*04/24/2019*

**BYGB/ISGB 7977: TEXT ANALYTICS**

**PROJECT FINAL REPORT**

# Fordham Sommelier: Analyzing Wine Reviews Using Text Analytics and Machine Learning

GROUP 10: FORDHAM SOMMELIER

*Charles Owens, Ila Srivastava, Sravya Katta, Katie Linh Cao, Yi Qian*

1. **Executive Summary**

Our primary project goal is to derive insights from WineMag dataset that is a list of 129,971 reviews of different wines by professional wine tasters at Wine Enthusiast, one of the preeminent wine magazines. We would like to identify the wine that receives the highest sentiment ranking, their location and characteristics just like a professional sommelier would through a tasting.

Focusing mostly on the textual description of these reviews, we intend to apply various text analytics techniques such as sentiment analysis, text classification, tokenization, stemming, stopword filtering and building a machine learning model to predict the wine variety based on the wine reviews. The tools used for our analysis are Python with its extensive libraries such as Scikit-learn, Pandas, Vader, NLTK; SPSS Modeler Text Analytics for building different machine learning models; and Tableau for visualization purpose.

Our dataset includes variables such as wine description, price, province, region, country, title, variety, etc. We have included tables and charts showcasing our analysis with respect to our research hypotheses along with our country-specific analysis. Tableau charts indicate correlation between different variables such as sentiment score, points, and sentiment rank. We have done detailed analysis regarding textual data to satisfy our problem statements, developed a strong neural network model to test our pre-analysis conclusions and hypotheses.

1. **Problem Statement**

*Background*

Wine is an alcoholic beverage made with the fermented juice of grapes. Variants of wine include red wine, white wine, rosé wine, fruit wine and other wine-based products. Technically, wine can be made with any fruit (i.e. apples, cranberries, plums, etc.) but most wines are made with wine grapes.

Wine tasting is the sensory observation and appraisal of wine. Nowadays sommeliers and wine buyers often use specialized terminology which is used to describe the flavors, aroma, and other characteristics of a wine. That said, a taster’s judgment can be influenced knowing details of a wine, creating what’s often being categorized into price bias, color bias, and geographic origin bias. The general assumption has always been wine produced in European countries like France is better than other country, or expensive wine tastes better than inexpensive ones. Another assumption is white wine is supposed to be refreshing while red wine more poignant. Many scientific researches have confirmed these biases through blind tasting experiments.

We want to perform exploratory analysis on a relatively large dataset on wine reviews of various types of wine from different countries in order to observe any interesting finding and test several hypotheses that we have regarding wine appreciation. Our analysis will help us determine the main criteria through which a quality in wine is determined by testing various features such as price, rating, review, region, etc. The value added from this research includes answering questions such as whether expensive wines really deserve their price tag or whether famous wine producing countries such France, Italy match their reputation.

When performing text analytics on wine reviews, it is important to note that wine grape varieties are variously evaluated according to a wide range of descriptors which draw comparisons with other, non-grape flavors and aromas. For example, common sensory descriptors for a signature red grape variety- Cabernet Sauvignon are eucalyptus, chocolate, tobacco; while descriptors for another white grape variety- Prosecco are apple, honey, citrus. Basic knowledge of wine appreciation enhances our understanding of text analytics result hence the effectiveness of our analysis for this project.

*Project proposal*

The proposal for our analysis has been detailed as follows:

* Our primary goal is to rank the wines based on sentiment derived from reviews and the points given to each wine by wine enthusiasts.
* Secondly, based on that sentiment rank, we would try to identify the top three countries from which the most popular wines are produced and perform text analysis specifically on each of those top countries as well as popular regions and provinces.
* We will also identify the most popular wine variety based on the reviews given to each wine.
* Finally, we would like to build a machine learning model that can predict and identify the variety of a wine based on review description.

*Hypotheses*

Looking at the dataset, we intend to test the following hypotheses:

* Wine price is positively correlated with the higher rank given to review.
* Popular wine producing countries (such as France, Italy) are likely to produce higher ranked wines.
* Wines with the perfect score i.e., 100 also have the best sentiment score.
* Better ranked wines also have a positive correlation with longer description.

1. **Data Description**

Our dataset is a 50.4 MB csv file containing wine reviews scraped from WineEnthusiast that our team downloaded from Kaggle website. The dataset contains reviews for 129,970 different kinds of wine from all around the world.

The variables are described in the table below.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| country | Country where the wine is from | string |
| description | Review of the qualities of the wine | string |
| designation | Vineyard within the winery where the grapes that made the wine are from | string |
| points | Number of points Wine Enthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80) | numeric |
| price | Cost for a bottle of the wine | numeric |
| province | Province or state that the wine is from | string |
| region\_1 | Wine growing area in a province or state | string |
| region\_2 | Sometimes there are more specific regions specified within a wine growing area | string |
| taster\_name | Name of wine taster | string |
| taster\_twitter\_handle | Twitter handle of wine taster | string |
| title | Title of the wine review, which often contains the vintage | string |
| variety | Type of grapes used to make the wine | string |
| winery | Winery that made the wine | string |

1. **Methodology**

*Data Collection/Preprocessing*

Our dataset is a list of 129,971 reviews of different wines by professional wine tasters at Wine Enthusiast, one of the preeminent wine magazines. The data was scraped from winemag.com during the week of June 15th, 2017. As the data was relatively clean, no preprocessing was required.

