
## viognier violet walla warm   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## washington way weight well   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## wet whiff white wild   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## will willamett wine winemak   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## wineri without wonder wood   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## wrap year yellow yet   
## 0.4824128 0.0000000 0.0000000 0.0000000   
## york young youth zealand   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## zest zesti zinfandel   
## 0.0000000 0.0000000 0.0000000

Here is the resut:

train\_svmLinear\_model

## L2 Regularized Support Vector Machine (dual) with Linear Kernel   
##   
## 19356 samples  
## 511 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 0.7796407 0.5579149  
## 0.25 L2 0.7830596 0.5646641  
## 0.50 L1 0.7800758 0.5588823  
## 0.50 L2 0.7825889 0.5637570  
## 1.00 L1 0.7805282 0.5599146  
## 1.00 L2 0.7818367 0.5622836  
##   
## 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 1661 404  
## good 633 2141  
##   
## Accuracy : 0.7857   
## 95% CI : (0.7739, 0.7972)  
## No Information Rate : 0.5259   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5681   
## Mcnemar's Test P-Value : 1.439e-12   
##   
## Sensitivity : 0.7241   
## Specificity : 0.8413   
## Pos Pred Value : 0.8044   
## Neg Pred Value : 0.7718   
## Prevalence : 0.4741   
## Detection Rate : 0.3433   
## Detection Prevalence : 0.4267   
## Balanced Accuracy : 0.7827   
##   
## 'Positive' Class : excellent   
##

We can see that accuracy is slightly increasing (just 0.3%). FN is slightly decreasing. Although the number of FP is increasing but overall, I conclude that the model is a little bit better.

Now double the point only for adjectives: 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 finger is 0.32…

wine\_train\_set[1,]

## accent acid add africa   
## 0.00000000 0.06589887 0.00000000 0.00000000   
## aftertast age alcohol almond   
## 0.42289717 0.00000000 0.00000000 0.00000000   
## almost along alongsid alreadi   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## alsac also although ampl   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## anis anoth appeal appel   
## 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 0.22373411 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   
## 0.00000000 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 broad brunello   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## brut burgundi burst butter   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## cab cabernet california can   
## 0.22520822 0.10774135 0.06468103 0.00000000   
## candi caramel carri cassi   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## catalonia cedar cellar central   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## champagn char charact chardonnay   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## cherri chewi chianti chile   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## chocol chunki cinnamon citrus   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## citrusi classic classico clean   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## close clove coast cocoa   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## coffe cola color columbia   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## combin come complex concentr   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## cool core counti cranberri   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## cream 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 douro dri   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## drink dusti earth earthi   
## 0.00000000 0.20366764 0.00000000 0.00000000   
## easi edg eleg element   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## end enjoy enough espresso   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## estat even excel exot   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## express extra extract fair   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## famili featur feel ferment   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## fill find fine finegrain   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## finger finish firm first   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## flavor fleshi floral flower   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## focus follow food foothil   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## forest forward fragrant frame   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## franc french fresh front   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## fruit fruiti full fullbodi   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## generous gentl germani get   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## give glass good grand   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## grape grapefruit graphit great   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## green grenach grill grip   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## gris grown hard heat   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## heavi herb herbal high   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## highlight hill hint hold   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## honey honeysuckl impress includ   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## integr intens interest intrigu   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## itali jam jammi juic   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## juici just keep lack   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## lake last layer lead   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## leaf lean least leather   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## leav lemon lend length   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## les licoric lift light   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## like lime linger littl   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## live 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 0.20891038   
## 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 mouth mouthfeel much   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## napa natur nebbiolo nectarin   
## 0.15066910 0.00000000 0.00000000 0.00000000   
## need new next nice   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## noir north northeastern northern   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## nose note now nuanc   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## oak oaki offdri offer   
## 0.00000000 0.00000000 0.00000000 0.11163263   
## old oliv one open   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## opul orang oregon overal   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pack pair palat part   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## paso peach pear peel   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pepper pepperi perfect perfum   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## persist petit pie piedmont   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pineappl pink pinot play   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pleasant plenti plum plump   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## plush polish pomegran portug   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## portugues potenti power premier   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## pretti price produc provid   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## provinc prune pure qualiti   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## quit raci raisin ranch   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## raspberri rather raw readi   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## red refresh region remain   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## reserv reserva rhnestyl rich   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## riesl right rioja ripe   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## riserva river roast robl   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## robust ros rose round   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## russian sage sangioves santa   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## sardinia sauvignon savori scent   
## 0.00000000 0.09764457 0.17173581 0.00000000   
## seem select sens set   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## sharp short show sicili   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## side sierra silki simpl   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## sip sirah skin slight   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## smell smoke smoki smooth   
## 0.00000000 0.00000000 0.17861632 0.00000000   
## soft soften soil solid   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## somewhat sonoma soon sour   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## south southern southwest spain   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## sparkl spice spici start   
## 0.00000000 0.00000000 0.15747439 0.00000000   
## still stone straightforward strawberri   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## streak strong structur style   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## subtl sugar suggest superior   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## suppl support sweet syrah   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## take tangerin tangi tannic   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## tannin tart tast tea   
## 0.07522370 0.00000000 0.00000000 0.00000000   
## tempranillo textur that there   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## thick though tight time   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## toast toasti tobacco togeth   
## 0.00000000 0.00000000 0.17242706 0.00000000   
## tomato tone top touch   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## tropic turn tuscani two   
## 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 0.00000000   
## vibrant vine vineyard vintag   
## 0.00000000 0.00000000 0.10226734 0.00000000   
## viognier violet walla warm   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## washington way weight well   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## wet whiff white wild   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## will willamett wine winemak   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## wineri without wonder wood   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## wrap year yellow yet   
## 0.24120640 0.00000000 0.00000000 0.00000000   
## york young youth zealand   
## 0.00000000 0.00000000 0.00000000 0.00000000   
## zest zesti zinfandel   
## 0.00000000 0.00000000 0.00000000

