# Predicting Wine Varietals from Professional Reviews

By Ron Tidhar, Eli Ben-Joseph, Kate Willison 11th December 2015

CS 229 - Machine Learning: Final Project - Stanford University

#### Abstract

This paper outlines the construction of a wine varietal classification engine. Through use of topic analysis, word stemming and filtering, a Naïve Bayes classification algorithm performed with a surprising degree of accuracy. This research, therefore, represents exciting first steps in applying Machine Learning techniques to an area not well studied in traditional research.

### 1 Introduction

While many of us enjoy a good glass of wine, it can be difficult at times to put a finger on what exactly draws us to any particular bottle. Given the qualitative breadth and scope of hundreds of different wine varietals¹ - ranging from the full-bodied Petit-Sirah, to the light and sweet Chenin-Blanc - it's no wonder that Sommeliers and laypeople alike have striven to share their experiences through developing a common vocabulary around the qualities and aromas they find in each glass.²

Although this vocabulary may be difficult to navigate for the uninitiated, among professional wine reviews, one often finds distinct and recurring descriptors for each varietal. As such, in the following study, we aim to use data from a large sample of professional reviews in combination with various Machine Learning techniques to build a classification model for a number of common wine varietals. This would not only enable categorization based on provided wine-tasting terms (which has applications for recommender models and blind-tasting<sup>3</sup> education),

but would also allow one to relate similar wines to one another.

## 2 Data

Data were scraped from http://www.klwines.com/using a BeautifulSoup<sup>4</sup>-based python script between 10-30-2015 and 11-04-2015. For each of 35 wine styles categorized by the site, data for at most 2000 unique examples was collected, including varietal, professional and nonprofessional reviews, name, country, region, appellation, alcohol content and persistent web address. While at most five reviews were collected for each wine, a large portion of the dataset had no associated reviews - these were removed from the final dataset. In total, therefore, 32,892 reviews were collected for use in the analysis.

# 3 Modeling

In order to measure a baseline performance, a simple multi-class one-against-all classification model was built. This model was implemented across all 35 wine varietals using Vowpal Wabbit<sup>5</sup>. Words were tokenized and grouped across reviews for a given wine, and analyzed as a simple "bag-of-words". Training 80% of the data in a single pass with a logistic loss function, the resulting model correctly classified 61% of the wines in the test set - far better than the approximate prior of 1/35 = 2.85%.

<sup>&</sup>lt;sup>1</sup>'Varietal' refers to the type of grape primarily used in making a wine, so that a wine labeled as 'Chardonnay' must be made from at least 75% Chardonnay grapes. This is in contrast to classification systems used widely in Europe, whereby blends are labeled by region rather than the grape variety (e.g. Bordeaux will commonly be a blend of Merlot, Cabernet Franc and Cabernet Sauvignon)

 $<sup>^2{\</sup>rm For}$  examples of common wine-descriptive words used by reviewers, see well known critic Robert Parker's wine glossary - https://www.erobertparker.com/info/glossary.asp

<sup>&</sup>lt;sup>3</sup>'Blind tasting' is the practice of tasting a wine without knowing any information about its origin, varietal or production, with the goal of guessing each from the qualities of the wine itself.

<sup>&</sup>lt;sup>4</sup>Richardson, L., "Beautiful Soup", Crummy, http://www.crummy.com/software/BeautifulSoup/, 2015

<sup>&</sup>lt;sup>5</sup>Langford, J., "Vowpal Wabbit", Microsoft Research, https://github.com/JohnLangford/vowpal\_wabbit/wiki

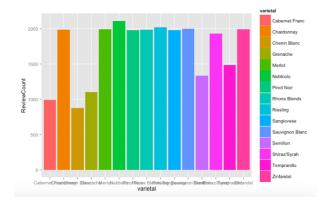


Figure 1: Top 10 wine varietals by number of collected reviews.

Given that the basic premise of the study was validated by this simple analysis, further model refinement and strengthening was sought. As a first step, a list of wine-review-specific stop words was created. The words listed were ones that indicated the varietal directly or indirectly (such as 'Chardonnay' or 'chateau'), or else represented information that wouldn't be available to a blind-taster (e.g. 'hectare'), and so were removed them from the data set.

In order to counter the high variance demonstrated in the initial learning curves, model simplification was implemented, through both class and feature set (i.e., input words) reduction. This was accomplished by building a 20-category topic model using Latent Dirichlet Allocation (LDA)<sup>6</sup> in MALLET<sup>7</sup> a Java-based package for statistical natural language processing - re-estimating Dirichlet parameters every 10 iterations. Following this, calculations were made for the cumulative probabilities of each word across all wine-related categories, as defined by the model output. By stemming both the resulting word list, as well as the words contained in the training data (so that, for example, 'spice', 'spicey' and 'spiciness' would map to the same feature), it was possible to then filter the training features. In addition, in the final model, only varietals for which there were at least 200 reviews were included; a total of 23 predicted categories.

