Data Science of Wine Survey Paper

Chiu-Yuan Wu
M.S. of Computer Science
Washington State University
Pullman, WA

Abstract—**Analyzing partial dependence plots on chemical attributes of wines and identify the attributes that affects wine quality. Build a recommending system by doing word embedding on wine reviews then apply nearest neighbors to the vectors.**

Keywords—wine; partial dependence; word embedding; nearest neighbors;

I. Introduction

There are over eight thousand types of grapes in the world, out of the eight thousand types, six thousand of them are suitable for winemaking. According to the Food and Agriculture Organization of the United Nations, an average of 36 billion bottles of wine[1] are produced per year around the world. The wine market is quite enormous and a recommending system will be quite nice for people that are new to wine or people that are learning about wine.

In this survey paper, I will explore two approaches of applying data science to wine, and try to find out which approaches might be suitable for building a wine recommending system. The first approach is to use the measured numerics, such as alcohol percentages, pH value, and density of wine. Use those chemical properties as features to train the model and output a wine quality score. We can then predict wine quality with the chemical properties. The second approach is to analyze professional wine reviews and convert the descriptors into quantitive features. Then those features can be used to create a wine recommending system.

II. Recommend with Chemical Properties

A. Wine Components

The dataset Wine Quality[2] from UCI Machine Learning Repository is used by most of the sources I found. The data contains 1599 wine samples, there are eleven measured variables and an output variable of wine quality. The input variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The output wine quality is a score between 0 and 10.

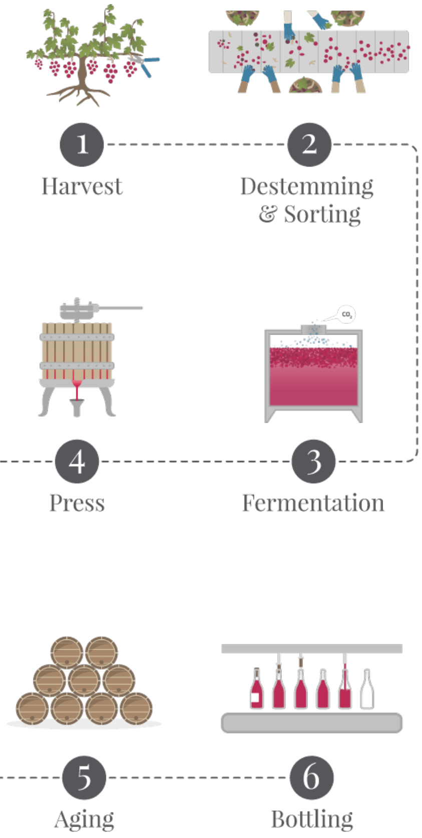
The most important components that wine is made up of are alcohol, acids, sugar, sulphur compounds, and chlorides. Over 80% of wine sell on the market contains 11%-15% of alcohols, but it is not restricted to lie within the range, some are higher and some are lower. Acidity makes the wine feel more interesting, wines that lack acidity tastes flat and considered one-dimensional, or boring. Two types of acidity might be preserved in the wine during the winemaking processes, fixed acidity and volatile acidity. Fixed acids originate from the grapes and they do not easily evaporate. These acids are the part that make wines differ from each other. The second is called volatile acidity, these acids come from the yeast in the fermentation process. Volatile acids will make the wine taste like vinegar and ruined the origin flavor of the grape. During the fermentation process, yeast produces alcohol by consuming sugar. Residual sugar is the amount of sugar that is left after fermentation. Residual sugar is the component that makes wine tastes sweet, and the opposite of sweet in wine language is called dry. Most of the wine in the market are considered as dry wine.

Fermentation Process:

$$Sugar + Yeast \rightarrow Alcohol + Carbon\ Dioxide$$

The next component of wine is sulfur dioxide. It is a compound that is used to prevent the wine from infiltration and oxidation. Sulfur dioxide can also be divided into two types, free sulfur dioxide and fixed sulfur dioxide. Usually there is only 35% to 40% of the sulfur dioxide is in free form when added to wine. When the percentage of free sulfur dioxide is kept high, the better the wine can be preserved. Fixed sulfur dioxide is created when free sulfur dioxide binds to microorganisms in the wine. And it might be the sign of the wine being oxidized or microbial spoiled. The last main component is chlorides. The amount of chloride varies among wine production location and climate. For example, vineyards that are closer to the sea will produce wine with higher percentage of chlorides than wine produced from vineyards in land.