And here is the result:

train\_svmLinear\_model

## L2 Regularized Support Vector Machine (dual) with Linear Kernel   
##   
## 19356 samples  
## 511 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 0.7785955 0.5558134  
## 0.25 L2 0.7821534 0.5626976  
## 0.50 L1 0.7791757 0.5570207  
## 0.50 L2 0.7824386 0.5632690  
## 1.00 L1 0.7794546 0.5575787  
## 1.00 L2 0.7823650 0.5631409  
##   
## 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 1662 413  
## good 632 2132  
##   
## Accuracy : 0.784   
## 95% CI : (0.7722, 0.7956)  
## No Information Rate : 0.5259   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5649   
## Mcnemar's Test P-Value : 1.544e-11   
##   
## Sensitivity : 0.7245   
## Specificity : 0.8377   
## Pos Pred Value : 0.8010   
## Neg Pred Value : 0.7713   
## Prevalence : 0.4741   
## Detection Rate : 0.3435   
## Detection Prevalence : 0.4288   
## Balanced Accuracy : 0.7811   
##   
## 'Positive' Class : excellent   
##

The accuracy is just increasing 0.2%. But the FN is decreasing 29 units, quite impressive. TP for the excellent class is higher. Although FN is also higher. **I conclude that the best model for “good” - “excellent” wine classify is SVM linear with double tf-idf point for adjectives.**

To explain why double the point for adj is better than double the point for nouns. When reviewing the dataset. I see there are two reasons: Firstly, most important nouns is the ingredients, country, region… They can use the same ingredients, but the experts may have some unique techniques to make it become excellent wine. So, nouns are not so helpful. But adjective, which means the tease or the properties of the wines will show clearly the difference between good and excellent wine. Secondly, because I steam when process data. Many original words are steamed to nouns but it is not the nouns in the original paragraph. (we can see that the number of nouns is higher than the number of adjectives) Which leads to the method is not as useful as expected.

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

train\_svmLinear\_model\_final1

## L2 Regularized Support Vector Machine (dual) with Linear Kernel   
##   
## 96780 samples  
## 508 predictor  
## 2 classes: 'excellent', 'good'   
##   
## No pre-processing  
## Resampling: Bootstrapped (25 reps)   
## Summary of sample sizes: 96780, 96780, 96780, 96780, 96780, 96780, ...   
## Resampling results across tuning parameters:  
##   
## cost Loss Accuracy Kappa   
## 0.25 L1 0.7916149 0.5817861  
## 0.25 L2 0.7933132 0.5850904  
## 0.50 L1 0.7925280 0.5836351  
## 0.50 L2 0.7934785 0.5854305  
## 1.00 L1 0.7930683 0.5847101  
## 1.00 L2 0.7934808 0.5854413  
##   
## 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\_final1)

