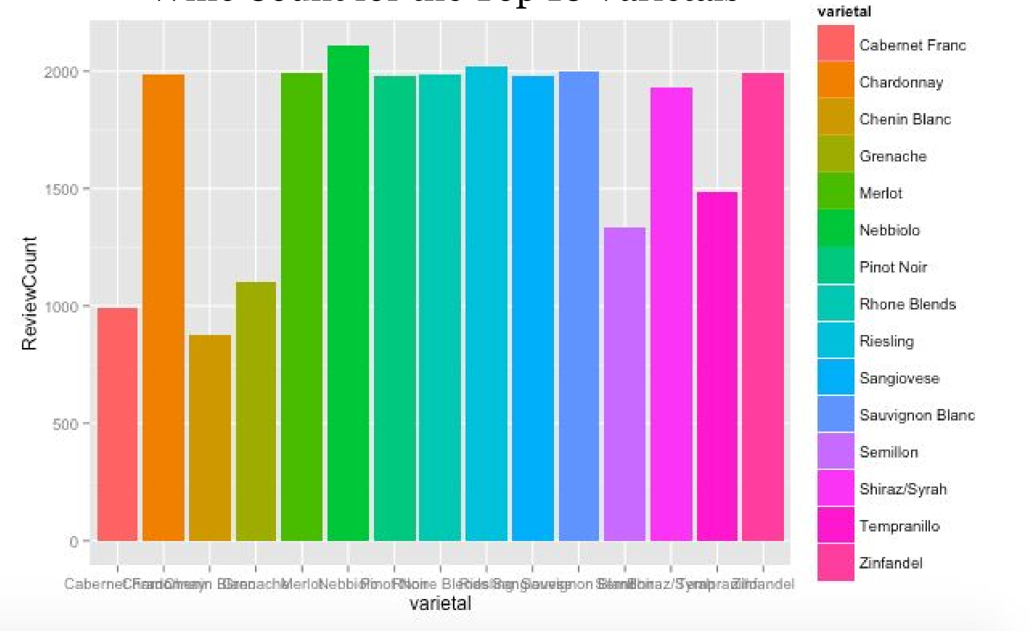
Predicting Wine Varietals from Professional Reviews

Eli Ben-Joseph | Kate Willison | Ron Tidhar

Project Goals

- Build a classifier to predict wine varietals based on reviews
- Learn to mine, clean, and process a large word-based dataset to prepare it for analysis
- Understand the differences between each model we train, including their pros and cons for our dataset, in order to produce a high-quality model

