Word 2 Wine

Karan Manwani (akp4he)

DS5001 Final Project

Introduction

Archeological records indicate that winemaking has been around for thousands of years and can be traced to as far back as 6000 or 7000 BC. Over time, winemaking has spread and evolved from region to region. With numerous grape varieties and the complexity involved in producing what ends up in a bottle of wine, it is not always easy to know what to expect when reading the grape variety on the label of a bottle, especially if it is one you have never tried or heard of. In this project, my goal is to use information from wine reviews to answer some questions most consumers might have when they see a bottle of wine at their favorite wine store or are browsing the wine list at a restaurant.

The data was pulled from a Kaggle dataset that scrapped over 130,000 reviews from Wine Enthusiast, which is a magazine and website specializing in providing information and reviews on different wines. Using natural language processing techniques, some of the areas I explored include:

- Analysis of high rated vs. low rated wines to see what words describe them.
- Which wine varieties are similar to each other?
- Keynotes that describe each type of wine variety
- What words are associated with each other in the subject of wine?
- Are there some wine varieties that have a more positive sentiment from wine critics, and are there some that tend to have a negative sentiment?

Preprocessing and Standard Text Analytic Data Model Tables

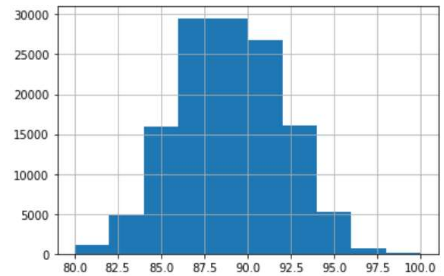
The first step was to preprocess the data by assigning IDs to country, variety, and region to give them numerical values. The data was also assigned an Ordered Hierarchy of Content Objects (OHCO) as follows: 'country_id', 'variety_id', 'region_id', 'points', 'price'. The NLTK package was then used to Tokenize the description column of the data as this column had the reviews written by wine critics for each wine they reviewed. This was then used to build the Token and Vocab tables where parts of speech, stop words, and stems were added to the token terms. The parts of speech tagging were obtained from the Penn Treebank. In addition to stop words from the nltk library, some custom stop words such as wine, drink, and fruit were added. Stop words were then excluded from the models.

How to describe wine

The first area I explored was the nomenclature used in the wine world before going further to analyze the top terms found in highly rated and low-rated wines. Starting with a Bag of Words model to create a Document Term Matrix and using the sum as the term frequency, I generated a Term Frequency–Inverse Document Frequency (TFIDF) matrix. Below is a list of the top 40 words from the corpus with words such as cherry, tannins, apple, peach, citrus, acidity, spice, soft. A lot of them touch on the notes and flavors you would normally hear when wine critics describe a wine.

term_id	term_str	n	num	stop	p_stem	pos_max	tfidf_sum
5571	cherry	29322	0	0	cherri	NN	6.828307
3299	black	29024	0	0	black	JJ	6.499725
27487	tannins	30878	0	0	tannin	NNS	5.520513
30667	white	12916	0	0	white	JJ	5.230045
22515	red	21784	0	0	red	JJ	5.147
15705	lemon	9596	0	0	lemon	JJ	4.945836
20141	peach	8728	0	0	peach	NN	4.805598
1795	apple	13581	0	0	appl	NN	4.791892
1992	aromas	39639	0	0	aroma	NN	4.700557
20975	plum	14991	0	0	plum	NN	4.470995
818	acidity	35003	0	0	acid	NN	4.41059
3115	berry	16983	0	0	berri	NN	4.355353
5872	citrus	11769	0	0	citru	NN	4.30342
4514	cabernet	10794	0	0	cabernet	NNP	3.964014
18826	nose	16963	0	0	nose	NN	3.908194
3304	blackberry	12840	0	0	blackberri	NN	3.884092
23252	ripe	27377	0	0	ripe	JJ	3.696822
27191	sweet	13444	0	0	sweet	JJ	3.598977
7367	crisp	12868	0	0	crisp	NN	3.560192
20156	pear	7992	0	0	pear	NN	3.554653
11499	fresh	17527	0	0	fresh	JJ	3.554532
25900	spice	19233	0	0	spice	NN	3.470792
25576	soft	13664	0	0	soft	JJ	3.332134
24095	sauvignon	7717	0	0	sauvignon	NNP	3.235057
17274	melon	4253	0	0	melon	NN	3.199202
4050	bright	11001	0	0	bright	JJ	3.199117
22322	raspberry	9508	0	0	raspberri	NN	3.18223
8974	dry	17222	0	0	dri	JJ	3.159414
23147	rich	17466	0	0	rich	JJ	3.127386
1842	apricot	3790	0	0	apricot	NN	3.049109
11663	full	16073	0	0	full	JJ	3.04341
11614	fruits	13550	0	0	fruit	NNS	3.035767
19120	offers	12675	0	0	offer	VBZ	3.033201
19313	orange	5841	0	0	orang	NN	3.024201
15855	light	12682	0	0	light	JJ	2.997401
12630	green	9711	0	0	green	JJ	2.978152
7816	dark	12403	0	0	dark	JJ	2.952556
12534	grape	2665	0	0	grape	NN	2.922906
29676	vanilla	11062	0	0	vanilla	NN	2.861391
17363	merlot	5998	0	0	merlot	NNP	2.83438