**Text Analysis Using Python**

Figure 1. Diagram Showing Overall Methodology Using Python

*Sentiment analysis using Vader*

VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative. VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is. In this study, we used VADER in python to analyze the sentiments based on reviews by taster. Based on the value of the positive sentiment the wines have been ranked in descending order. The better the positive score, the better the wine.

We created numerous functions in python using NLTK which is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to [over 50 corpora and lexical resources](http://nltk.org/nltk_data/) such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. We created python functions for performing tokenization, stopword removal, stemming, POS tagging, n-grams to understand our data better for the most popular wine countries based on our rank generated in python. Our study focuses on France, Portugal and US for further analysis and in order to do that we compiled the description for each of these countries for rank less than 50.

*Tokenization*

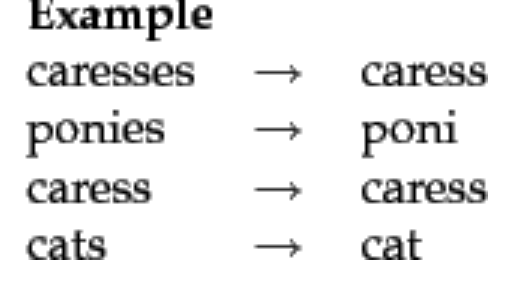
Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Before a program can process natural language, we need to identify the words that constitute a string of characters. This is important because the meaning of text generally depends on the relations of words in that text.

*Stopword removal*

Stopwords are words that from non-linguistic view do not carry information or a set of commonly used words in any language, not just English. The reason why stopwords are critical to many applications is that, if we remove the words that are very commonly used in a given language, we can instead focus on the important words instead. For example :A, The, And etc. We have imported stopwords from NLTK library and remove stopwords from our reviews.

*Stemming*

Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. The most common algorithm for stemming English, and one that has repeatedly been shown to be empirically very effective, is Porter's stemmer. Porter's algorithm consists of 5 phases of word reductions, applied sequentially.



*POS tagging*

The primary target of Part-of-Speech (POS) tagging is to identify the grammatical group of a given word. Whether it is a NOUN, PRONOUN, ADJECTIVE, VERB, ADVERBS, etc. based on the context. POS Tagging looks for relationships within the sentence and assigns a corresponding tag to the word. We have retained words that have tag starting with ‘N’ as we consider them to be of more relevance.

*n-grams*

N-grams of texts are extensively used in text mining and natural language processing tasks. They are basically a set of co-occurring words within a given window and when computing the n-grams you typically move one word forward (although you can move X words forward in more advanced scenarios). It is known that Bigrams are the most informative N-Gram combinations. Adding bigrams to feature set will improve the accuracy of text classification model.

**Analysis Using Machine Learning Model with SPSS Modeler Text Analytics**

Figure 2. Diagram Showing Overall Methodology Using SPSS Modeler Text Analytics

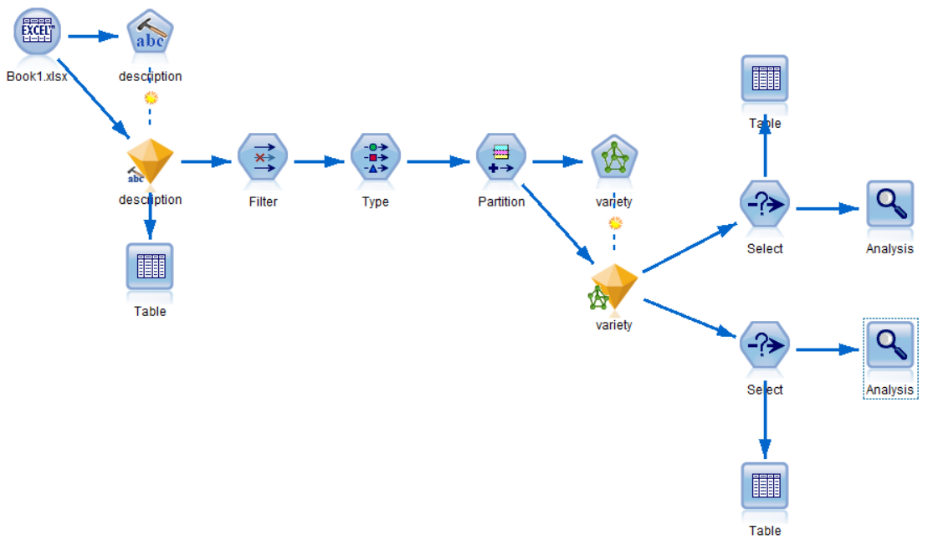


Figure 3. SPSS Text Analytics Model Stream

*Data preprocessing*

In order to run the model successfully, we selected records of the top 10 wine varieties as a new dataset to train the model. These top 10 wine varieties are defined as ones with the most records, each beyond 1000 records in our dataset. This by no means indicates that these varieties are the most well-received, but rather they were chosen because of their prevalence and to make the model work. We wanted to analyze customers’ wine review, so we chose description column as text field. Also, we unchecked concepts that occur in too many records, percentage of which was beyond 95.

