Wine Profile Project NLP Ontology K-Mean

For this project, I would like to understand wine reviews since I am working for an alco hol beverage organization. In this paper, I will break down my analysis into following 6 sections. EDA, Data Processing, Classification, Modeling, Ontology and Further Analysis. My goal for this project is to identify distinct flavor profile in the data.

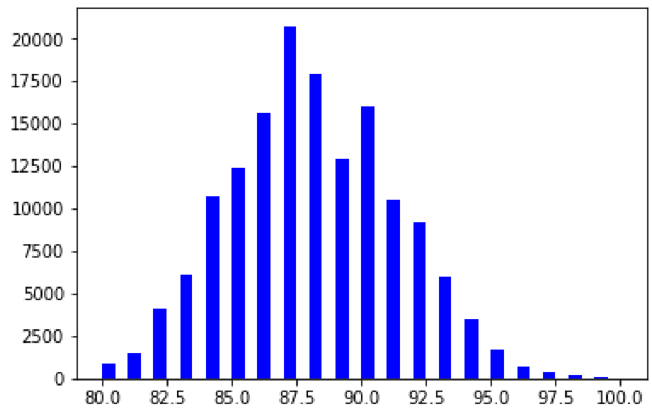
I used wine reviews data off of Kaggle.com. This data has a rich 150,000 different review entries based on wine location, price, description, winery, etc.

1. EDA

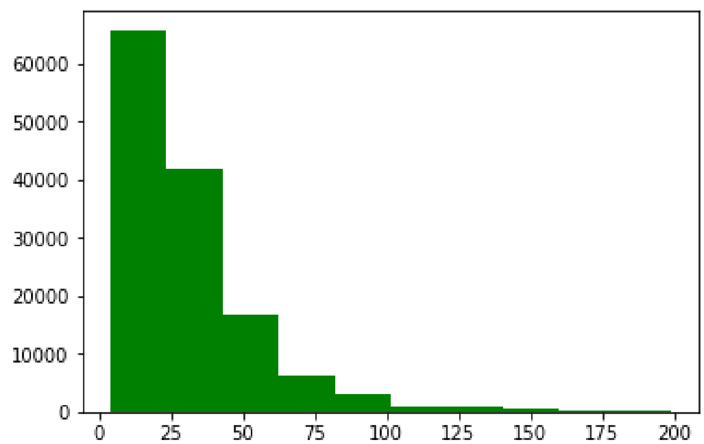
I identified top varieties of the wines as well as the geographic countries by volume of entries. Not surprised, US is dominant in this data, while Italy, France Spain, Chile following based on volume count. Grouping by variety, Chardonnay, Pinot Noir and Cabernet Sauvignon are the top three. I set number of entries at 4,000 to get to a list of top 9, rather than looking at all varieties represented from this data set.

```
array(['Cabernet Sauvignon', 'Sauvignon Blanc', 'Pinot Noir',
       'Chardonnay', 'Syrah', 'Red Blend', 'Riesling',
       'Bordeaux-style Red Blend', 'Merlot'], dtype=object
```

EDA revealed most wines are assigned above 80 points, I broke down all reviews into 5 levels with new column “wine_quality”. Level 1, 80 to 84 points are labeled as “Average”; level 2, 84 to 88 points is “Good”; level 3, 88 to 92 points is “Very good”; then level 4 assigned to 92 to 96 points as “Great”. Level 5 is “Excellent”, assigned to wine points received between 96 to 100.



Points based on graph is evenly distributed. When looking at price of the wine, it's more skewed. The describe () function revealed 75% of the wine is at $40, while mean is at $31 across all dataset.




```
In [69]: df_wine.head(20)
Out[69]:
```

	country	description	points	price	variety	wine_quality	wine_rank	price_val
39623	France	Made from organically grown grapes and boastin...	89	14.0	Rhône-style Red Blend	Very Good	3	0-20
96742	New Zealand	With only 6.5 g/l of residual sugar, this weig...	90	19.0	Pinot Gris	Very Good	3	0-20
34380	Israel	Fleshy black plum and berry aromas open the bo...	90	40.0	Cabernet Sauvignon	Very Good	3	30-50
47819	Argentina	An intense Malbec-led marauder with rubbery bl...	91	50.0	Red Blend	Very Good	3	30-50
71476	Argentina	Overt oak is the first thing you encounter on ...	88	20.0	Malbec	Good	2	0-20
122812	Italy	Here's a standout Sangiovese-based super Tusca...	92	120.0	Sangiovese	Very Good	3	Above 100
32589	New Zealand	This is an amply endowed, round wine, with no ...	91	40.0	Pinot Noir	Very Good	3	30-50
147428	France	A second label of the famous second growth, Ch...	85	50.0	Bordeaux-style Red Blend	Good	2	30-50
57127	Australia	Heady and superripe, this is a huge, mouthfill...	95	50.0	Shiraz	Great	4	30-50
45056	Argentina	Fiery and clipped on the nose, but then it set...	85	12.0	Cabernet Sauvignon	Good	2	0-20
137840	Italy	Typical of this hot vintage, La Fiammenga is a ...	87	28.0	Nebbiolo	Good	2	20-30

While one Italian wine goes as high as $120 classed as "very good". We can find similar points for wines at a fraction of the cost. This suggests price value may not be a good predictor since complexity of the production, availability, and other factors not available here can all play a part to how wine is being priced.