With such data treatment techniques, a simple Naïve Bayes classification algorithm was run on the

Rank	Topic	Weight	Rank	Topic	Weight	Rank	Topic	Weight
1.	fruit	577.6	11.	drink	263.9	21.	good	178.6
2.	finish	480.5	12.	black	263.8	22.	fresh	168.9
3.	flavors	460.9	13.	dark	250.9	23.	licorice	161.5
4.	tannins	346.3	14.	long	245.8	24.	bright	155.6
5.	red	336.7	15.	oak	225.6	25.	full-bodied	151.1
6.	aromas	320	16.	ripe	215.1	26.	berry	143.2
7.	notes	309.1	17.	acidity	213.8	27.	raspberry	137.4
8.	cherry	275.5	18.	rich	204.2	28.	blackberry	134.4
9.	palate	267	19.	nose	203.8	29.	floral	130.1
10.	sweet	266.3	20.	spice	190.3	30.	white	126.7

Figure 2: Table showing a sample of the 30 top wine-related words, as classified by the unsupervised LDA algorithm.

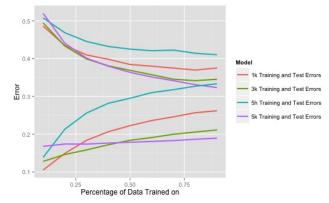


Figure 3: Learning curves for 500, 1000, 3000, and 5000 feature word-based models.

data. Cross validation was used to assess the optimal number of word features (selected from the LDA topic analysis) to be used in the model.

## 4 Results

#### Learning Curves:

Given that the initial model exhibited high variance, one clear strategy for improvement was to reduce the feature set. To test this, cross validation was used to find the most suitable number of word features. The ranked list shown in Figure 2 was used to filter four Naïve Bayes bag-of-words models fit using ten-fold cross validation (10% holdout); one each for the top 500, 1000, 3000 and 5000 words. Using a set of learning curves resulting from this analysis, algorithm success rates were compared. As can be seen in Figure 3, using 5,000 of the top descriptive words to train and test the model yielded the best results. Discussion of this somewhat surprising result follows later.

<sup>&</sup>lt;sup>6</sup>Blei, David M., Ng, Andrew Y., Jordan, Michael I (January 2003). Lafferty, John, ed. "Latent Dirichlet allocation". Journal of Machine Learning Research 3 (4–5): pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

<sup>&</sup>lt;sup>7</sup>McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

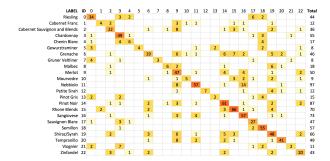


Figure 4: Test set confusion matrix. Rows represent true varietal value, while columns correspond to the predicted category for each example: cell values represent the total count of examples that correspond to each true/predicted value.

### Model quality:

The confusion matrix is useful to further assess the performance of the most accurate model. In so doing, it can be observed which varietals are most accurately classified (i.e. have a higher proportion of their row sum in the cell along the matrix diagonal), as well as which varietals they are most often misclassified as (cells that lie outside of the matrix diagonal).

The confusion matrix shows us that those wines that are misclassified are more often than not assigned to a varietal that is descriptively similar. For example: in the test set, Riesling is classified correctly 77% of the time (34/44), but misclassified as Sauvignon Blanc in 14% of examples. This is fairly understandable, as both are pale green-gold wines, mostly unoaked, and often not very high alcohol, with (depending where they are grown) 'green' fruit flavors and high acidity.<sup>89</sup> Given this perspective, the 68% accuracy that the model achieves on the test set is all the more impressive.

### Varietal proximity:

Another insight afforded by the LDA topic analysis involves varietals' characteristic proximity to one another: because the algorithm is unsupervised, topics are each generally mapped to reviews of wines from more than one varietal. Intuitively, a topic composed primarily of two varietals indicates that those varietals are likely similar in the dimensions indicated by the highly-weighted words associated with that topic.

Figure 5 demonstrates an example of this with a subset of topics from the LDA analysis.

# 5 Applications

There are various applications of the wine-classifier model outside the realm of academia. The most straightforward application is a simple wine recommender: given a set of descriptors that represent one's general tastes (in terms of flavors, textures, and aromas), the model can recommend wines that best fit that profile in a rank-ordered list (see Figure 6 for an example of this in action).

This would allow someone to consider and gain exposure to wines which they may not otherwise have been acquainted. For many who are interested but new to the world of wine, understanding the nuances in tastes and aromas can seem like a daunting task that presents a barrier to enjoying wine to its fullest.

Another application of the model is as a tool for blind tasting. The wine classifier could serve as a decision guide: as the user inputs more descriptors, the model would update the likely matches and use the coefficients to provide some motivation for why a particular varietal is likely. Though the model is not a perfect predictor, this would nevertheless be a valuable educational tool.

Lastly, as many of the reviews analyzed also contained recommended food pairings, the wine classifier model could be modified to recommend wine-food pairings. During data preprocessing for our main model, these food mentions were filtered out, as the associated words were relatively uncommon (and therefore did not make the 5,000-word cut). By modifying the feature inputs to include food-related words, it would be possible to build a model that would recommend top food pairings for a given varietal. This could be useful for anyone looking to pair a wine with a nice meal or vice versa, including both restaurants and home chefs.