*Variable selection*

After we selected the typical words in description, we used filter node to remove variables except series of concepts. Because here we wanted to predict wine varieties just according to wine review. We also kept C1 column as ID field and variety column as target.

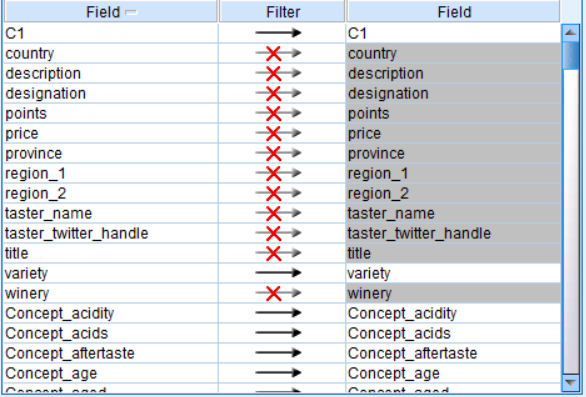


Figure 4. Feature Selection in SPSS

*Model building – Artificial Neural Network*

We separated the dataset into 80% training and 20% testing data. After trying different machine learning models, namely Artificial Neural Network (ANN), Random Forest, Decision Tree, we decided to go with ANN for our prediction because of its high accuracy on both training and testing data. The model ran successfully with an accuracy of 72.6%.

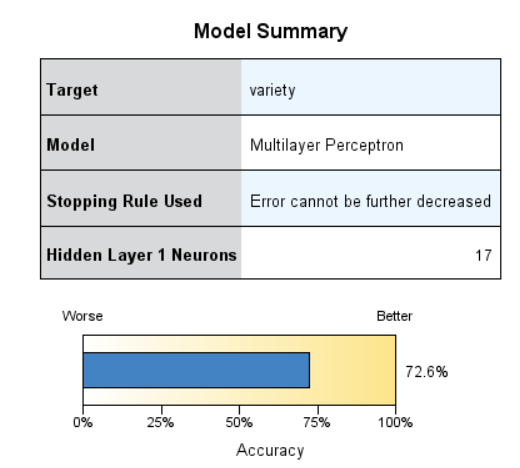


Figure 5. ANN Model Summary

1. **Results & Discussion**

Below is a table containing the text analysis of the three most popular wine producing countries-France, Portugal, and the United States based on the wine varieties that have the highest sentiment ranking that we have identified using Vader Sentiment Analysis.

