Introduction

Have you ever found a recipe that you loved? Have you ever found a bunch more after that weren't so great? How can we learn from the recipes that we know are successes to prevent ourselves from finding failures? One way is to pull multiple successful recipe authors from the cooking section of the New York Times, determined to be successes by their reviews. The reviews were the ten randomly chosen authors from pescetarian-friendly recipes on the site, acquired by searching "vegetarian" on the main page. The raw dataset includes 192 recipes, varying in length, with titles, introduction summaries, review number, recipe steps and nutritional information. Added features include recipe number for tracking purposes and full text field for analyzing a combination of both the summary and recipe steps.

	Recipe_Num	Title	Author	Time	Summary	Steps	Reviews	Calories	Fat	Carbs	Protein	Text
0	0	Summer Squash Fritters With Garlic Dipping Sauce	KIM SEVERSON	85	David Venable, the most popular host on QVC, h...	PREPARE THE DIPPING SAUCE: Heat oven to 375 de...	155	253.0	22.0	9.0	3.0	David Venable, the most popular host on QVC, h...
1	1	Creamy Ramp Pesto Pasta	KIM SEVERSON	30	Ramps are one of those items that seem so appe...	Bring a large pot of water to a boil for pasta...	47	516.0	21.0	62.0	18.0	Ramps are one of those items that seem so appe...
2	2	Mushroom Risotto With Peas	MARTHA ROSE SHULMAN	50	If you are ever at a loss for what to make for...	Bring stock or broth to a simmer in a saucepan...	709	NaN	NaN	NaN	NaN	If you are ever at a loss for what to make for...
3	3	Mussel Risotto	MARTHA ROSE SHULMAN	45	I usually keep a good supply of arborio rice o...	Clean the mussels. Inspect each one carefully ...	23	603.0	9.0	93.0	27.0	I usually keep a good supply of arborio rice o...
4	4	Asparagus Frittata With Burrata and Herb Pesto	DAVID TANIS	30	Frittata, the savory Italian egg dish, can be ...	Rinse asparagus, and pat dry. Cut into 1-inch ...	275	373.0	33.0	3.0	15.0	Frittata, the savory Italian egg dish, can be ...

Figure 1. A depiction of the first five rows of the dataset including all raw fields and created fields.

Clustering

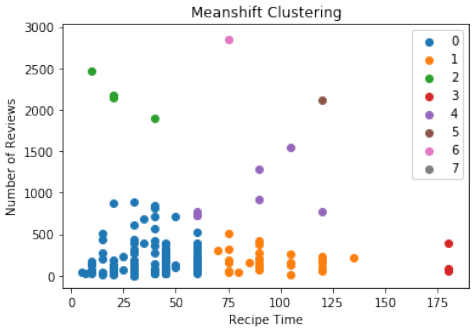

Figure 2. A depiction of the 7 clusters estimated and predicted by Meanshift clustering.

To start diving into the data, I looked at clustered formed by looking at the numerical data available in the raw data set. The first subset of data I clustered was recipe time and number of reviews. Using Meanshift clustering, 7 clusters were predicted. With 10 authors in the dataset I was under no illusion that my Meanshift clustering had perfectly identified each author by their recipe time and number of reviews alone, but I was interested to see the breakdown of each cluster and what I could learn from the model. By a visual inspection of the graph, it appears as though there are two main clusters and

five clusters of outliers. Upon closer inspection, the two groups are distinguished by a time to reviews ratio of 0.2 (left group) and ratio of 0.5 (right group). When comparing Meanshift clustering to KMeans, Spectral clustering, and Affinity Propagation, similar clusters were formed, splitting the authors evenly between them. Affinity Propagation performed the least like the others as it predicted 19 clusters. I moved forward with applying Meanshift to the data, because it both identified the two main groups that a majority of the data was split into and broken down the outliers, instead of grouping them together similar to Spectral clustering. While we can't learn much about who wrote the recipe from the time and the number of reviews alone, it is interesting to learn that a good majority of recipes take 30-90 minutes and a ratio to reviews of 0.2-0.6.

The second set of data that was clustered was the numerical data previously looked at including the nutritional information of calories, carbohydrates, fat, and protein of the recipes. Approximately 50 recipes had to be dropped due to missing nutritional information. Clustering was performed with Spectral Clustering with five clusters. About a third of the recipes are grouped into a cluster with the lowest time, calories, carbohydrates, fat, and protein. With that being said time did not play a large factor in the clustering as the time averages all fell within 45-70 minutes. The rest of the recipes are split evenly among the clusters with around 20 recipes per group. A second cluster has the highest calories, carbohydrates, fat, and protein. A third has a low-medium amount of calories and carbohydrates at 300 and 40 respectively with low protein and fat. A fourth has a medium amount of calories and fat, with a similarly low amount of carbs and protein. The final has a medium amount of calories, high carbohydrates, and equally low fat and protein.