My next step was to take a closer look at the words associated with describing high-rated wines versus low-rated wines. I first looked at the distribution of the points assigned by the wine critics to determine how to separate the population of high-rated vs. low-rated wines. The below histogram shows a good summary of this, and anything with a score above 91 points (75th percentile) was considered a highly rated wine. Anything below 86 points (25th percentile) was considered a low-rated wine.



After filtering on the points to get two token tables for the two categories and running the TFIDF function similar to the process described above, I decided to focus on adjectives since those would be the terms used to describe a wine. Below are the top 20 adjectives for each of the two categories.

Top 20 adjectives in high rated wines

term_id	term_rank	term_str	n	num	stop	p_stem	pos_max	tfidf_sum
1904	1	black	7707	0	0	black	JJ	2.072002
13054	4	ripe	5986	0	0	ripe	JJ	1.245601
12993	7	rich	4974	0	0	rich	JJ	1.188223
4351	9	dark	3621	0	0	dark	JJ	1.191889
6646	10	full	3451	0	0	full	JJ	1.014851
12621	11	red	3401	0	0	red	JJ	1.05561
5046	19	dry	2558	0	0	dri	JJ	0.792325
6560	25	fresh	2459	0	0	fresh	JJ	0.876863
9158	28	long	2371	0	0	long	JJ	0.983287
3657	29	concentrated	2365	0	0	concentr	JJ	0.696236
6134	33	fine	2175	0	0	fine	JJ	0.748926
3607	34	complex	2035	0	0	complex	JJ	0.772937
15341	35	sweet	2034	0	0	sweet	JJ	1.041414
7173	38	great	1966	0	0	great	JJ	0.812476
17234	39	white	1964	0	0	white	JJ	1.323447
4503	45	delicious	1838	0	0	delici	JJ	0.761542
8914	50	lemon	1717	0	0	lemon	JJ	1.104208
5265	53	elegant	1654	0	0	eleg	JJ	0.726083
2300	56	bright	1572	0	0	bright	JJ	0.74127
9940	57	mineral	1542	0	0	miner	JJ	0.739516

Top 20 adjectives in low rated wines

term_id	term_rank	term_str	n	num	stop	p_stem	pos_max	tfidf_sum
9003	4	red	3583	0	0	red	JJ	3.789385
10979	5	sweet	3511	0	0	sweet	JJ	2.655784
10283	6	soft	3289	0	0	soft	JJ	2.650626
3566	7	dry	3094	0	0	dri	JJ	2.352801
6316	9	light	3006	0	0	light	JJ	2.516272
5077	12	green	2482	0	0	green	JJ	2.277887
4634	14	fresh	2414	0	0	fresh	JJ	2.467559
9311	15	ripe	2353	0	0	ripe	JJ	2.160009
10072	17	simple	2175	0	0	simpl	JJ	1.783055
1231	23	black	1939	0	0	black	JJ	2.806336
5360	26	herbal	1717	0	0	herbal	JJ	2.260479
12457	27	white	1657	0	0	white	JJ	3.051654
4969	28	good	1635	0	0	good	JJ	1.634103
1529	37	bright	1337	0	0	bright	JJ	1.862056
3647	43	easy	1269	0	0	easi	JJ	1.698847
4715	44	full	1256	0	0	full	JJ	1.63176
6408	47	little	1244	0	0	littl	JJ	1.319641
1225	51	bitter	1163	0	0	bitter	JJ	1.540235
2282	54	clean	1101	0	0	clean	JJ	1.692061
11093	55	tannic	1083	0	0	tannic	JJ	1.397696

