**An Examination of Recipes from Around the World**

By: Ben Sturm

# Project Summary

**Project Design**

For this project, I examined recipes from around the world through the lens of a data scientist. I was really interested to see if I could learn something about the relationships of different cuisines throughout the world. In order to explore this topic, I chose to use recipe data. In particular, I used the list of ingredients for ~12,500 different recipes and ran different machine learning models which I will discuss further in this summary.

**Tools**

Since this project was focused on text data, I used many of the typical natural language processing tools in order to feed my data into a machine learning model. Such tools included tokenization and bag-of-words processing using sklearn’s CountVectorizer. Removing words that occur very frequently using a combination of tf-idf analysis as well as my own common sense. I also did stemming to remove plural forms of ingredients by representing words with their corresponding stem. After preprocessing the data, I could run some machine learning models. In particular, I ran a number of unsupervised learning algorithms including k-Means Clustering, Principal Component Analysis (PCA), and Latent Dirichlet Allocation (LDA).

**Data**

The recipe data I used for this project came from the Yummly.com site. I was granted a student license to Yummly’s API, so I was able to do queries to search for recipes directly from my ipython notebook. Yummly supports doing searches based on cuisine type. The following are the supported list of cuisines:

“American, Italian, Asian, Mexican, Southern & Soul Food, French, Southwestern, Barbecue, Indian, Chinese, Cajun & Creole, English, Mediterranean, Greek, Spanish, German, Thai, Moroccan, Irish, Japanese, Cuban, Hawaiin, Swedish, Hungarian, Portugese”

In total, I downloaded approximately 500 recipes for each of the 25 cuisines supported. This lead to ~12,500 different recipes. A few lines from the dataframe I used for my analysis is displayed in Figure 1.

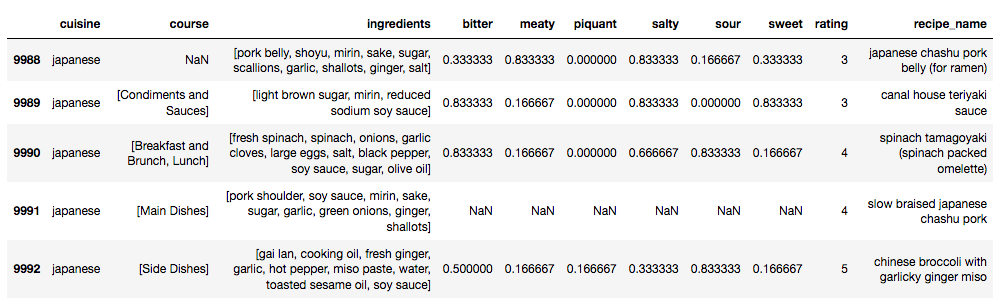


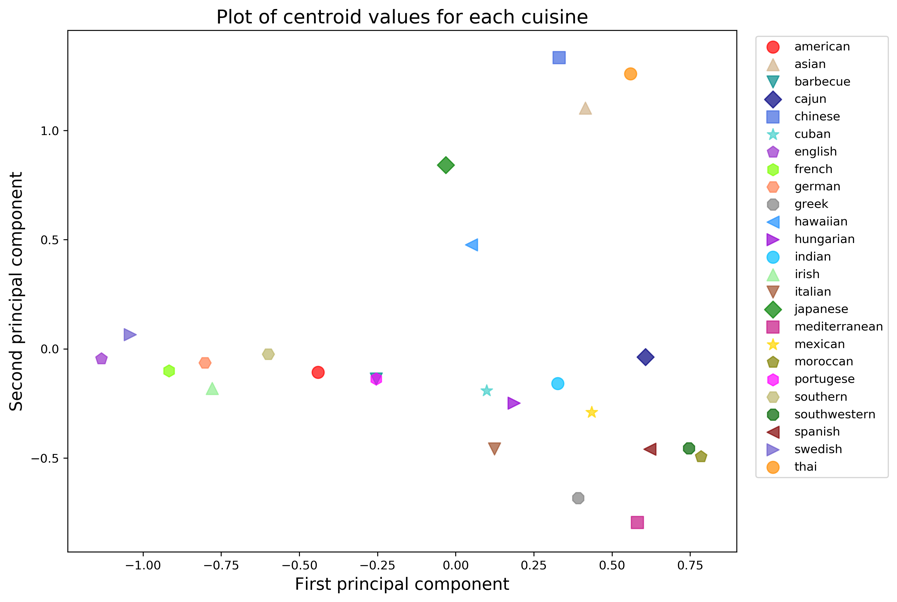
Figure 1. The dataframe containing the Yummly recipe data. I focused on the ingredients column for all of my unsupervised learning models.

**Algorithms**

The algorithms I focused on were all unsupervised learning algorithms. I did k-Means Clustering to see if I could cluster recipes together based on the cuisine type, however clustering wasn’t super helpful for my analysis, because it was unclear what the different clusters represented. Instead, I focused my attention on PCA analysis as well as LDA which I will discuss further in the Results section.

**Results**

Applying PCA to the recipe data was really insightful. Originally, the recipe dataframe had a size of (12492, 1985). This corresponds to the 12,492 recipes and 1985 ingredients. After PCA, the dataframe was of size (12492, 2). When plotting all of the recipes in this 2-D principal component space, I didn’t learn a lot, because many of the data points were overlapping, so it was difficult to see any structure in the data. However, by grouping the recipes based on the cuisine and taking the centroid values along



**E**

**D**

**C**

**B**

**A**

Figure 2. A plot of the centroid values for each of the different cuisines along the first and second principal components. Group (A) is associated with Asian cuisines, (B) consists of Japanese and Hawaiian cuisines, (C) and (D) are European and American cuisines, respectively. Group (E) is a mixed bag of cuisines from all over the world including Cuban, Mexican, Indian, and Spanish.

the first two principal components, I could see some interesting structure in the data. A plot of this is shown in Figure 2. We can observe that the centroid values tended to group the recipes based on similar cuisines. For instance, group (A) in Figure 2 consists of Chinese, Thai, and Asian, which could all be classified as Asian foods. I found groups (B) and (E) to be particularly interesting, because group (B) consisted of Japanese and Hawaiian cuisines. Both of these cuisines place a strong emphasis on fish, so it makes sense that they are closely grouped together. Group (E) was also interesting, because it was a mixed bag of many different cuisines from all over the world. This included Cuban, Mexican, Indian, Spanish, and Southwestern. When I think of these cuisines, I think of big, bold flavors, so it makes perfect sense that these cuisines would be closely grouped together.

**Things I’d do differently next time**