Wine Count for the Top 10 Varietals



MALLET

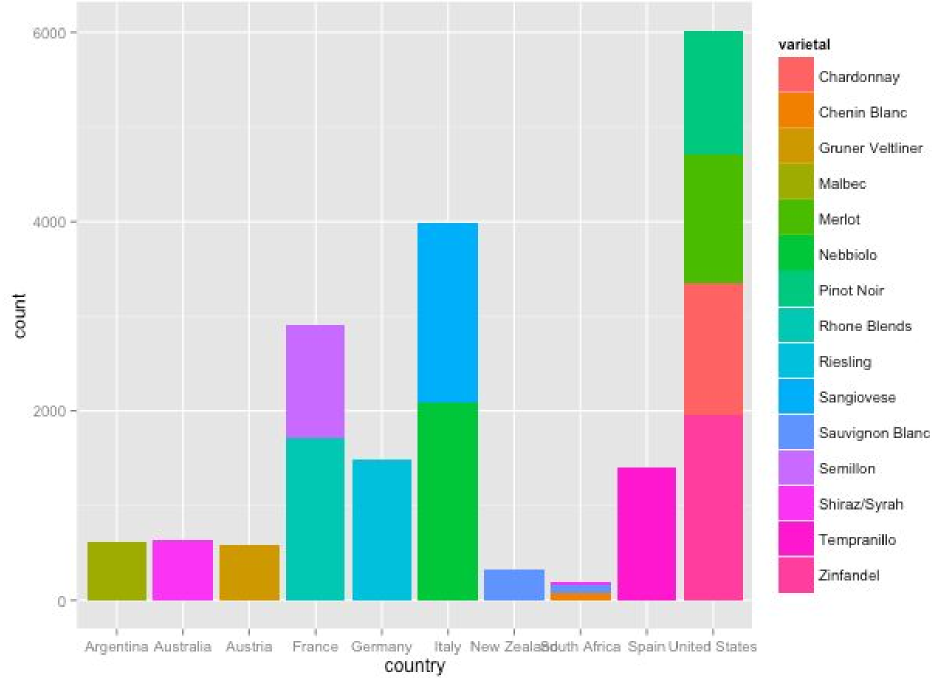
machine learning for language toolkit

LELAND STANFORD JUNIOR UNIVERSITY

DIE LUFT DER FREIHEIT WEHT

1891

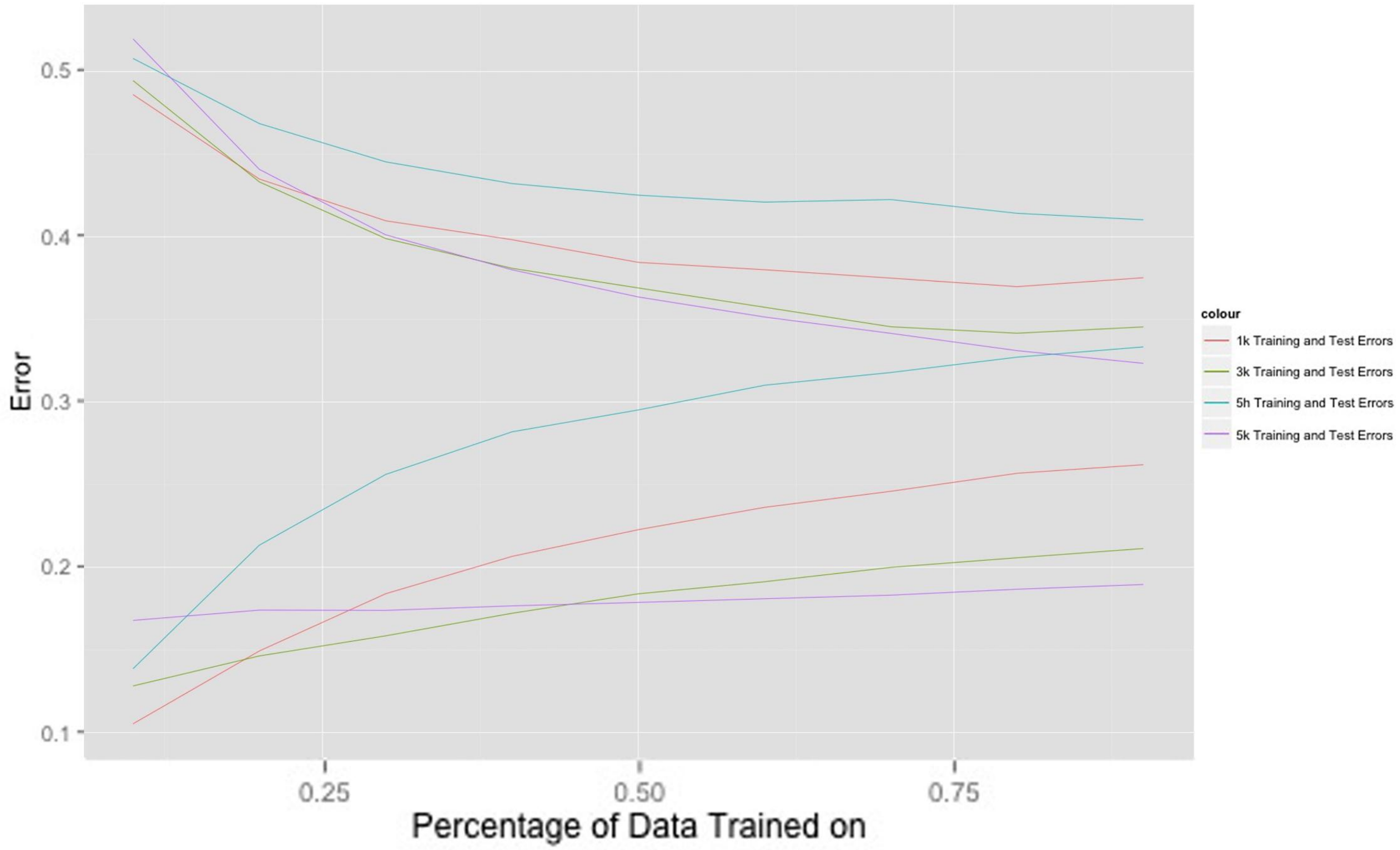
Wines with >50% Share per Country



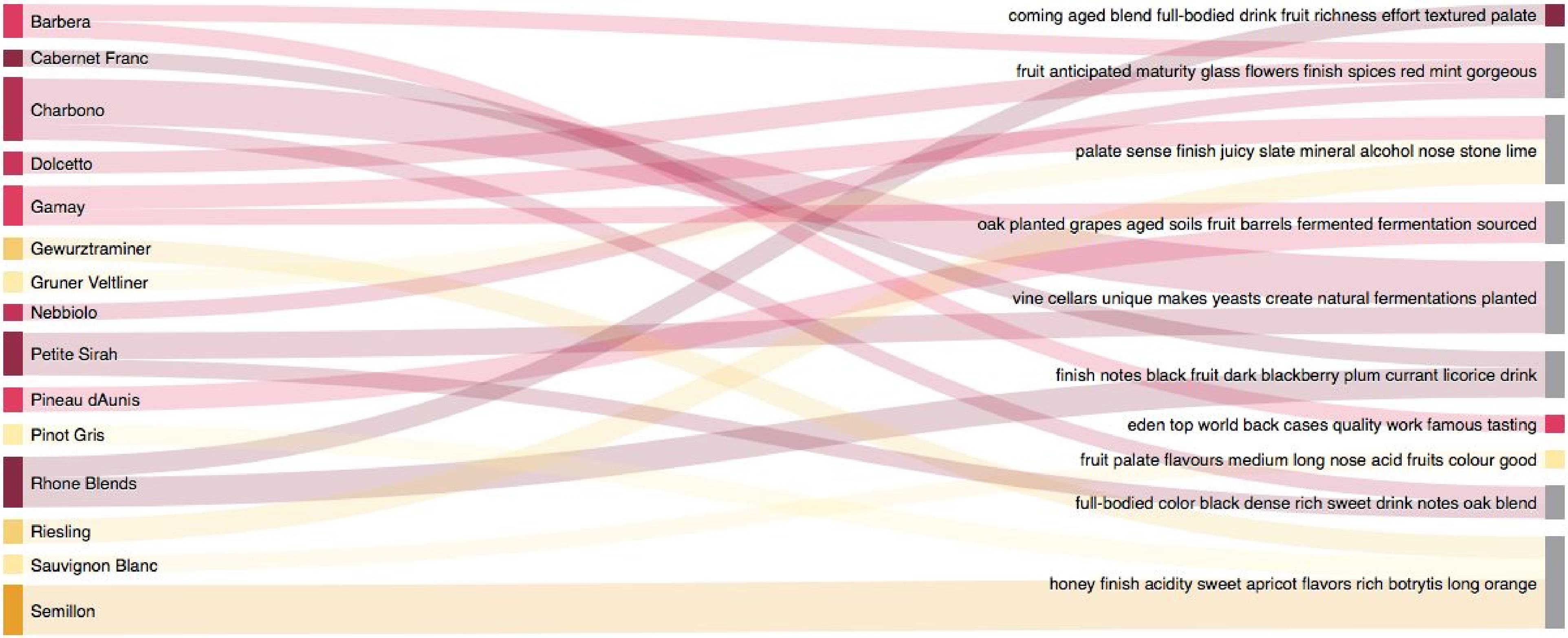
Models

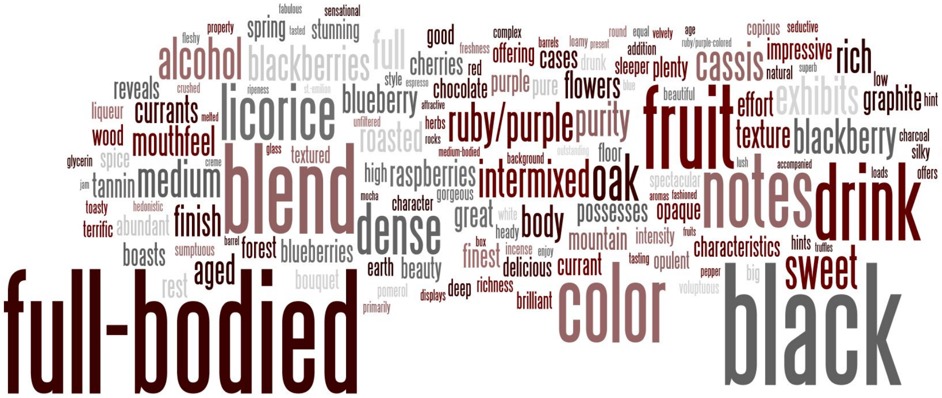
- Preliminary Naïve-Bayes model built with Vowpal Wabbit using all reviews for a given wine, and word counts as features
- Second Naïve bayes model built in MALLET by removing both generic and wine-specific stopwords
- Third Naïve bayes model built using bigrams from the 2nd model's input
- Unsupervised Topic model built using Latent Dirichlet Allocation in MALLET
- Final Naïve-Bayes models built by filtering based on the most relevant features in the output from the LDA topic model (5K, 3K, 1K, and 500 words), and stemming the input features