The following figure shows the process of wine making[3].



Wine Making Process

B. Partial Dependence

A way to see how each properties effects wine quality is to use partial dependence plot, which uses the average value as input except for the one feature we want to focus on. Let's assume the training model as a function *f*, inputs are the features *x* and output a prediction wine quality $f(x_S)$. To focus on one feature x_S, we marginalize the model over the other features x_C's distribution, which is $P(x_C)$. The following is the partial dependence function on a feature.

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C)$$

We can do approximation on the function with the number of data as follow.

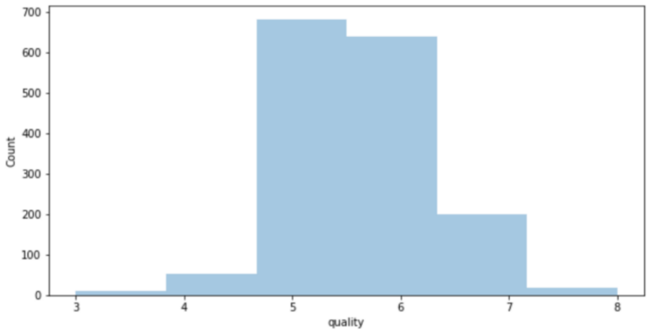
$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$

As an example, if we want to get the partial dependence plot for alcohol percentage on predicting wine quality, the equation looks like this:

$$\hat{f}_{x_S}(x_S)=\frac{1}{n}\sum_{i=1}^{n}\hat{f}(alcoholcontent, otherfeatures^{(i)})$$

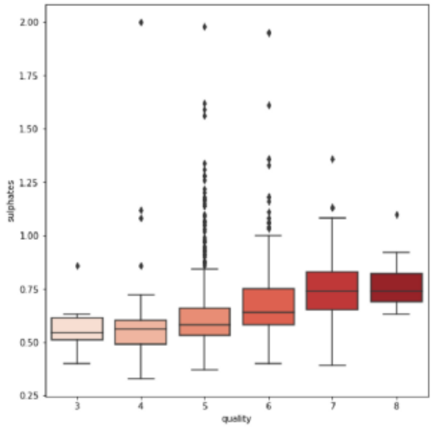
C. Data Exploration [4]

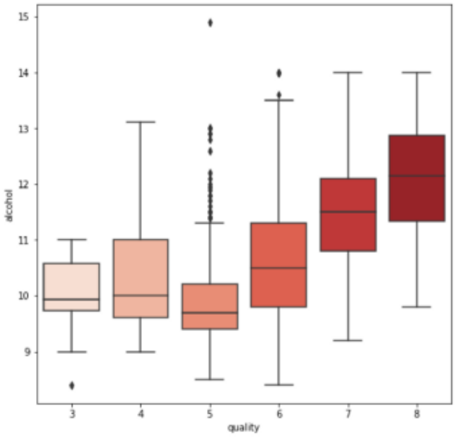
Most of the wine contains in the data fell within mid-tier quality, as the plot shown below.



Count of Wine Quality Score

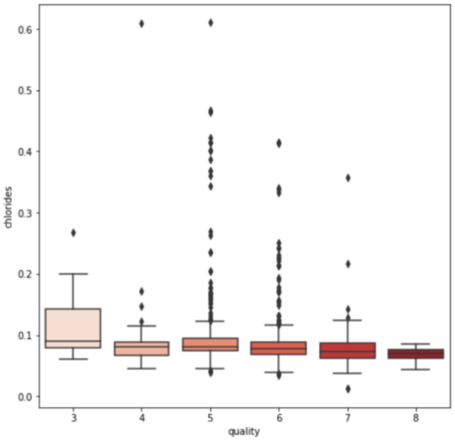
Some features like alcohol and sulphates might be correlated to wine quality, shown in the following figures.

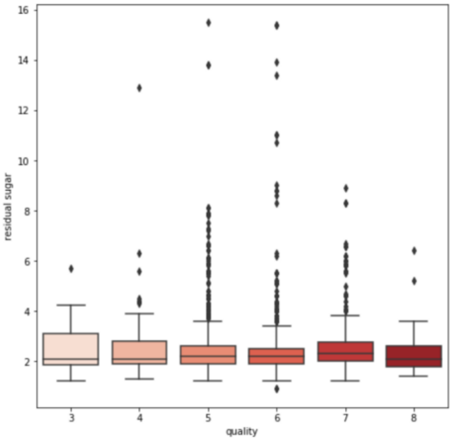




Correlation Plots

While some other features like residual sugars and chlorides seem to have low correlation with quality.