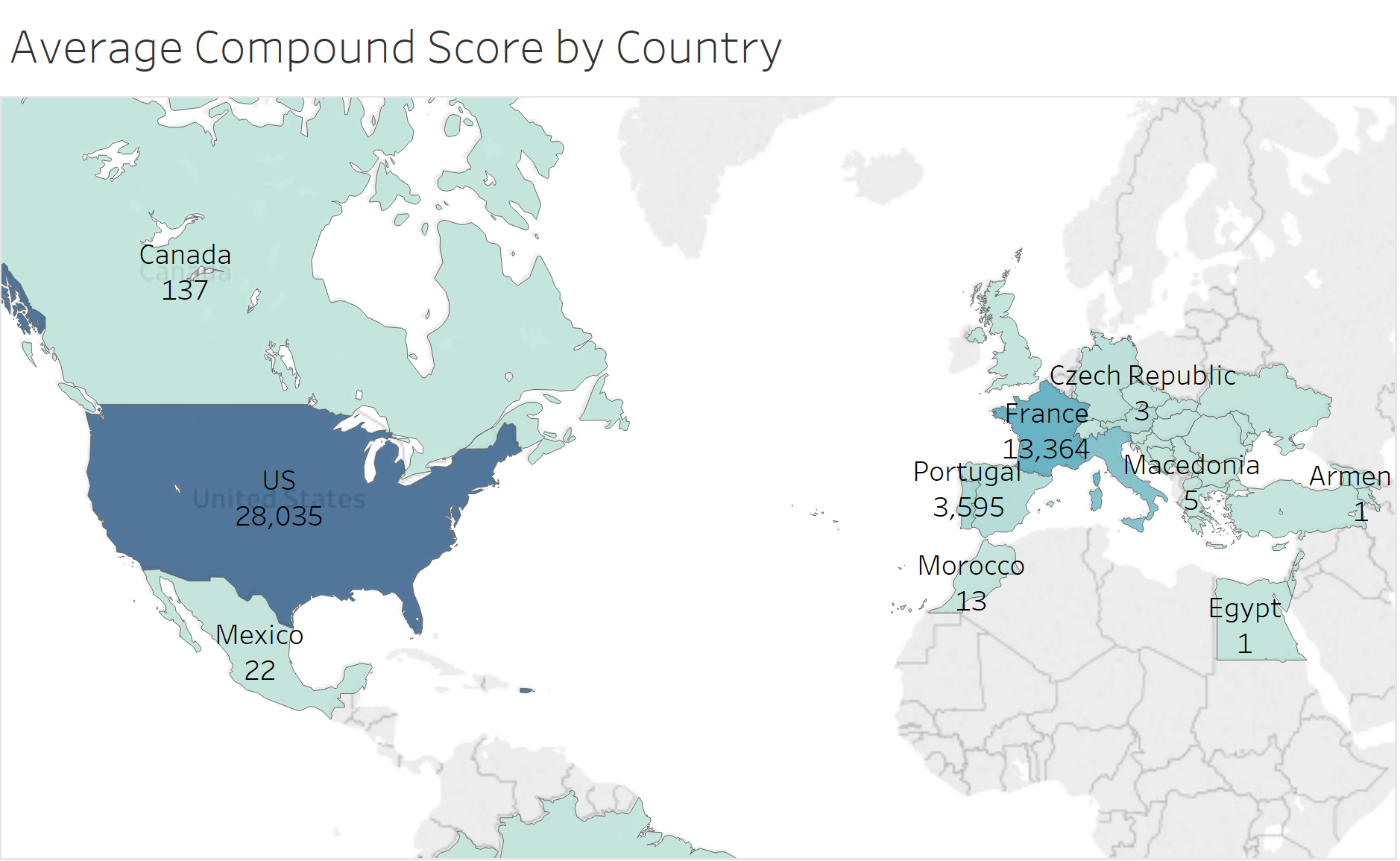


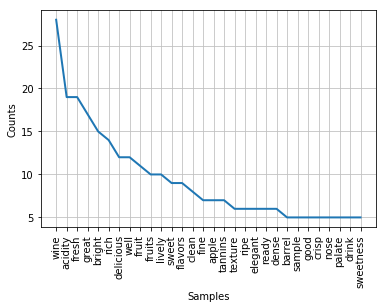
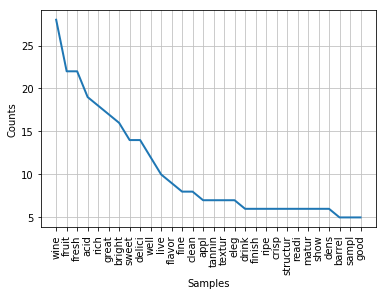
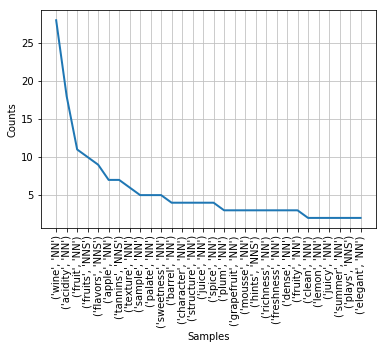
Figure 6.0. Average Vader Compound Score by Countries

This shows that the U.S., France, and Portugal generally have wine varieties with high sentiment ranking.

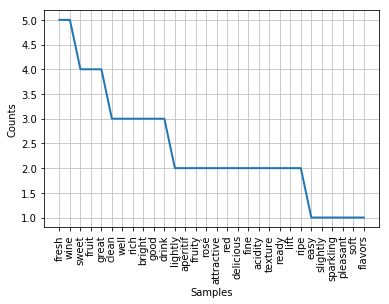
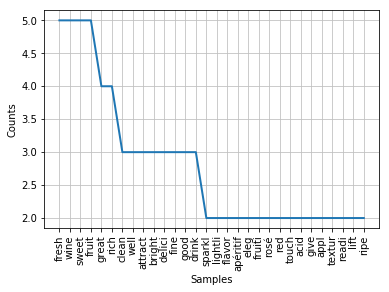
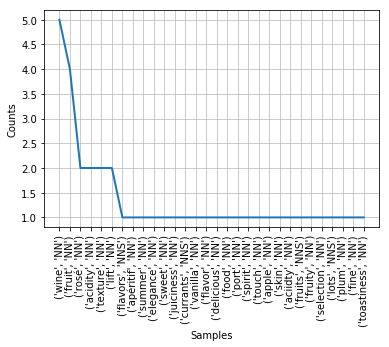
|  |  |  |  |
| --- | --- | --- | --- |
|  | **France** | **Portugal** | **United States** |
| **Tokenization** | Words like and, a, with, wine have higher frequencies. | Words like a, with, and, is have higher frequencies. | Words like flavors, and, a, clean have higher frequencies. |
| **Stopword removal** | After stopword reduction we can see France wines have characteristics like fresh, acidity, bright, rich, lively, sweet. | After stopword reduction we can see Portugal wines have characteristics like sweet, rich, lightly, etc. | After stopword reduction we can see US wines have characteristics like pleasant, rich, aromatic etc. |
| **Stemming** | Helps in retaining root word like acid, sweet, rich, fine | Helps in retaining root word like light, rose. Rosé wine is preferred. | Helps in retaining root word like soft, clean, aroma, pleasant, spice etc. |
| **POS tagging** | Apple, plum, grapefruit, strawberry are among noticeable words. | Rose and vanilla are famous ones. Texture of wines is good. | Melon, blackberry, orange, grapefruit appear to be top descriptors. |
| **n-gram** | The most common bi-gram is “fine” and clean”. Other notable common bi-grams include “intense” and “sparkling”, “apple” and “barrel”, “sweet” and “fruits”, “delicious” and “sweet”. | The most common bi-gram is “easy” and “fresh”. Other notable common bi-grams are “slightly” and “sparkling”  “pleasant” and “soft”,  “lightly” and “sweet”. | The most common bi-gram is “honey” and “orange”. Other notable common bi-grams include “vanilla” and “flavors”, “rich” and “balanced” and “elegant” and “clean”. |