2. Data processing

I created a separate data frame based on added categories and labels, dropping geographic attributes for simplicity. Using existing code to take out stop words, punctuations and standardize to all lower case. Instead of 4, I used len > 5 hoping to bring the data size down. TF-IDF analysis is a bit challenging due to the sizing. But looking at the outlier results. “acidity” has a score of 0.029339, which was marked as outlier. But acidity is a key part of the flavor profile and should not be eliminated.

3. Clustering

I started with K=9 to kick things off. I noticed two out of the three varieties returned in the key words. “Chardonnay”, “Cabernet Sauvignon”. “Merlot” was ranked 9th based on entries count. Perhaps cluster 6 suggests, these 2 varieties share similar key words and should be grouped together.

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
chardonnay	creamy	flavors	cherry	citrus	blackberry	cabernet	fruits	cherries
pineapple	flavors	aromas	raspberry	flavors	flavors	sauvignon	acidity	blackberries
flavors	texture	finish	flavors	finish	tannins	merlot	tannins	currants
acidity	aromas	palate	tannins	aromas	cherry	tannins	character	raspberries
vanilla	finish	acidity	finish	palate	chocolate	flavors	structure	flavors
buttered	vanilla	tannins	aromas	acidity	finish	cherry	texture	tannins
pineapples	acidity	theres	palate	grapefruit	aromas	cassis	structured	spices
orange	chardonnay	bright	acidity	tropical	tannic	verdot	fruity	chocolate
peaches	citrus	herbal	pepper	sauvignon	palate	blackberry	flavors	tannic
finish	palate	offers	offers	offers	cassis	sangiovese	attractive	pepper

Since cluster 0 and cluster 1 both share “Chardonnay”, I am trying K=4 and see if I can get a more distinct clustering. Running K=4 with key words parameter change from 10 to 15, the results returned were less distinctive than anticipated. In one of the clustering, “leather” and “lico rice” came up as key terms. That’s amusing, since I am not sure what leather really tastes like.

Cluster 0	Cluster 1	Cluster 2	Cluster 3
acidity	cherry	flavors	cabernet
fruits	flavors	finish	sauvignon
flavors	blackberry	aromas	merlot
tannins	tannins	palate	tannins
citrus	raspberry	citrus	flavors
character	finish	vanilla	cherry
fruity	aromas	cherries	blackberry
texture	palate	tannins	currant
bright	pepper	theres	cassis
balanced	chocolate	herbal	finish
structure	leather	simple	chocolate
delicious	licorice	offers	aromas
finish	tannic	creamy	verdot
pineapple	offers	chardonnay	palate
chardonnay	cassis	pineapple	blackberries

K=6 parameter however was doing a good job. I then able to classify them into wine profiles/flavors, which I will use them later for my ontology analysis.

Acidity Cherry	**Tannins Structrued Fruits**	**Tropical Vanila**	**Bright Aroma Citrus**	**Spice Pepper Berry**	**Chocolate Blackberry**
Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
cherry	fruits	pineapple	flavors	cherries	blackberry
raspberry	acidity	chardonnay	aromas	blackberries	cabernet
flavors	tannins	flavors	finish	raspberries	flavors
tannins	character	vanilla	palate	currants	tannins
finish	fruity	acidity	citrus	flavors	cherry
aromas	flavors	apricot	acidity	tannins	merlot
palate	texture	creamy	theres	spices	sauvignon
acidity	structure	tropical	offers	pepper	chocolate
pepper	attractive	buttered	tannins	finish	finish
offers	citrus	finish	bright	acidity	aromas