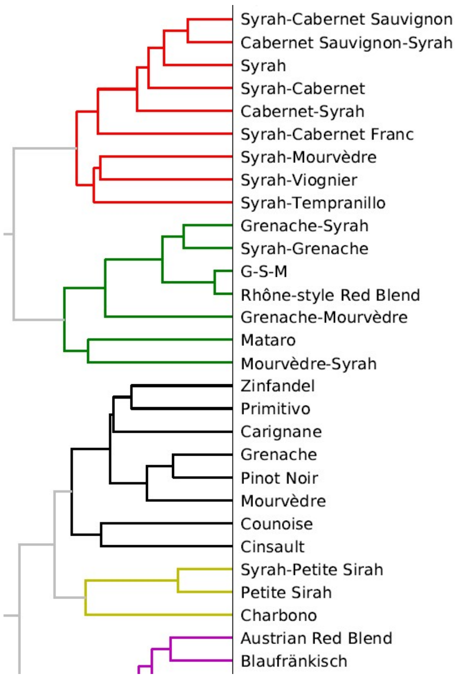
As can be seen from these tables, some of the top adjectives that tend to be used for describing high-rated wines are ripe, rich, dark, full, fresh, long, concentrated, complex, and elegant. Whereas, for low-rated wines, they tend to be sweet, soft, simple, light, herbal, dry, and bitter.

Similarities between different Varieties - Clustering

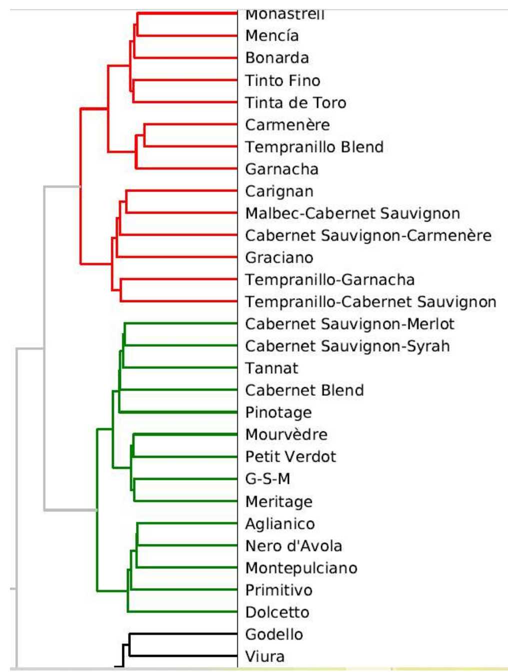
Next, I explored which of the over 700 varieties in the data set appeared to be similar to one another based on the descriptions. The idea behind looking at this is to help determine what other wine someone might like based on an existing variety they are fond of or what to stay away from if there is any particular variety they don't care for.

After creating a Token table grouped by variety_id as the index and using this to set up the TFIDF matrix, I built a document pair table. The Jaccard and Cosine methods were then used to compute the distances between each pair and hence comparing each variety with each other. The results of the two can be seen in dendrograms, and below is an extract of it (the full dendrograms are attached separately given their size)