*3 Country-specific analyses: France, Portugal, United States*

France Analysis

**  **

Portugal Analysis

United States Analysis

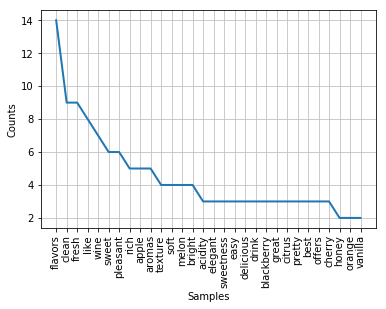
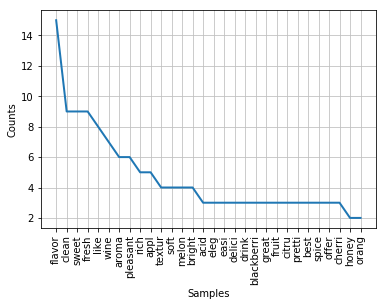
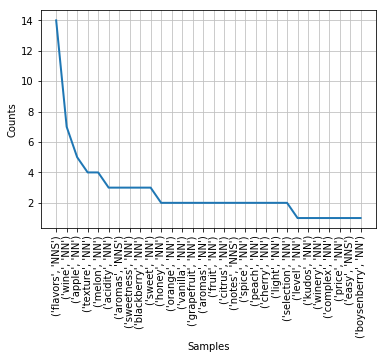
  

Figure 6. From left to right: Subplots showing word frequency of Tokenization after Stopword removal, Stemming, POS tagging

**Analysis Based on Visualization**

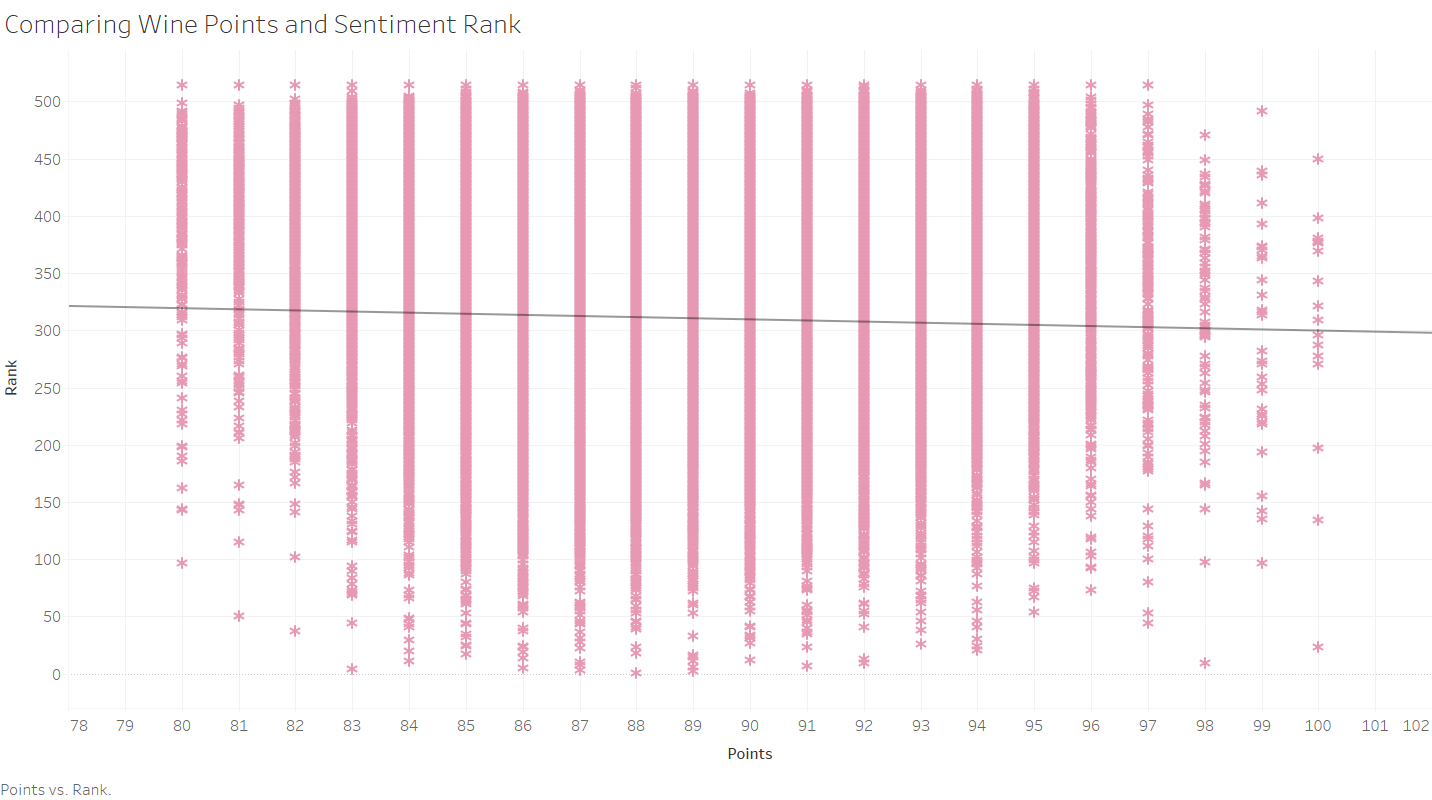
****

Figure 7. Scatterplot Comparing Wine Points and its Corresponding Sentiment Ranking