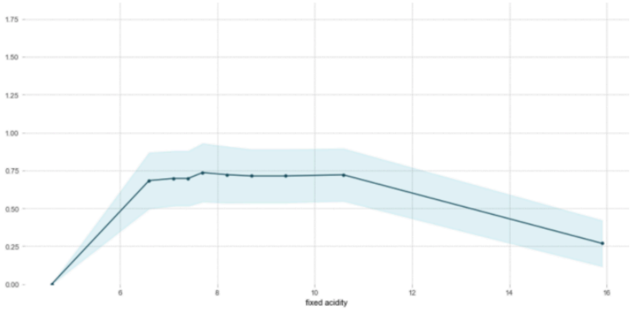
Correlation Plots

D. Model

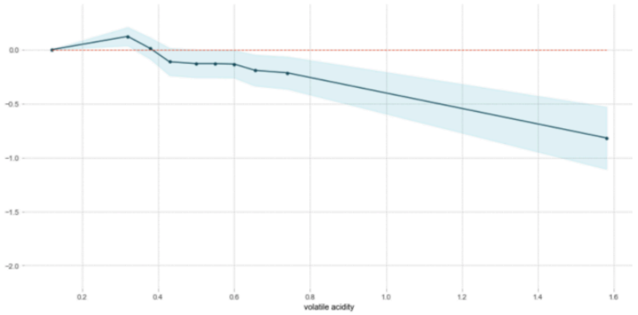
This training model is provided in a post at Towards Data Science by Travis Tang [4]. He feeds the data into a Random Forest Regressor and show the partial dependence plots of each variable.

Take the following two plots as examples, when fixed acidity is around 7 to 10, it will top out its contribution at around 0.75. It means that wine tasters like their wine with higher acidity, and when the acidity is around 7 to 10, the quality increases at a highest rate. When the acidity passes 10, the quality will still increase, but at a slower rate as it increases. This is discussed when introducing components in wine, tasters like acidity comes from the grape fermented, it can make the wine more dimensional, and thus more interesting.

Volatile acidity, on the other hand, has a negative impact on wine quality. The higher it goes, the lower the quality becomes. This is also mentioned in the components paragraph, volatile acidity comes from the fermentation process and it will create a vinegar like taste in the wine, which will ruin the original flavor.

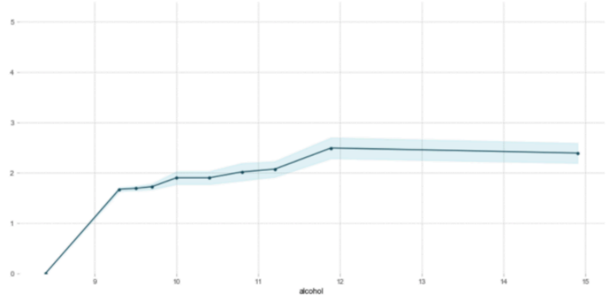


Partial Dependence Plot for Fixed Acidity

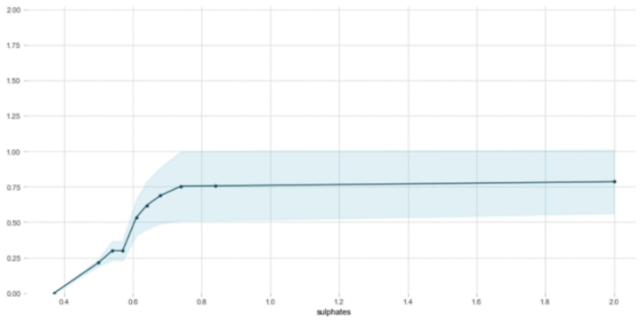


Partial Dependence Plot for Volatile Acidity

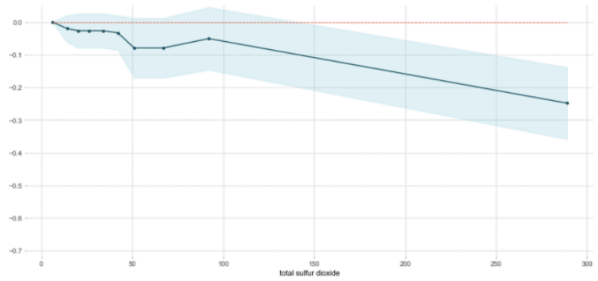
Most of the plots shows a result as we expected, alcohol percentage and sulphate content have a positive effect on wine quality. Sulfur dioxide has a negative impact on wine quality. Residual sugar and chlorides have nearly no impact on wine quality.