Cosine Dendrogram



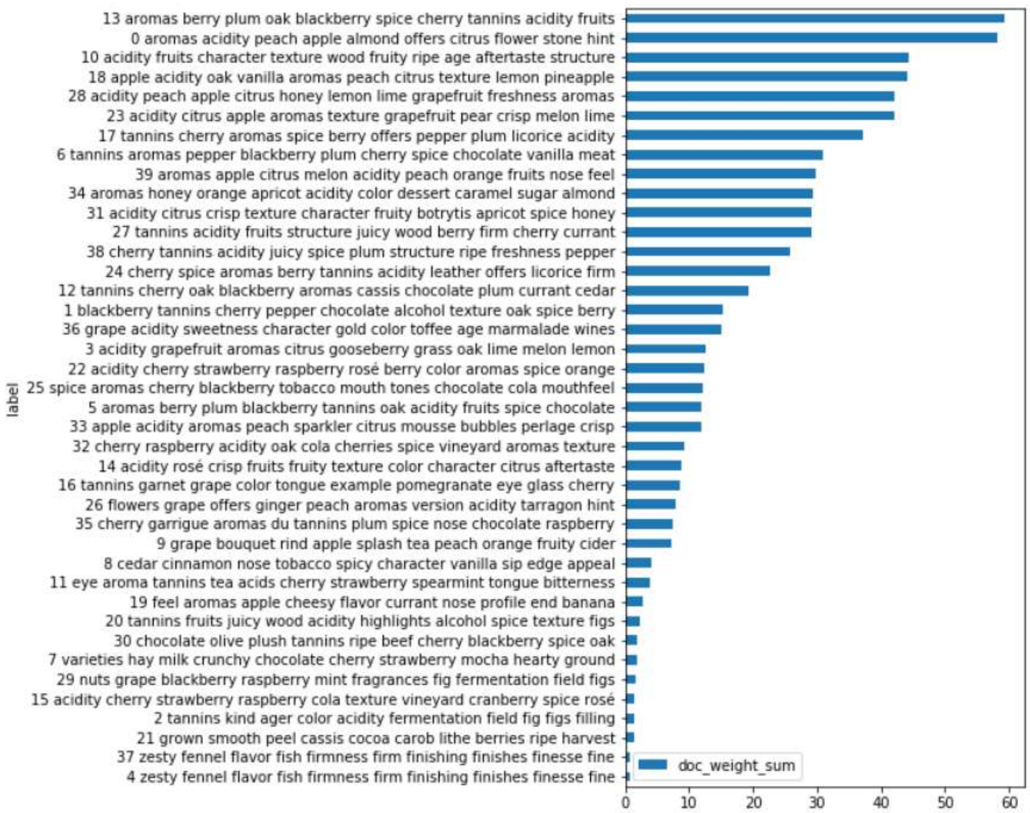
Jaccard Dendrogram



Some interesting results can be found here. First, looking at the Cosine results, Pinot Noir matches with Grenache, and after doing some research, it does appear there are similarities between these two varieties. Here is a comment on these two varieties from a website called garnachagrenache.com “Both grapes planted on sandy soils give fresher, lighter wines with more aromas, while red clay ones have longer flavors and more structure.” Another publication, called Wine folly also mentions the similarities between these two grapes. Wine folly also mentions the variety called Primitivo, which is from Southern Italy being similar to Zinfandel, and the cosine dendrogram above also picked up on this. The Jaccard results didn’t stand out as much, although more research is needed to determine this. That being said, both models could be improved if we consolidated varieties that are the same but show up differently in the data, such as Syrah – cabernet sauvignon and cabernet sauvignon – Syrah or the same grapes that have multiple names, such as Tinta del Toro and Tinta del Pais. In addition, it might be useful to remove varieties where there are limited data points.

Key notes by Variety – Topic Modeling

My next goal was to identify the keynotes that describe each type of wine variety, and I was able to achieve this using Topic modeling. Starting with the token table that was grouped by variety_id as the index, I grouped all the terms for each variety together. This ended up being a table with over 700 rows with each row containing all the tokens for the corresponding variety. The Latent Dirichlet Allocation (LDA) model with 5 iterations and 40 topics was used to build the topic model. The document to topic (theta) and a term to topic (phi) matrices were built with the LDA model to perform this analysis. The top 10 terms were selected for each topic based on their weight to build the Topics table. The document weight score was then pulled into this able from the Theta table, and below is a summary of the document weight for each topic in the entire corpus:



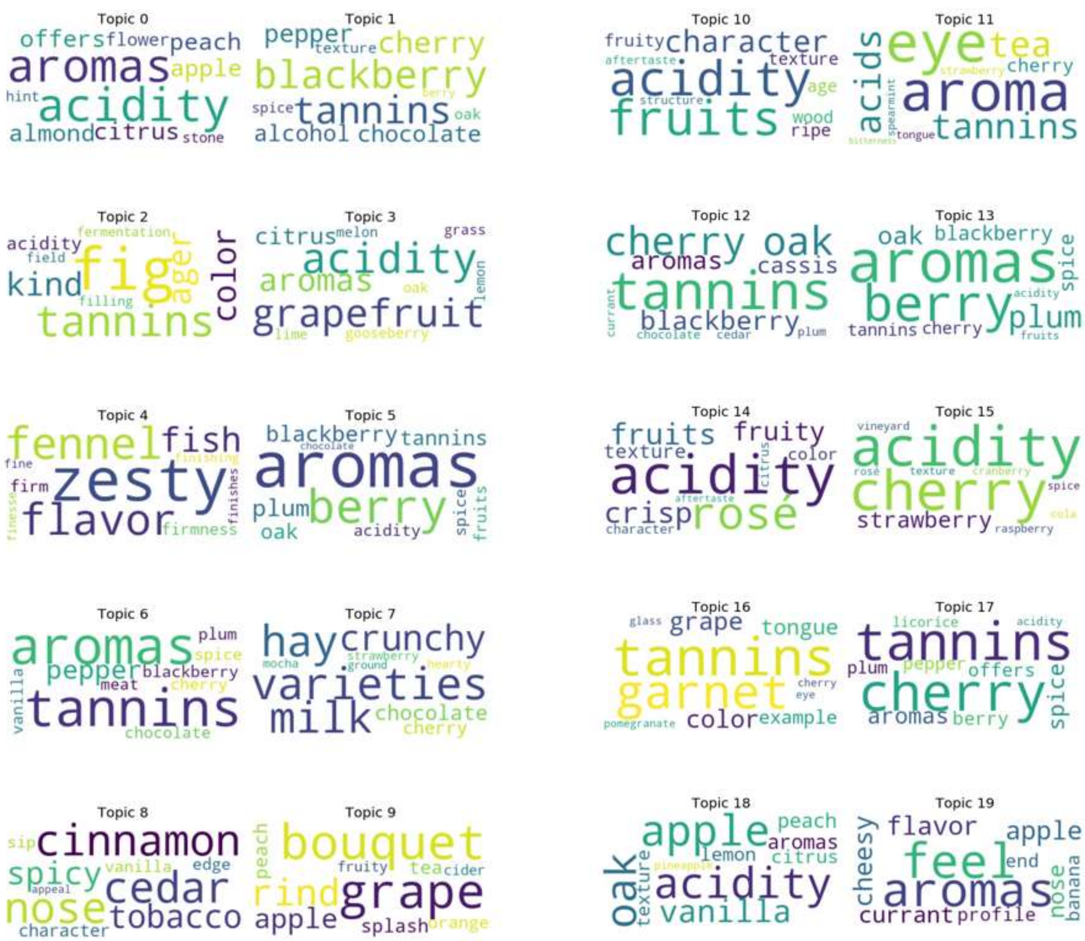
As can be seen above, topic 13 and topic 0 have the highest weights. The first one seems to be words you would find in red wine and the second one looks like white wine terms. There are also other topics with a lot of weight in the document. To explore further, I summarized a table with the top topic by variety and saved it as a Document to topic concentrations table. I have pulled some of the more commonly known wine varieties from that table and split them into Reds and Whites in the table below.

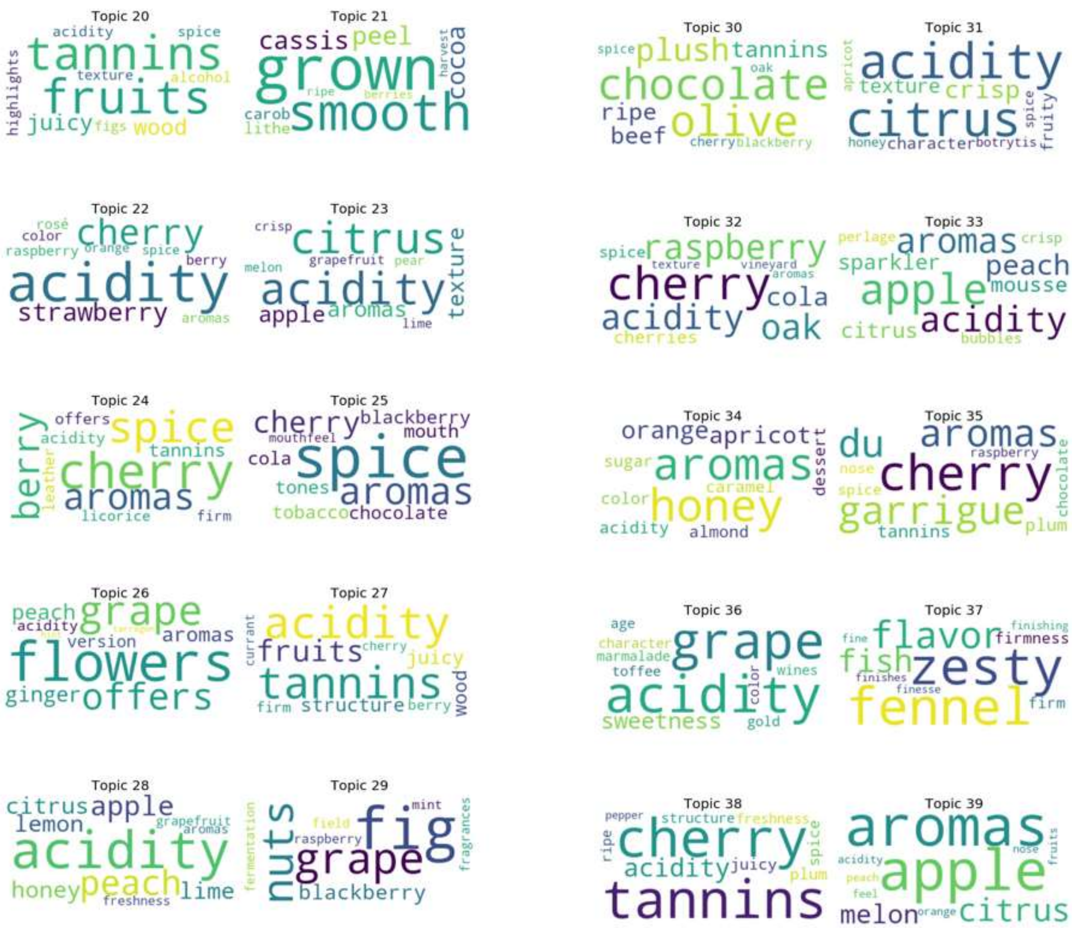
Red Wines			**White Wines**	
variety	**top_topic**		**Variety**	**top_topic**
Barbera	17		Albarino	39
Cabernet Sauvignon	12		Chardonnay	18
Merlot	12		Chenin Blanc	18
Nebbiolo	17		Gewurztraminer	28
Pinot Noir	32		Godello	39
Sangiovese	17		Pinot Grigio	0
Syrah	6		Riesling	28
Tempranillo	13		Sauvignon Blanc	23
Zinfandel	1		Verdejo	39