# 6 Conclusions and Future Work

With a final classification accuracy of 68%, it is clear that there is still room for improvement. The learning curve for the 5,000 word feature model indicates a large separation between the training and testing error rates. Given a desired performance on the order of 80-90%, the curve implies that the model still exhibits a high degree of overfitting (i.e., variance). To rectify this problem, two strategies may help.

 $<sup>^8 \</sup>rm Gregutt,~Paul.~``White Wine Basics'' Wine Enthusiast (2011) http://www.winemag.com/2011/03/16/white-wine-basics/$ 

<sup>&</sup>lt;sup>9</sup>Laube, James; Molesworth, James. "Varietal Characteristics" Wine Spectator (1996) http://www.winespectator.com/webfeature/show/id/Varietal-Characteristics\_1001

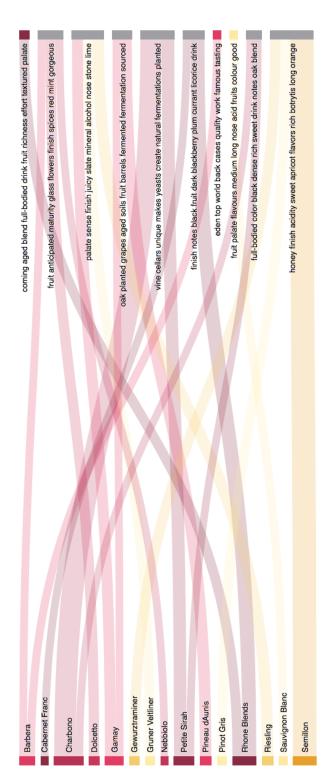


Figure 5: Subset of topics demonstrating result of LDA analysis. An interactive version of this figure can be found at http://web.stanford.edu/~kawi/wine\_model/category\_vis.html

your input: full-bodied blackberry pepper blueberry spice dates plum

Shiraz/Syrah : 0.665593470162 Zinfandel: 0.078221896624 Rhone\_Blends : 0.0632306537527 Grenache: 0.0431642074862 Petite\_Sirah : 0.0304376567416 Merlot: 0.0274437006157

Cabernet\_Sauvignon\_and\_Blends : 0.025783071175 Tempranillo : 0.0196141402865

Mourvedre : 0.0151156202084 Sangiovese : 0.0133922511295 Malbec: 0.0102571300437 Pinot\_Noir: 0.00376997569854 Cabernet\_Franc : 0.00266901004762 Nebbiolo: 0.00130366154699 Gruner Veltliner: 2.87890203521e-06 Riesling: 2.00106922034e-07 Gewurztraminer: 1.98052949445e-07 Chenin\_Blanc : 9.81887695192e-08 Pinot\_Gris : 6.5835614815e-08 Semillon: 6.03794598567e-08 Chardonnay: 2.2310794508e-08 Sauvignon\_Blanc : 1.71910324801e-08 Viognier: 1.35137078632e-08

Figure 6: A sample output predicting what a user may enjoy based on their input.

First, increasing the size of the training data set will help reduce variance, and will serve to increase model robustness. This is a relatively straightforward improvement, and can be done by finding other review sites for which scraping is permissible.

Secondly, reducing the size of the feature set (i.e., training on less words) is also likely to improve the model. Though this was tried (and rebuffed) with the cross-validation analysis, there are still some further optimisations to be considered.

Many descriptive words used in wine tasting require qualifiers or modifiers in order to be most meaningful. For example, while "acidity" may be picked up as a feature, it is most descriptive with a modifier. The difference between a "high acidity" and "low acidity" wine is significant. This is most likely the cause of the counter-intuitive cross validation analysis. As a result, a selected bigram analysis may serve to reduce the feature set, by allowing for a smaller set of more descriptive features.

Ultimately, this paper presents promising first steps towards building a robust wine varietal classification engine. By implementing the suggested further improvements, many of the useful applications can easily be realised.

## 7 Bibliography

Blei, David M., Ng, Andrew Y., Jordan, Michael I (January 2003). Lafferty, John, ed. "Latent Dirichlet allocation". Journal of Machine Learning Research 3 (4–5): pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

Gregutt, Paul. "White Wine Basics" Wine Enthusiast (2011) http://www.winemag.com/2011/03/16/white-wine-basics/

Langford, J., "Vowpal Wabbit", Microsoft Research, https://github.com/JohnLangford/vowpal\_wabbit/wiki

Laube, James; Molesworth, James. "Varietal Characteristics" Wine Spectator (1996) http://www.winespectator.com/webfeature/show/id/Varietal-Characteristics\_1001

McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

Parker, R., "A Glossary of Wine Terms", eRobertParker.com, https://www.erobertparker.com/info/glossary.asp

Richardson, L., "Beautiful Soup", Crummy, http://www.crummy.com/software/BeautifulSoup/, 2015