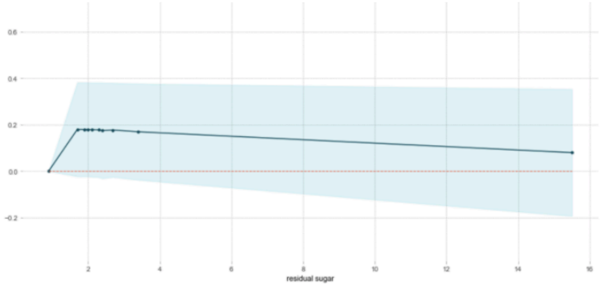
Partial Dependence Plot for Alcohol



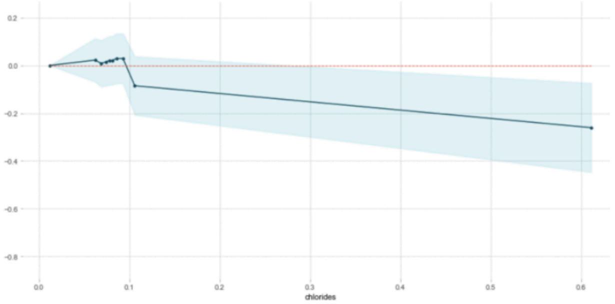
Partial Dependence Plot for Sulphate contents



Partial Dependence Plot for Sulfur Dioxide



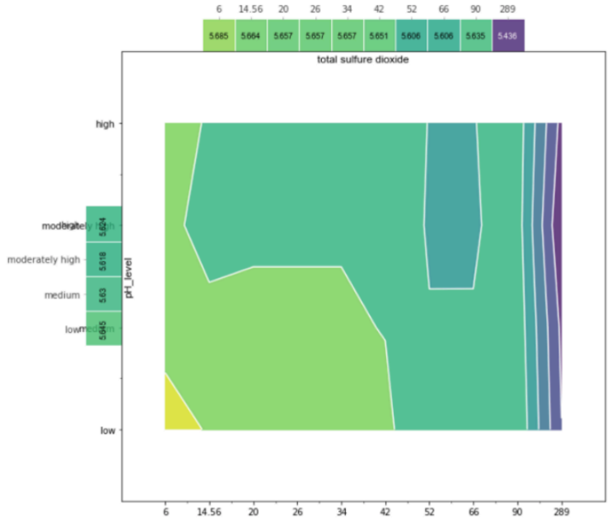
Partial Dependence Plot for Residual Sugar



Partial Dependence Plot for Chlorides

For some one that is new to wine, it might be interesting to see residual sugar has nearly no impact on wine quality, and the contribution becomes even lower when the amount of residual sugar goes higher. It is because sweet wine are usually put in a different category, and people usually do not like wine that is too sweet. For the wine with more sugar in it, it will require even higher acidity in it to make the wine taste balance. For a casual daily drink, dry wines works the best, that's why most of the wine in the market are labeled as dry wine.

In the last part of the post, the author also provided a combined dependence plot of pH level and sulfur dioxide.