The below are word clouds to help visualize what are the top terms in each topic. Looking at Merlot for instance, we see Topic 12 best describes it – cherry, oak, blackberry, tannins, chocolate, cedar, plum. Sangiovese comes up with topic 17, which is cherry, spice, tannins, plum, berry, licorice, pepper. In the whites, Riesling is described by topic 28 which is lemon, acidity, apple, honey, peach, lime, grapefruit, freshness. Chardonnay is described by topic 18, which is apple, vanilla, acidity, peach, citrus, lemon, pineapple, oak. This can also be used to see which wines are similar to each other, for instance, Albarino and Verdejo are both covered by topic 39. Another interesting topic is number 34 with words such as honey, orange, apricot, dessert, caramel, sugar, almond. This seems to describe the dessert or fortified wines, and here we will find varieties such as Sherry and White Port.

There are some areas to explore further to determine if the model needs to be improved or if the data needs pre-processing improvements. For instance, topic 19 had two varieties – Franconia and Macabeo-Chardonnay. One is a red and the other is a white so this needs to be explored further to determine if the assigned topics are accurate representations, but it could potentially be an issue of limited data points available for these wines in this dataset.

Word Clouds for Topics





Word Embeddings – Word 2 Vec

This section examines what words are associated with each other in the discourse of wine. For instance, when we type the word pinot, what are we expecting to follow it? A good guess would be noir. A word2vec model can help answer this. Starting with the Token table, and grouping it by variety_id and then splitting the terms into a list for use in the word2vec model. Some of the parameters used for the word2vec model were 5 for window, and 500 for the minimum count given that a lot of words are repeated in this corpus. Each word was then assigned coordinates in vector format and the tsne library was used to transform the top 2 pca components into x and y coordinates for plotting.

Below are some interesting results from the model. Starting with the first three words which are the first word in varieties. The words next to pinot are noir and gris which are expected since these are second word of two popular varieties. Same with Petit where Verdot is associated with it. Interestingly for Cabernet, Sauvignon does not appear, but this could be due to the model parameters where by this word was dropped from the model as it may not have appeared over 500 times. But in all three examples, other varieties appear associated with them in addition to the words we commonly expect with them.

```
model_variety.wv.most_similar('pinot')

[('chardonnay', 0.7205592393875122),
 ('noir', 0.7156832218170166),
 ('gris', 0.6623498797416687),
 ('sparkling', 0.5744326114654541),
 ('sparkler', 0.5231398344039917),
 ('champagne', 0.5206406116485596),
 ('rosé', 0.5021917223930359),
 ('nero', 0.5001140832901001),
 ('delicately', 0.4940997064113617),
 ('70', 0.46875929832458496)]
```

```
model_variety.wv.most_similar('petit')

[('verdot', 0.958683967590332),
 ('malbec', 0.8525865077972412),
 ('cab', 0.7792518734931946),
 ('mourvèdre', 0.7379230260848999),
 ('parts', 0.719452977180481),
 ('merlot', 0.7182372212409973),
 ('syrah', 0.7176245450973511),
 ('includes', 0.703104555606842),
 ('sirah', 0.6925644874572754),
 ('8', 0.6890406608581543)]
```

```
model_variety.wv.most_similar('cabernet')

[('cab', 0.8544741868972778),
 ('merlot', 0.8157801032066345),
 ('malbec', 0.7703526020050049),
 ('verdot', 0.6977696418762207),
 ('syrah', 0.689529538154602),
 ('petit', 0.6603595018386841),
 ('tempranillo', 0.6238324642181396),
 ('parts', 0.5817668437957764),
 ('blended', 0.5793178677558899),
 ('equal', 0.5781236290931702)]
```

Below are a few more examples. The words associated with meat are cured, smoked, beef, grilled. Similarly with leaf, we get oregano, dill, herbs and with years we get five, several, 3, and 4.