There does not appear a correlation between the sentiment rank and points given for each wine. Using Tableau and assessing all red wine that has 100 points, we found that they are all described as black or dark.

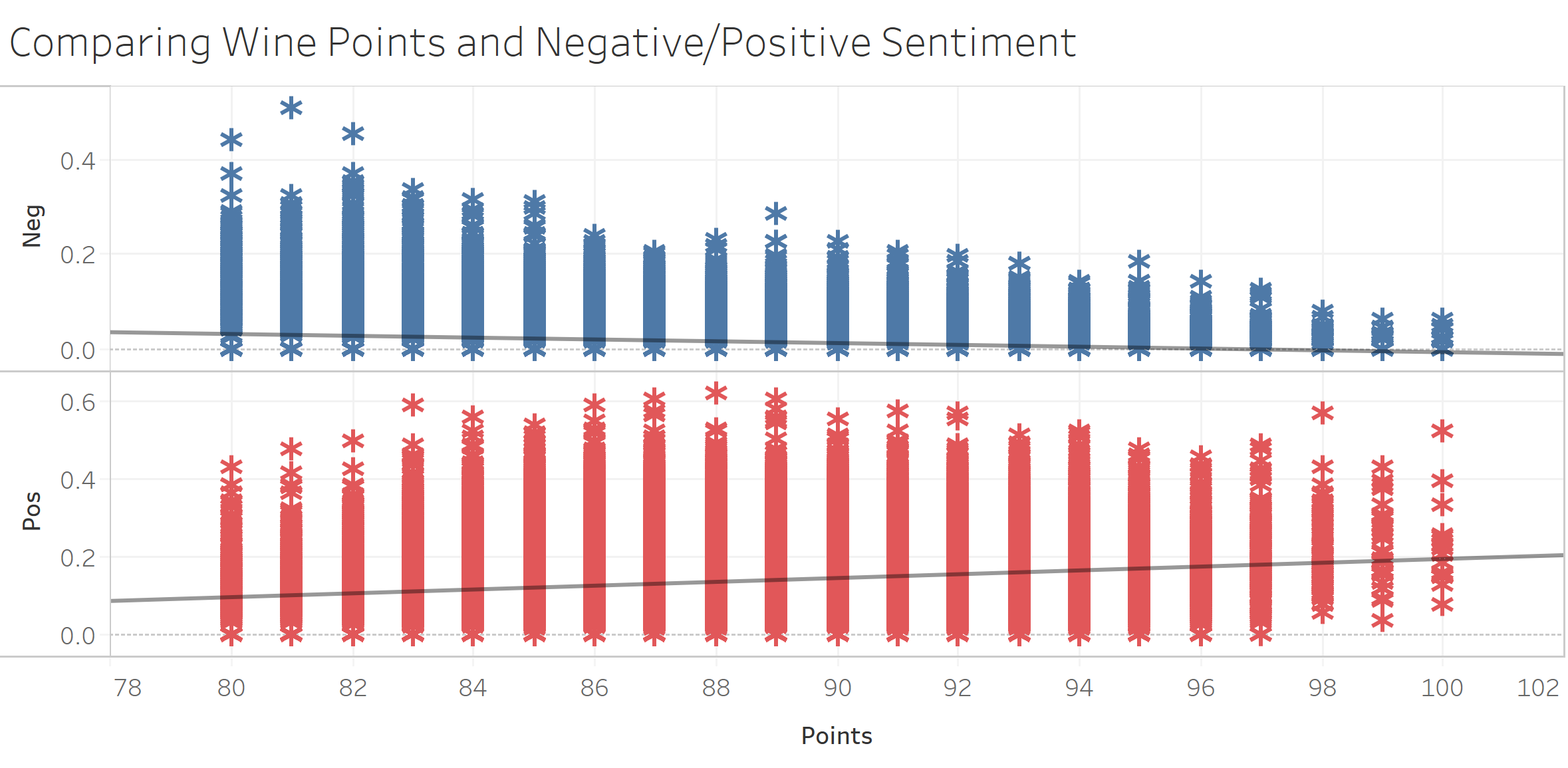


Figure 8. Scatterplot Comparing Wine Points and its Negative/Positive Sentiment

There appears a strong negative correlation between wine points and negative sentiment score given to each wine. However, the correlation for positive sentiment and wine points is not statistically significant.