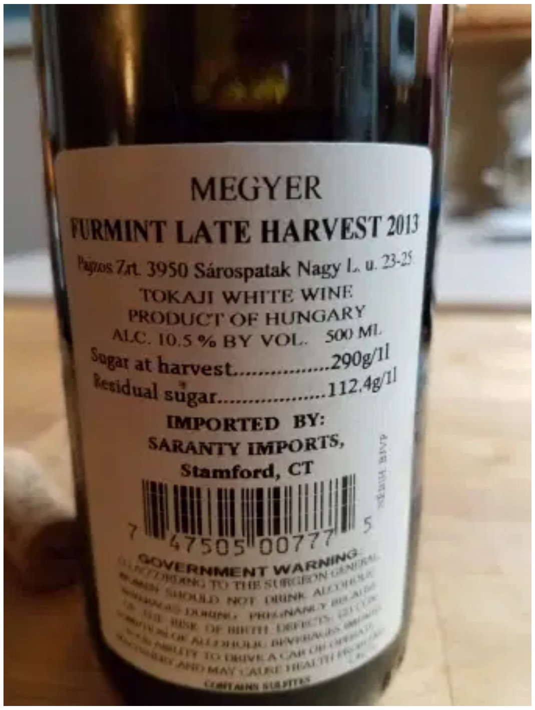


Sulfur Dioxide and pH Combined

The y-axis are 4 pH levels equally divided and the x-axis is sulfur dioxide. The lighter color indicates a prediction of higher wine quality, so the bottom left corner is the best, which means wines with lower pH and lower sulfur dioxide. This matches the previous dependence results. But as the amount of sulfur dioxide goes higher, the pH level seems to become irrelevant. Since the lines between color become vertical, that means there is nothing to do with pH levels.

E. Analysis

By going through the UCI wine quality dataset, we can know which chemical properties usually effect a taster's preference. We may be able to apply this into the recommending system. Perhaps some tasters have different preference on the chemical properties and we could consider this in the system. There is one problem with this approach. That is we don't usually have access to the chemical properties of wines. Alcohol and sugar percentages will be labeled on the bottle but the other ones might required testing to get the data. So this approach might not be easy to put in use.


MEGYER
FURMINT LATE HARVEST 2013
Pajzos Zrt. 3950 Sárospatak Nagy L. u. 23-25
TOKAJI WHITE WINE
PRODUCT OF HUNGARY
ALC. 10.5 % BY VOL. 500 ML
Sugar at harvest.................290g/1l
Residual sugar..................112.4g/1l
IMPORTED BY:
SARANTY IMPORTS,
Stamford, CT
7 47505 00777 5
GOVERNMENT WARNING:
CONTAINS SULFITES

Example Wine Label [5]

III. RECOMMEND WITH WINE REVIEWS

A. Wine Reviews

An experienced wine taster will definitely have some experience dealing with wine reviews or have their own wine tasting notes. A good wine review can let people who forgot whether he or she had tasted that bottle of wine before to recall the taste of the wine if they had tasted it. It can also let new tasters to imagine the flavor of the bottle of wine and decide whether to choose it. The context should be an important role of a wine recommending system.

Ronald Schuring posted a paragraph on Towards Data Science [6] about breaking down professional wine reviews and build a model to train with the reviews. We can eventually build a recommending system with the reviews.

In the post Schuring provided an example wine review for a 2016 Pinot Noir made by Point & Line in Sebastiano Vineyard Reserve [6].

Dried red flowers and sagebrush combine for an elegant aromatic entry to this bottling by two business partners who have worked in Santa Barbara's restaurant scene for many years. Tarragon and intriguing peppercorn flavors decorate the tangy cranberry palate, which is lightly bodied but very well structured.

B. Preprocessing

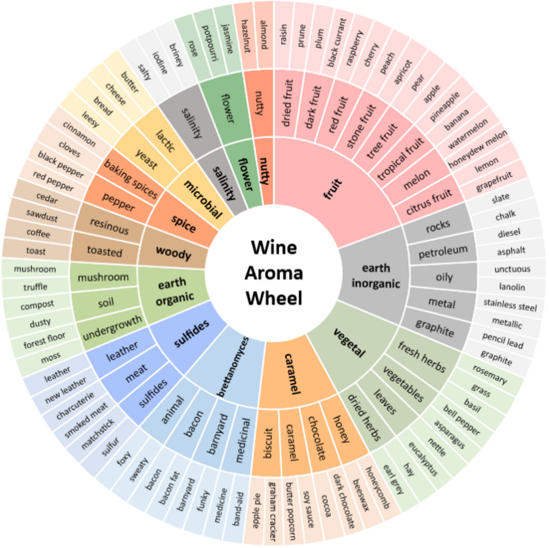
The dataset for this section is from Wine Enthusiast [7], which contains 180,000 wine reviews with wine labels.