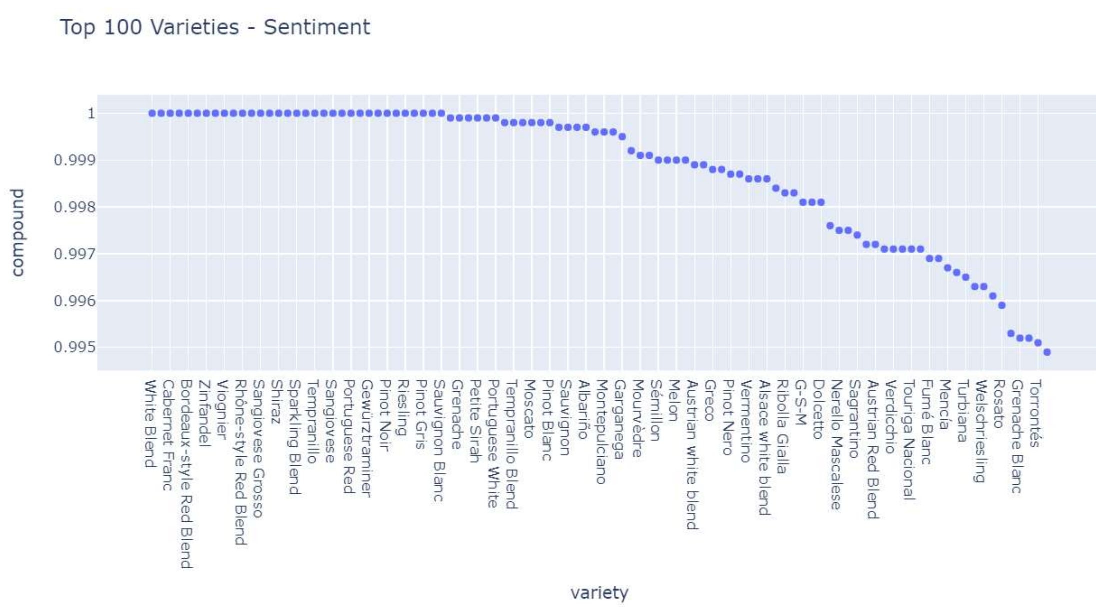
```
model_variety.wv.most_similar('meat')

[('cured', 0.8590074777603149),
 ('smoked', 0.8571118116378784),
 ('beef', 0.8128323554992676),
 ('meats', 0.7970245480537415),
 ('sauce', 0.7884708642959595),
 ('bacon', 0.7444738149642944),
 ('grilled', 0.7282095551490784),
 ('steak', 0.7215409278869629),
 ('mushroom', 0.7212794423103333),
 ('game', 0.6736063361167908)]
```