## Confusion Matrix and Statistics  
##   
## Actual class  
## Predicted class excellent good  
## excellent 7898 1846  
## good 3435 11016  
##   
## Accuracy : 0.7817   
## 95% CI : (0.7765, 0.7869)  
## No Information Rate : 0.5316   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.558   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.6969   
## Specificity : 0.8565   
## Pos Pred Value : 0.8106   
## Neg Pred Value : 0.7623   
## Prevalence : 0.4684   
## Detection Rate : 0.3264   
## Detection Prevalence : 0.4027   
## Balanced Accuracy : 0.7767   
##   
## 'Positive' Class : excellent   
##

# Second aspect: High value wine

## The problem

In most of the time, the price of the bottle can reflect the quality of the wine. For example, If 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 solveable. Let see the scatter plot of price vs points for wines 100$ and under:

As we can see. There is a positive relation 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 point while many 25 dollar wine just have less than 85 points.

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

## Aim

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

## Method and result

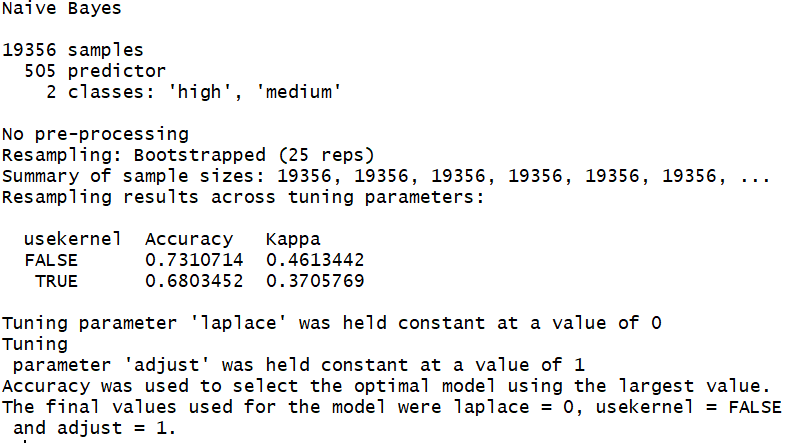
Let’s take a look at the price of our data:

Because the price range quite big. Prices have a lot of outliers and high value of SD. otherwise, the log of the price looks better and quite similar to the point.

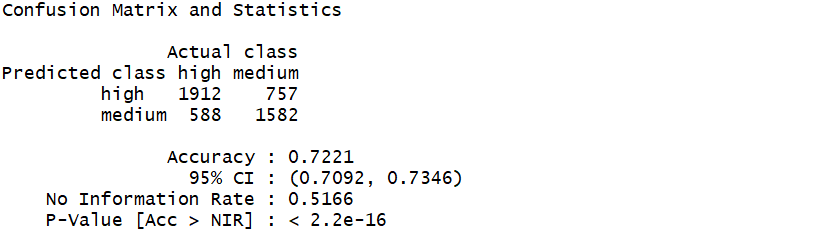
I define the Here is the value:

The median number for value is ~27. So the bottle with a higher value than 27 will have a “high” benefit. The rest have “medium” benefit. This way of define class will make all data become interesting. For example. there is a bottle with just 4$ but have 80 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.

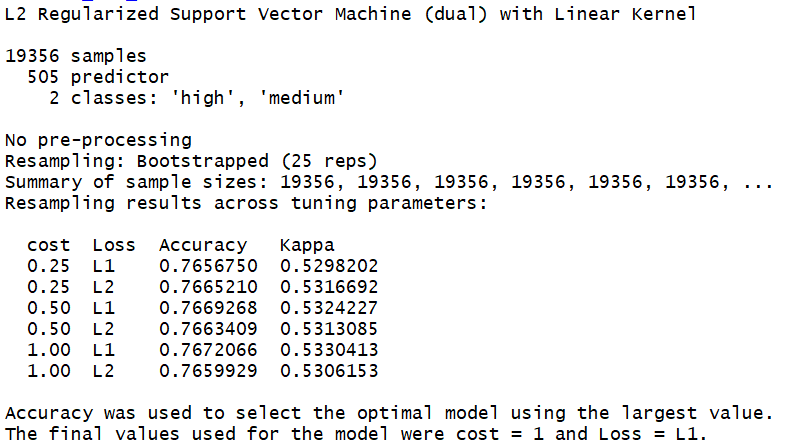
I do the same preprocessing for the text. Then continue to compare models with sample 20% of the data. Here is the result.