There are some preprocessing we have to do with the review before feeding it to a machine learning model. First thing first is to normalize the words, also removing punctuation and stop words. After this process, the review becomes:

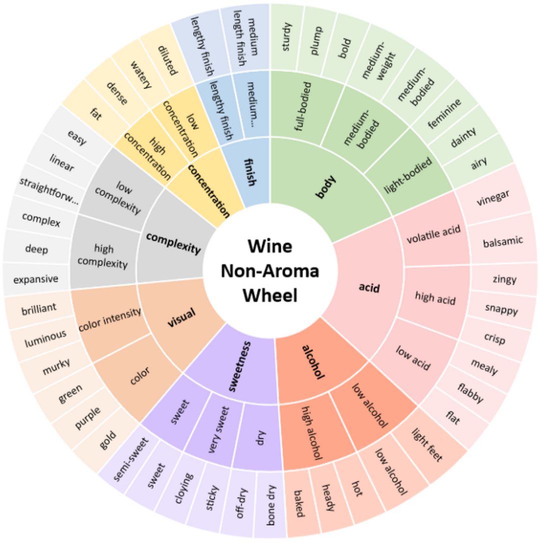
dri red flower sagebrush combin eleg aromat entri bottl two bus partner work santa barbara restaur scene mani year tarragon intrigu peppercorn flavor decor tangi cranberri palat light_bodi veri well structur

The next step is to find out words combinations or phrases that often appears in wine reviews. The author uses the package Phrases to do the task. Terms like *santa_barbara, mani_year,* and *light_bodi* are combined as a term since they usually show up next to each other.

Then comes the two famous wine wheels for wine language. Wine reviewers often create their own language when describing wine aromas, flavors, and structures. Several researchers developed an aroma wheels and a non-aroma wheels to better categorize the word and wine language used to describe wines.



Wine Aroma Wheel [6]



Wine Non-Aroma Wheel [6]

These wheels are the words that used most frequently as descriptors in wine reviews, they are called RoboSomm wheels. The aroma wheels, as the name suggests, are descriptors for aromas. And the non-aroma wheels are words for other characteristics like acid levels, sweetness, and body.

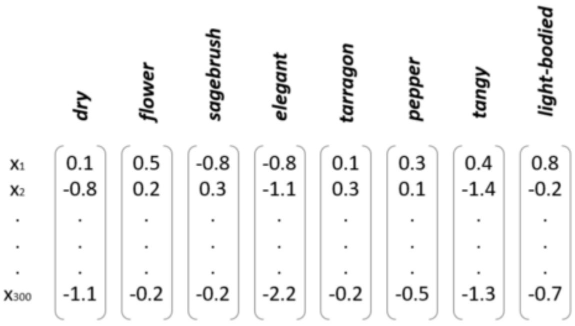
With the wheels, we can have a standardized version of wine review. If we highlight the words that shows up in the wheels for the example review we normalized, it will look like this:

dry *red* ***flower sagebrush*** *combin* ***elegant*** *aromat entri bottl two bus partner work santa_barbara restaur scene mani_year* ***tarragon*** *intrigu* ***pepper*** *flavor decor* ***tangy cranberry*** *palat* ***light_bodied*** *veri well structur*

Now we got the most informative words in the wine review and we can move on to model building part.

C. Model

A trivial approach to quantify the descriptors is to mark them as 0 or 1 representing their absence or presence. But this approach might not take account of the semantic similarities. So the author used the approach of creating word embeddings, which represents words or phrases with vectors. A technique called Word2Vec is used, for every term, it will create a 300 dimensional embedding.

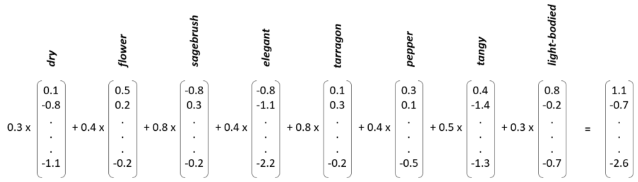


Word Embeddings [6]

With the preprocessing, we already got the words that we care about in the wine review, which are shown in the figure above.

In the next step, TF-IDF is used to weight each words. Term Frequency-Inverse Document Frequency (TF-IDF) will consider both the amount of the relevant words show up in one single paragraph and the amount in the entire document. Take wine reviews as an example, if there are 25 descriptors show up in one review but there are only 10 descriptors in other 5 reviews altogether, the descriptors in the one review should be weighted less than the ones in the other reviews.

Then we multiply the embeddings with the TF-IDF weightings and sum them up to get one vector of embeddings for the wine review.