**Analysis of Machine Learning Model**

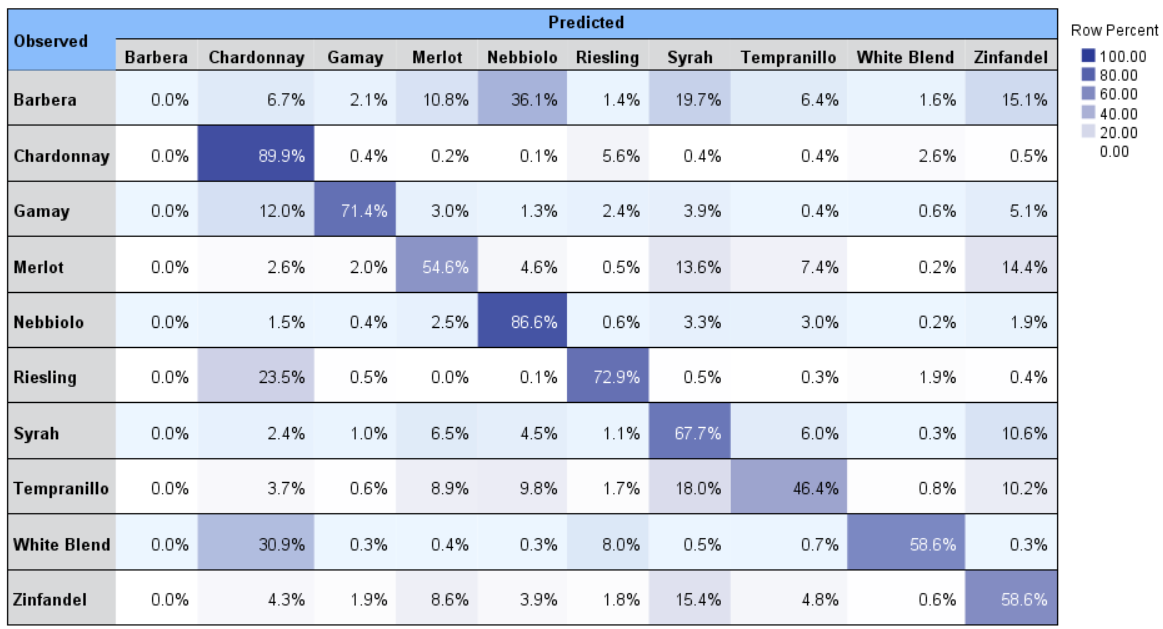


Figure 9. Prediction Accuracy of Each Wine Variety

We can see that Chardonnay is the highest prediction accuracy of 89.9% and Nebbiolo comes second highest with an accuracy of 86.6%. The third one is Riesling with that of 72.9%.

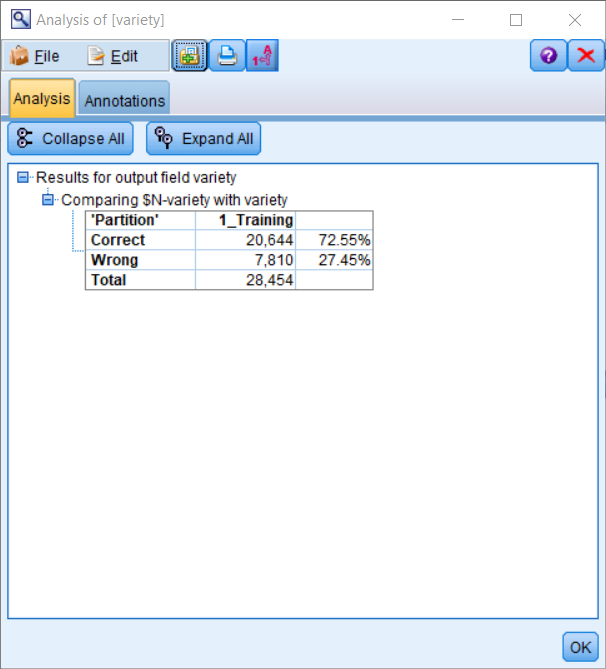
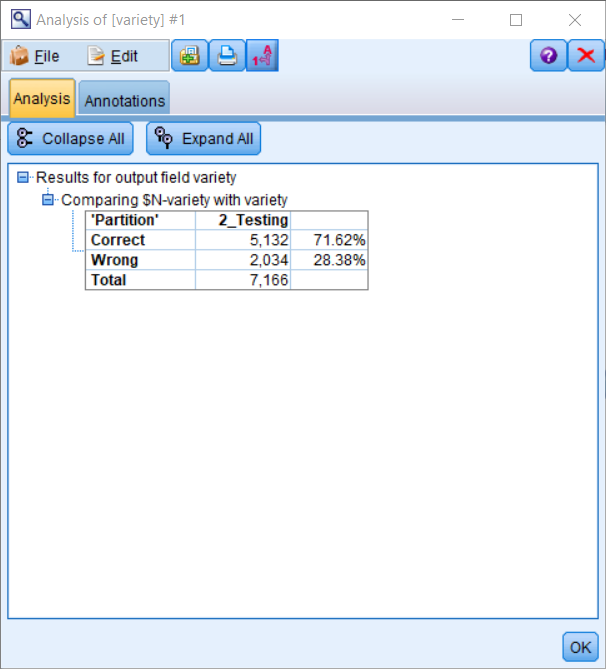
 

Figure 10. Accuracy of Training and Testing Data

We put an Analysis node to see the results of the ANN model. Accuracy of training data is 72.55%, which is relatively good. Accuracy of testing data is 71.62%, which is also relatively good. Because the accuracy of the training data and testing data are relatively the same, we can conclude the model performs well overall.