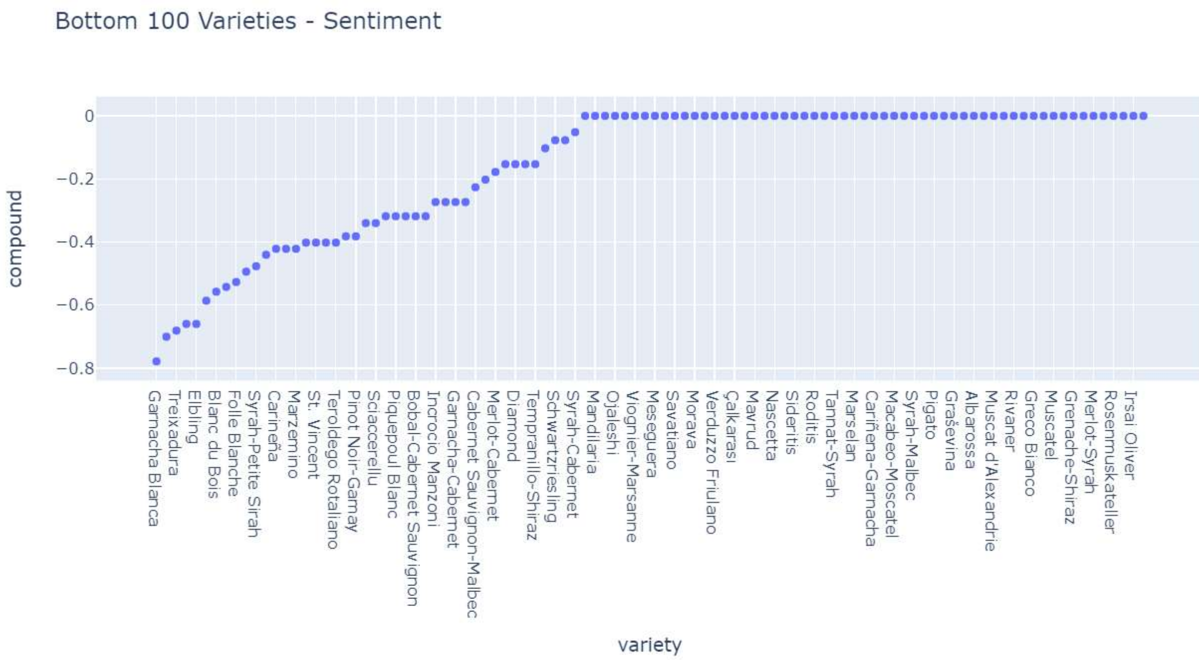
```
model_variety.wv.most_similar('leaf')

[('tomato', 0.8462241888046265),
 ('oregano', 0.8023450374603271),
 ('dill', 0.7938407063484192),
 ('chopped', 0.7328243255615234),
 ('herbs', 0.71958327293396),
 ('thyme', 0.7137711644172668),
 ('bramble', 0.712658166885376),
 ('eucalyptus', 0.7055050134658813),
 ('fennel', 0.6750075221061707),
 ('mint', 0.664734423160553)]
```

```
model_variety.wv.most_similar('years')

[('five', 0.896599292755127),
 ('several', 0.866423487663269),
 ('3', 0.8475190997123718),
 ('4', 0.8430639505386353),
 ('next', 0.8370513916015625),
 ('another', 0.8242892026901245),
 ('least', 0.81653892993927),
 ('2', 0.7969988584518433),
 ('four', 0.7809898853302002),
 ('two', 0.7608671188354492)]
```

Sentiment Analysis on Varieties

Finally, this section describes the results from sentiment analysis done on the varieties to see if there are varieties that tend to have a positive or negative sentiment by wine critics in this dataset. The Vader sentiment analyzer was used for this analysis. The sentiment analyzer was run on the table which grouped all the descriptions of each variety together (similar table as the one used in topic modeling). The Vader results produced a positive, negative, neutral, and compound score for each variety. Below charts are the top 100 and bottom 100 varieties based on the compound score. The compound score is normalized between -1 and +1.





In the top 100, we see wines such as Pinot Noir, Sangiovese, Tempranillo. These are some of the popular varieties and perhaps they tend to have more reviews and a lot of quality wines. On the flip side, some of the varieties with negative sentiments are Garnacha Blanca, and a blend of Tempranillo-Syrah. Taking a look at the Garnacha Blanca, here is a review of a wine with a rating of 81: “This fruitless Garnacha Blanca smells a bit too much like rotting compost. A plump palate holds bitter flavors in front of a pithy finish. Overall, there's very little that's good about this.” And here is a review of another one with a rating of 87: “vanilla aromas give this Garnacha Blanca a cookie-like nose. Plump and round across the palate, this tastes pithy like citrus peel, with a mildly bitter aftertaste. A simple finish doesn't offer anything new.” These are not very positive reviews and can explain some of the reasons why the sentiments are low for this wine. But further analysis is needed to validate this across the various varieties and determine if data quality issues are distorting the results.

Conclusion and Future work

As mentioned at different points in the paper, there are areas to explore further to help improve the results of the models and analysis. I would like to explore if more stop words from the corpus should be added, consolidate different varieties that are the really the same but have different names in the dataset, and perhaps drop some varieties where we have limited data points. I would also like to further examine some of the results from areas such as sentiment analysis. To conclude, I believe the results from this analysis can help with understanding what to expect when trying a certain type of wine without have to sample every single variety. Whether it is using topic modeling or similarities from clustering to determine which wine to pick up based on known preferences or deciding to try something new based on sentiment analysis, this study can help with those decisions.

Bibliography

5 Value Alternatives to Your Favorite Wine Varieties

https://winefolly.com/lifestyle/5-cheap-alternative-wine-varieties/

The new Pinot Noir?

https://garnachagrenache.com/the-new-pinot-noir/