Embeddings with TF-IDF weighting

D. Building Wine Recommender [6]

By calculating the cosine distance between the embedding vectors of wine reviews, we can build a simple nearest neighbor model to recommend wines. We simply return the wine with the closest vector. The dataset of 180,000 wine reviews are quite large, and the outcomes are surprisingly well with just some simple tricks.

The author used the wine from the wine review example, a 2016 Pinot Noir from Sta. Rita Hills, California, as an input for the recommending system, the input looks like:

```
Wine to match: Point & Line 2016 John
Sebastiano Vineyard Reserve Pinot Noir
(Sta. Rita Hills)
Descriptors: [dry, flower, sagebrush,
elegant, tarragon, pepper, tangy,
cranberry, light_bodied]
```

The top three suggestions are outputted as following :

```
Suggestion 1: Chanin 2014 Bien Nacido
Vineyard Pinot Noir (Santa Maria Valley)
Descriptors: [hibiscus, light_bodied,
cranberry, dry, rose, white_pepper,
light_bodied, pepper, underripe,
raspberry, fresh, thyme, oregano,
light_bodied, fresh]

Suggestion 2: Hug 2016 Steiner Creek
Pinot Noir (San Luis Obispo County)
Descriptors: [fresh, raspberry, thyme,
pepper, rosemary, sagebrush, dry, sage,
mint, forest_floor, light_bodied,
cranberry_pomegranate, tangy]

Suggestion 3: Comartin 2014 Pinot Noir
(Santa Cruz Mountains)
Descriptors: [vibrant, tangy, cranberry,
hibiscus, strawberry, pepper,
brown_spice, pepper, spice, bay_leaf,
thyme, herb, underripe, raspberry,
cranberry, fruit]
```

We can see that all three wines shown above are Pinot Noirs from California. Consider that grape variety and produced location are not set as features, the result is actually really impressive. In the Jupyter Notebook[8] that provides the source code, they listed out not only three, but ten suggestions. All of them are from California, and nine of them are all Pinot Noirs.

E. Results

There are two possible explanations on why the result of the system perfectly suggesting the geographical location.

When reviewers already knew the grape variety or the produce location, they might be prejudiced on the taste of the wine. Thus they might already had several descriptors coming to their mind once they see the bottle. Then they might ending up with giving similar reviews to wines that produced in the same region. For example reviewers might tend to use the word 'sagebrush' to describe California Pinot Noirs.

On the other hand, maybe even the reviewers all do the reviews blindfolded, they will still give out similar reviews. As mentioned at the beginning of this survey paper, in the wine component paragraph, production area do effects the flavor of wine. Wines that are produced in the same area might actually let the reviewers use similar descriptors consistently.

IV. Extensions

For the first approach, I think we could change the features to attributes that are more accessible, such as grape variety, production region, winery, and vintage year. We can still apply the same method of analyzing partial dependence plot.

For the second approach, there are still quite some more to play with. Once the data is trained, we can let the user to choose a few descriptors that he wants for the wine and the system could suggest wines that are closest to the descriptors. Or we can also put grape variety and production region into consideration. Then we can get a better understanding of correlations between the two factors and certain descriptors. This could result in getting even better suggestions.

V. Conclusion

In this survey paper, I summarized two completely different approach to apply data science to wine. The first is a more intuitive way that simply take the attributes we have for wine into a machine learning model and analyze the results. The results are not so interesting since they mostly fell into our expectations. But I think with the correct chosen attributes, one can still build a recommending system with this approach.

The second approach is much more interesting. This approach breaks down professional wine reviews into relevant descriptors, and calculate the nearest neighbors. The results are incredible. With the descriptors, the suggestions do seem taste similar with the input. Plus, this approach can still be improved and has a few things left to be explored.

References

1. https://www.winespectator.com/articles/how-many-bottles-of-wine-are-there-in-the-world-46410
2. https://archive.ics.uci.edu/ml/datasets/wine+quality
3. https://www.smwewinecompanion.com/page/winemaking-process
4. https://towardsdatascience.com/what-makes-a-wine-good-ea370601a8e4
5. https://wydke.home.blog/2019/01/27/tokaji-furmint-fur-sure/amp/
6. https://towardsdatascience.com/robosomm-chapter-3-wine-embeddings-and-a-wine-recommender-9fc678f1041e
7. https://www.winemag.com
8. https://github.com/RoaldSchuring/wine_recommender/blob/master/creating_wine_review_embeddings.ipynb