Figure 3. A chart of the original Meanshift clustering applied to the test set of data.

The initial clustering was run on the test set of the data, the graph of which is shown in Figure 3. The test data fit right into the clustering predicted above, but with such a small subset of data in the test set there were not many outliers to fit into the clusters. While there are no outliers in the test data to test over-fitting occurring there, it does not appear there was any over-fitting in a majority of the data, split between clusters 0 and 1.

Analyzing Text Data

Being able to organize our data into clusters based on their nutritional information or their time to reviews ratio can be informative, however we are ultimately looking to predict the author of a recipe assuming that authors with multiple well-reviewed recipes continue to produce great recipes. To accomplish this, we must look beyond the given numerical data and dive into the text itself. For this modeling, we will exclusively be using the created feature "Text" column of the dataset that includes both the introduction summary and the recipe steps.

I first broke down the data text into sentences giving a dataset of approximately 4000 data points and applied a bag of word function to those sentences, grouping together the 20 most common words in each sentence. Using Random Forest Classifier with 50 estimators, the correct author could be correctly predicted an average 54% of the time,

which is a good model for our data that has a chance of randomly guessing the author at 10%.

Next, I took an unsupervised approach to the data learnings and applied tf-idf vectorization to the sentence breakdown of the text. In this model, Logistic Regression was applied to the model and an average of 33% of the time the author was correctly predicted based on the sentence. To be able to visualize this data, it was reduced to two components with PCA. With the text breakdown, there are significantly less outliers and a majority of the text broken into three clusters. The smaller outlier cluster with the x-axis greater than 0.60 are sentences about heating the oven to a temperature in the 300s and the small outlier cluster with the y-axis greater than 0.38 are sentences about seasoning with salt and/or pepper. The main three clusters deal with heating, chilling, and mixing ingredients together.

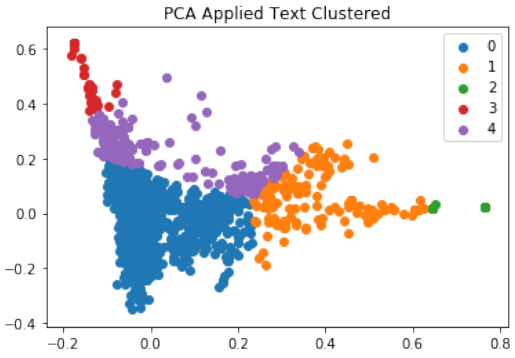

Figure 4. A chart of the clustered text data after PCA has been applied.

Finally, the above models were applied on the testing set of data to test their consistency and identify any overfitting. It can be seen in Figure 5 that both models have a bit of overfitting seen by their significant increase in performance of the train set over the test set. Furthermore, it appears as though bag of words performs better overall.

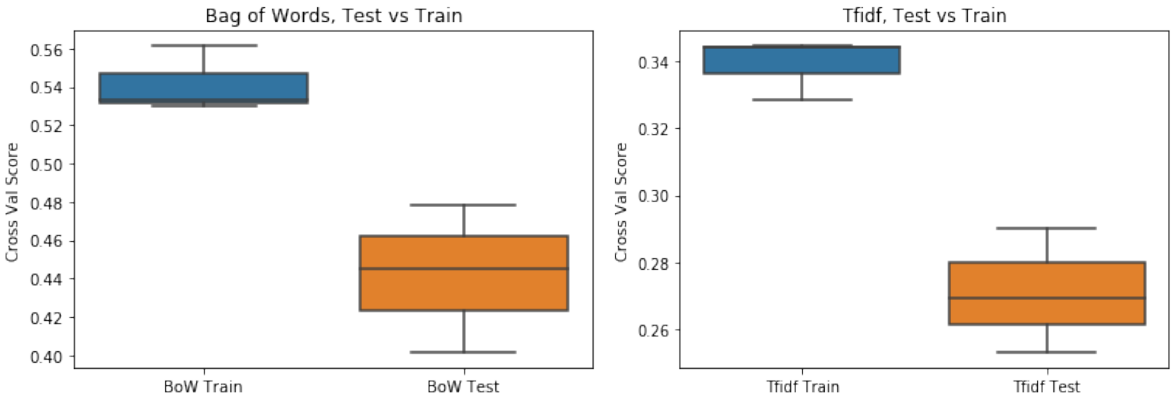

Figure 5. Boxplots of train and test sets comparing the supervised bag of words model and the unsupervised tfidf model.