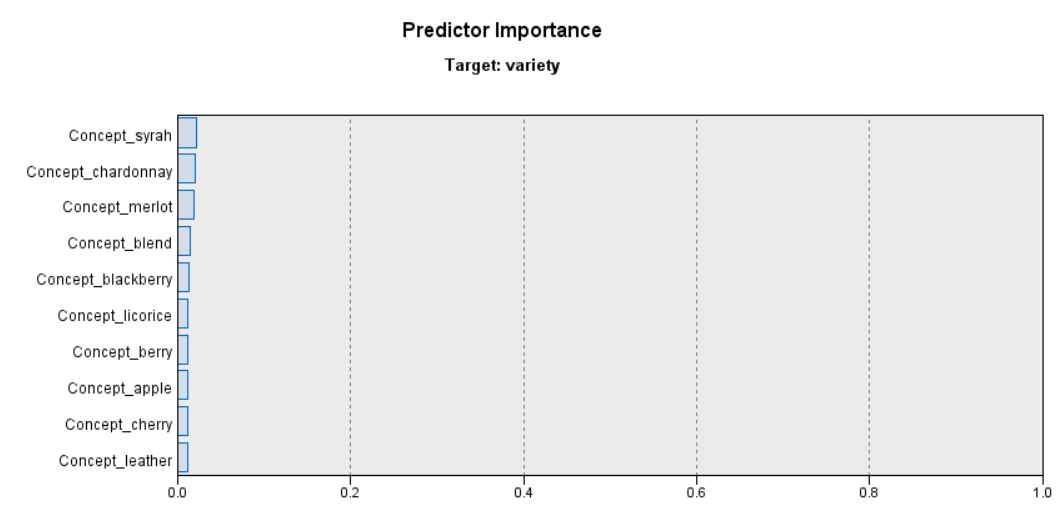


Figure 11. Predictor Importance for Wine Variety

Assessing the predictor importance, if the wine review contains the name of wine variety (for example, Syrah, Chardonnay, Merlot), it would increase the accuracy of the prediction. This should not come as a surprise. The next important factor is common sensory descriptors (what wine tasters usually compare the aroma and taste of the wine variety to a non-grape flavor). As mentioned earlier in the background section of our project, if the reviews mention common sensory descriptors such as fruity descriptors- blackberry, apple, cherry, or material aroma like leather, it will increase the accuracy of the prediction.

1. **Conclusions**

Regarding our initial hypotheses and the performance of our machine learning model:

* Our hypothesis regarding price and wine rank is not confirmed as most of the top ranked wines are not considered expensive.
* There appears to be a strong negative correlation between wine points and negative sentiment score given to each wine indicating that wines with high negative sentiment scores are also given low points.
* Significant correlation was not found between positive sentiment score/sentiment rank and wine points.
* Neural network model has given good accuracy with regard to the prediction of wine variety with an accuracy of 71.62%
* Important predictors for wine variety are common sensory descriptors like blackberry, apple and cherry.

Interesting findings about wine based on our text analysis of all the reviews:

* The top 10 wine varieties are White Blend, Riesling, Chardonnay, Merlot, Gamay, Zinfandel, Syrah, Nebbiolo and Barbera.
* France, Portugal, and the United States frequently produce the top ranked wines. Our hypothesis related to France producing top wines has been satisfied, yet Italy does not rank high.
* Some of the most popular wines are Domaine Huët 2005 Pétillant (Vouvray), Château Haut-Simard 2009 Barrel sample (Saint), Straight Line 2011 Sauvignon Blanc.
* Provinces that produce highly-preceived and highly-ranked wines are Loire Valley, Bordeaux, California and Vinho Verde, to name a few.
* French wines are usually described as fresh, acidity, bright, rich, lively, sweet whereas American-produced wines as pleasant, rich, and aromatic.

1. **Future Recommendations & Further Exploration**

This project provides great insight and added-value to wine tasting- a popular topic that has long been considered rather subjective and not backed by scientific research yet embedded in the culture and history all over the world for thousands of years. By performing multiple text analytics techniques, we were able to point out what contributes to a popular wine variety and what drives the rating and the sentiment of wine tasters based on hundred thousand of reviews. A professional sommelier could look at the result of this analysis to further understand what other sommeliers think about certain wine varieties and the aromas of them. This could help further fine-tune the art of wine and food pairings using large database. A wine buyer would look at this analysis and be able to navigate through various wine varieties and pick something that suits her taste the best without having to try it. It is also interesting to point out that price does not dictate the outcome of the wine points and its sentiment.

Availability of data from other magazines beside WineMag and professional sommelier networks would improve the analysis and the prediction of any model used in the future. Further deep-dived sentiment analysis into the different vintages of one variety (this is called vertical wine tasting) and same vintage of different wine varieties (horizontal wine tasting) would also be interesting.