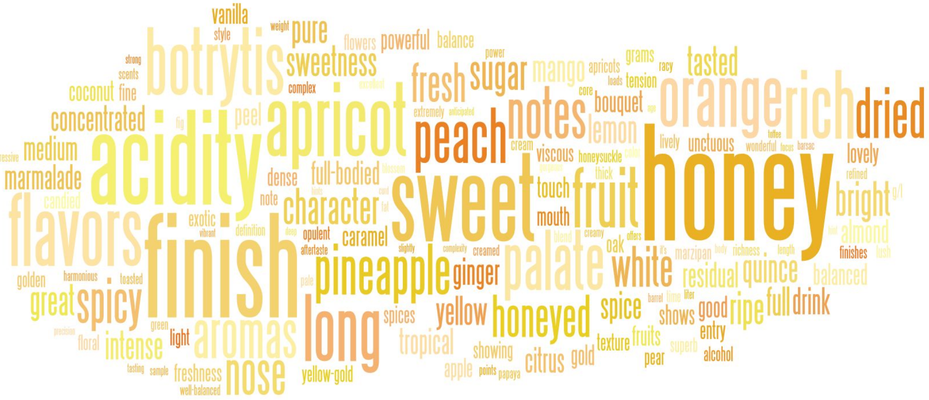
Learning Curves Across Models



Topic Analysis was used to filter out less important words for the final model







Confusion matrix for the final model

Using the top 5,000 words from the topic analysis

	LABEL	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	total
0	Riesling	34							3										4	2					43
1	Cabernet Franc			3							7		1			1	1				2				15
2	Cabernet Sauvignon and Blends		1	22						1	10					4		2			2				42
3	Chardonnay				46														6	1					53
4	Chenin Blanc	2			8	9													1	5					25
5	Gewurztraminer						2												1	1				1	5
6	Grenache			1				14			3		1			5	8				4	2		1	39
7	Gruner Veltliner	3			1				5										1			1			11
8	Malbec			1						9	4					1					1	2			18
9	Merlot		1	3							34					2		5			1				46
10	Mourvedre		1					2				5	1				3	3	1		2	1			19
11	Nebbiolo		1					2			2		49			1		15							70
12	Petite Sirah										2		1	9					1	1	5				19
13	Pinot Gris	1			1	1	2						1		3				1						10
14	Pinot Noir	1	1	1	1			1			3		6	1		70	3	1			3				92
15	Rhone Blends	2						1			4		2			6	52		1	1	6			1	76
16	Sangiovese			1							6		5			1		52				2			67
17	Sauvignon Blanc	1			6														38	4					49
18	Semillon				2						1								3	54					60
19	Shiraz/Syrah			1							2					5	3	2			51			2	66
20	Tempranillo			1							5					3	3	1			3	34		1	51
21	Viognier	1			3	1													1				6		12
22	Zinfandel			1							4			3		1	1	2			3			27	42

Conclusions

- Preprocessing data is critical to developing a high-quality model
    - Filtering out stop-words, narrowing the feature set to common wine-centric words, and stemming the words helped improve the Naïve-Bayes algorithm
- Learning curves are helpful in surfacing the flaws in our preliminary models
    - For example, we found that using bi-grams was not a good modeling approach as the training error remained very low, indicating the model was overfitting to the large feature set
- Our best model used Naïve-Bayes with 5,000 features and resulted in a testing accuracy of 69.0% across the 23 wine varietals (using 10-fold cross validation)
    - The top performer in our model was Chardonnay, with an 87.5% testing accuracy