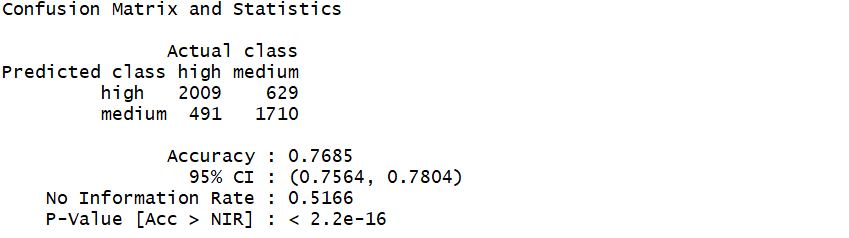
NB benefit



NB benefit predict

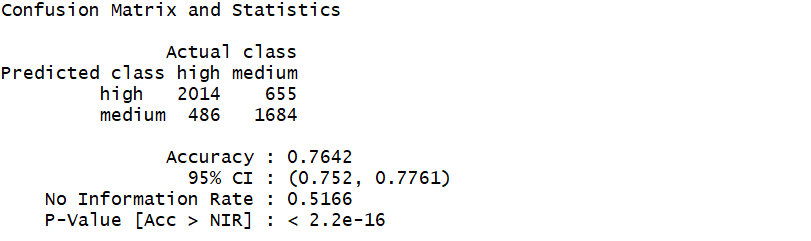


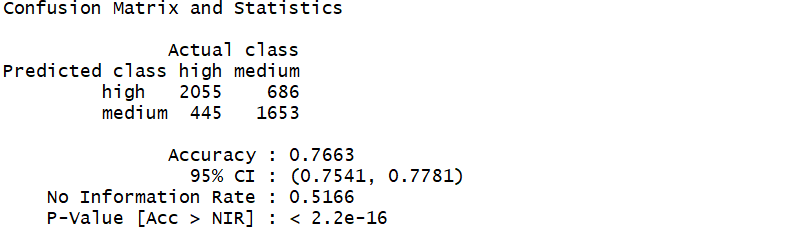
SVM benefit



SVMLinear\_benefit

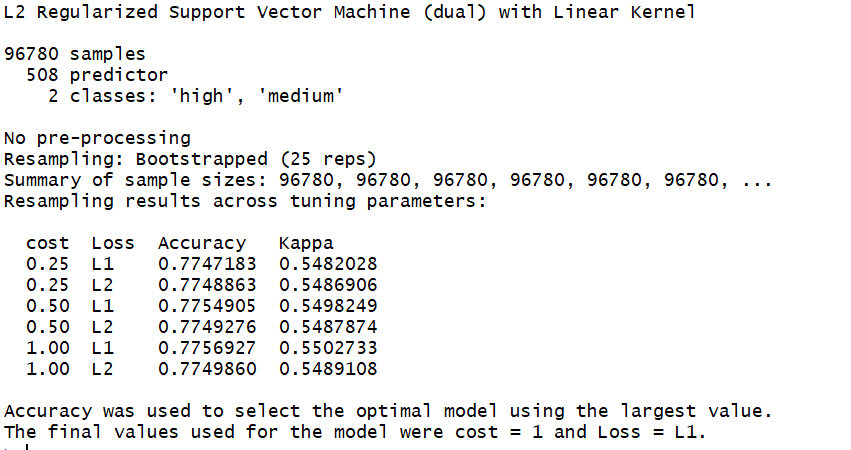
As we can see, SVM Linear is better than Nb in both accuracy and how the bottle is classified. Now I’ll try with POS Tagger

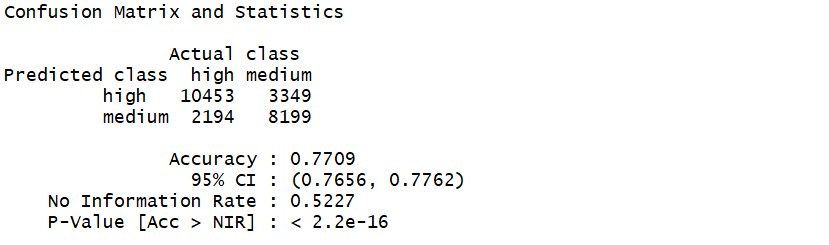
Here is the result when double the point for nouns: 

And here is the result when double the point for adjectives: 

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 is quite strange. I still don’t too much ideal about how the description should be for each class.