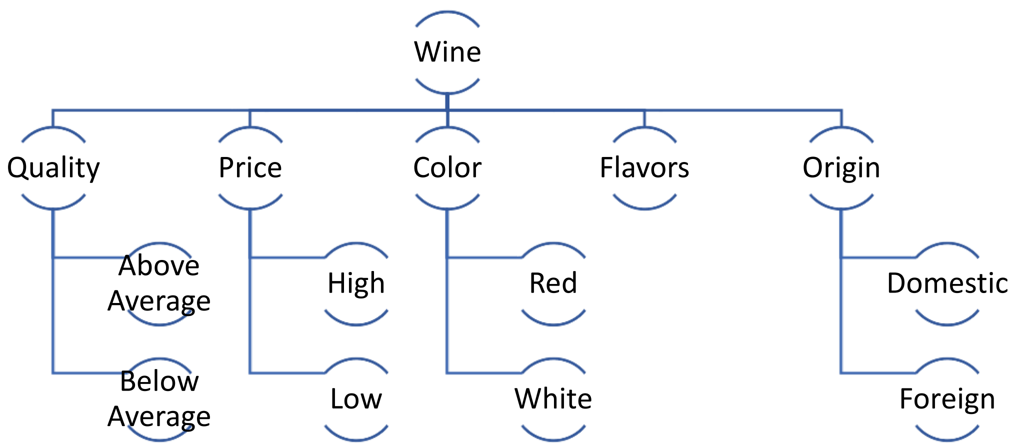
4. Modeling

Stepping into a customer perspective, I would use “points” as a simple reference for purchase, especially if I haven’t tried the product before. From that logic, I decided to try to model between 5 level of point categories against wine descriptions and see how fit is the model. Output was interesting. While I saw a 1.00 precision over at level 5, which is the range of 96-100 points, recall is only 0.50. Level 1 demonstrated the second highest precision at 0.64. None the less, overall prediction accuracy is not as high as I have hoped.

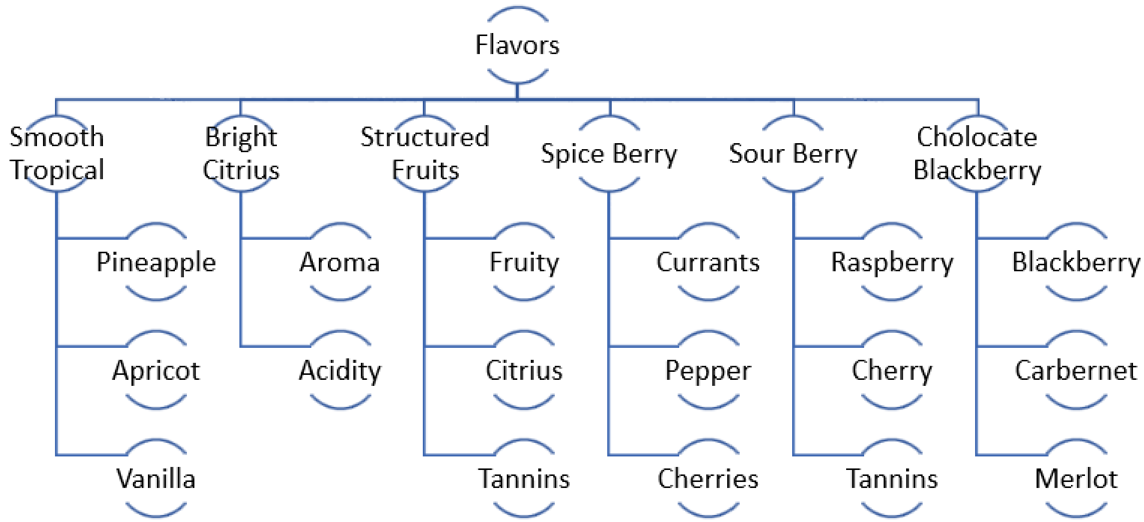
	precision	recall	f1-score	support
1	0.64	0.43	0.51	122
2	0.57	0.77	0.65	304
3	0.51	0.46	0.48	204
4	0.50	0.04	0.07	50
5	1.00	0.50	0.67	2
avg / total	0.56	0.56	0.53	682

5. Ontology analysis

Here is the highest-level attempt in outlining our wine reviews data.



Now focusing one key node here which is the agenda of the analysis, identify distinct favor profiles among the data. Color and even variety are no longer defining element to the flavor profile which making this a independent node on its own. And I would imagine myself classifying my favorite wine, Kim Crawford as “Bright Citrus” as a start. Another nuance is bias around the flavor profiles. Could one person define the same wine using different key terms? Was the wine being paired with food when the review was captured? Even the trained winemasters can choose different words to describe the exact same wine? How do we control consistency among amateurs? As a conclusion, there can be human bias and perception introduced unintentionally in the wine review data.



6. Further analysis:

For further analysis, with some parameter tuning along with additional feature engineering, I hope to improve the accuracy by potentially comparing against other models such as Naive Bayes, SVM, LSTM, etc. Getting more data than the current set can further verify the 6 distinct profiles I have classified.

In addition, if we can get relevent time stamp with the review entries, we should be able to see an overlay of flavor profile versus year which can indicate consumer preference shift. Using the historic trend, geographic location, timeline, we can potentially model the preference shift by location and time, which can support new product launch and positioning.

Additional sentiment analysis can be performed and see how positive, neutral or negative based on our wine descriptions. That can be used as another prediction model.

References

Goutay, O. (2018). Wine ratings prediction using machine learning. *Towards Data Science - Online*, Retrieved from https://towardsdatascience.com/wine-ratings-prediction-using-machine-learning-ce259832b321