**I conclude that the final model for “high” - “medium” wine value classify is SVM linear with tf-idf, bag of words**

Here is the final result for this model: 



SVM benefit predict full data

# Futher discusssion

## Suggession 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 is the result from the SVM Linear for “good” - “excellent” wine model: The first cloud is for good wine and the second one is for excellent wine

wordcloud(rownames(excellent), excellent[,1], max.words=100, colors=brewer.pal(8, "Dark2"))



wordcloud(rownames(good), good[,1], max.words=100, colors=brewer.pal(8, "Dark2"), main= "Top tfidf good class")

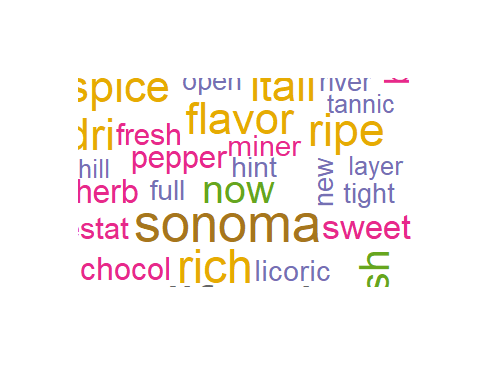
 Suggesstion for choosing wine:…

Here is the top influences 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, "Dark2"))



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



Suggesstion for choosing wine:…

## 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.
* 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 about improve the model is limited.
* Improve the POS tagger function.

# Conclusion

The issues raised at the beginning are solved with a good result. We figure out that the description and other information printed in a bottle can bring us some ideal 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 satified bottle. In the NLP and ML aspect. SVM linear with double tf-idf point for the adjective is the best model for predicting the quality with 78.24% accuracy. For predicting the benefit of a bottle, SVM linear is an acceptable model, with accuracy 77.09%.

# References

1. Slide of this course
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>

# Apendix

Here is the code used in this project. For some models that require a long time for trainning, I run it independently and attach the pitures of results.

library(dplyr)  
library(tm)  
library(stringr)  
library(NLP)  
library(openNLP)  
library(caret)  
library(wordcloud)  
library(RColorBrewer)  
knitr::opts\_chunk$set(echo = TRUE, warning=FALSE, message=FALSE, include=FALSE)  
wine <- read.csv("C:/Users/Duong Minh Duc/Documents/GitHub/Text-Mining-Project/wine.csv")  
head(wine)  
boxplot(wine$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$designation, wine$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)  
id = sample(1:n, floor(n\*0.8))  
train\_full = wine[id,]  
test\_full = wine[-id,]  
  
id2 = sample(1:n, floor(n\*0.2))  
wine\_sample <- wine[id2,]  
wine2 <- wine[id2,]  
n2 = length(id2)  
id\_test = sample(1:n2, floor(n2\*0.8))  
train = wine\_sample[id\_test,]  
test = wine\_sample[-id\_test,]  
  
##clean function  
clean <- function(text\_vector)  
 {  
 wine\_corpus = VCorpus(VectorSource(text\_vector))  
 wine\_corpus = tm\_map(wine\_corpus, removePunctuation)  
 wine\_corpus = tm\_map(wine\_corpus, content\_transformer(tolower))  
 wine\_corpus = tm\_map(wine\_corpus, removeNumbers)  
 wine\_corpus = tm\_map(wine\_corpus, removeWords, stopwords )  
 #wine\_corpus = tm\_map(wine\_corpus, stripWhitespace)  
 wine\_corpus <- tm\_map(wine\_corpus, stemDocument)  
   
   
 return(wine\_corpus)  
 }  
  
##create the train set  
wine\_train\_set <- clean(train$description)  
  
train$description[1:5]  
wine\_train\_set[1:5]  
  
train\_dtm\_tfidf <- DocumentTermMatrix(wine\_train\_set, control = list(weighting = 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 attach 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 = FALSE)  
}  
  
wine\_train\_set <- clean(train$description)  
  
train\_dtm\_tfidf <- DocumentTermMatrix(wine\_train\_set, control = list(weighting = 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 = FALSE)  
}  
  
wine\_train\_set <- clean(train$description)  
  
train\_dtm\_tfidf <- DocumentTermMatrix(wine\_train\_set, control = list(weighting = 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)  
#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(weighting = 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  
adj  
  
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)  
#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)  
 {  
 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)  
wine\_train\_set\_full <- clean(train\_full$description)  
  
train\_dtm\_tfidf\_full <- DocumentTermMatrix(wine\_train\_set\_full, control = list(weighting = weightTfIdf))  
train\_dtm\_tfidf\_full <- removeSparseTerms(train\_dtm\_tfidf\_full, 0.99)  
  
#create the test set  
wine\_test\_set\_full <- clean(test\_full$description)  
wine\_test\_set\_full <- DocumentTermMatrix(wine\_test\_set\_full, control = list(dictionary = Terms(train\_dtm\_tfidf\_full) ,weighting = weightTfIdf))  
#wine\_test\_set <- DocumentTermMatrix(wine\_test\_set, control = list(dictionary = Terms(train\_dtm\_tfidf) , tokenize=NLP\_tokenizer))  
  
#create matrix for training  
wine\_train\_set\_full <<- as.matrix(train\_dtm\_tfidf\_full)  
wine\_test\_set\_full <- as.matrix(wine\_test\_set\_full)  
wine\_test\_set\_full <- wine\_test\_set\_full[,Terms(train\_dtm\_tfidf\_full)]  
#create the test result  
wine\_testing\_result\_full <- test\_full$quality  
  
tag\_full <- tagPOS(Terms(train\_dtm\_tfidf\_full))  
tag\_full <- tag\_full$POStags  
adj\_id\_full <- which( tag\_full=="JJ")  
adj\_full <- colnames(wine\_train\_set\_full)[adj\_id\_full]  
  
column\_id\_full <- c()  
  
#multify for noun  
for (i in 1:dim(wine\_train\_set\_full)[2]) {  
 check <- colnames(wine\_train\_set\_full)[i] %in% adj\_full  
 if(check)  
 {  
 column\_id\_full <- c(column\_id\_full, i)  
 }  
}  
  
wine\_train\_set\_full[,column\_id\_full] <- wine\_train\_set\_full[,column\_id\_full]\*2  
wine\_test\_set\_full[,column\_id\_full] <- wine\_test\_set\_full[,column\_id\_full]\*2  
  
  
train\_svmLinear\_model\_final1 <- train(x= wine\_train\_set\_full, y=train\_full$quality , method = 'svmLinear3')  
  
model\_svmLinear\_result\_final1 <- predict(train\_svmLinear\_model\_final1, newdata = wine\_test\_set\_full)  
  
  
conf\_svmLinear\_train\_final1 <- table(model\_svmLinear\_result\_final1, wine\_testing\_result\_full)  
names(dimnames(conf\_svmLinear\_train\_final1)) <- c("Predicted class", "Actual class")  
  
train\_svmLinear\_model\_final1  
confusionMatrix(conf\_svmLinear\_train\_final1)  
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 = " Boxplot of log of the price")  
boxplot(wine$value, main = "Value of the price")  
  
good<- which(train\_full$quality=="good")  
good <- wine\_train\_set\_full[good,]  
good = data.frame(sort(colSums(good), decreasing=TRUE))  
excellent<- which(train\_full$quality=="excellent")  
excellent <- wine\_train\_set\_full[excellent,]  
excellent = data.frame(sort(colSums(excellent), decreasing=TRUE))  
  
wordcloud(rownames(excellent), excellent[,1], max.words=100, colors=brewer.pal(8, "Dark2"))  
  
wordcloud(rownames(good), good[,1], max.words=100, colors=brewer.pal(8, "Dark2"), main= "Top tfidf good class")  
  
#remove double points for adj  
wine\_train\_set\_full[,column\_id\_full] <- (wine\_train\_set\_full[,column\_id\_full]/2)  
  
  
high<- which(train\_full$benefit=="high")  
high <- wine\_train\_set\_full[high,]  
medium<- which(train\_full$benefit=="medium")  
medium <- wine\_train\_set\_full[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, "Dark2"))  
  
wordcloud(rownames(medium), medium[,1], max.words=100, colors=brewer.pal(8, "Dark2"))