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 several 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.

	cuisine	course	ingredients	bitter	meaty	piquant	salty	sour	sweet	rating	recipe_name
9988	japanese	NaN	[pork belly, shoyu, mirin, sake, sugar, scallions, garlic, shallots, ginger, salt]	0.333333	0.833333	0.000000	0.833333	0.166667	0.333333	3	japanese chashu pork belly (for ramen)
9989	japanese	[Condiments and Sauces]	[light brown sugar, mirin, reduced sodium soy sauce]	0.833333	0.166667	0.000000	0.833333	0.000000	0.833333	3	canal house teriyaki sauce
9990	japanese	[Breakfast and Brunch, Lunch]	[fresh spinach, spinach, onions, garlic cloves, large eggs, salt, black pepper, soy sauce, sugar, olive oil]	0.833333	0.166667	0.000000	0.666667	0.833333	0.166667	4	spinach tamagoyaki (spinach packed omelette)
9991	japanese	[Main Dishes]	[pork shoulder, soy sauce, mirin, sake, sugar, garlic, green onions, ginger, shallots]	NaN	NaN	NaN	NaN	NaN	NaN	4	slow braised japanese chashu pork
9992	japanese	[Side Dishes]	[gai lan, cooking oil, fresh ginger, garlic, hot pepper, miso paste, water, toasted sesame oil, soy sauce]	0.500000	0.166667	0.166667	0.333333	0.833333	0.166667	5	chinese broccoli with garlicky ginger miso

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

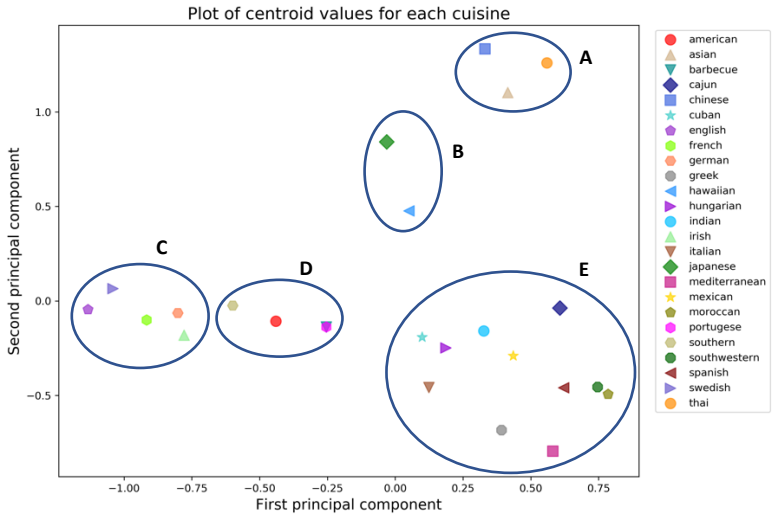

Figure 2. A plot of the centroid values for each of the different cuisines along the first and second principal components. Group (A) is associated with Asian cuisines, (B) consists of Japanese and Hawaiian cuisines, (C) and (D) are European and American cuisines, respectively. Group (E) is a mixed bag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

A question the reader may ask is, which features (e.g., ingredients) are most strongly linked to the first and second principal components? This can be visualized in Figure 3. This plot provides a visual representation of the dominate features for each of the two principal components. Ingredients such as chicken, garlic, onion, and tomato have strong associations along the positive direction of component one. These flavors have strong ties to cuisines such as Spanish or Indian. On the other hand eggs, butter, flour, milk and sugar have strong associations along the negative direction of component one. These are ingredients found typically in French or English dishes. Likewise, soy, sauce, and rice have strong associates with the positive direction of component two. These ingredients are common in Asian cuisines. Finally, cheese, lemon, olive oil and tomato have strong associations with the negative direction of component two. These flavors are very common in Italian and Greek cuisines. Figure 3 helps to explain the structure of Figure 2, namely why certain cuisines were clustered together in particular regions when plotted along the first and second principal components.

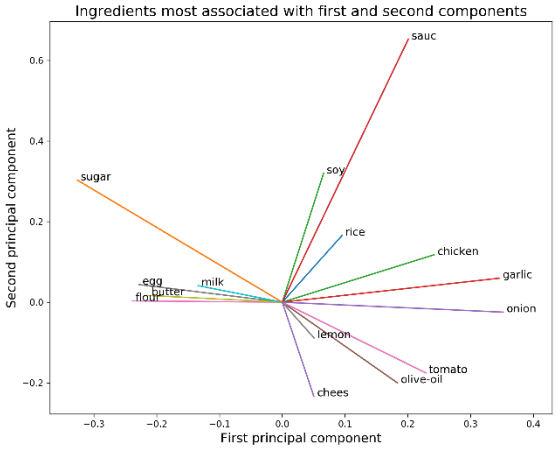

Figure 3. A plot demonstrating the ingredients most strongly associated with the first and second principal components.

Finally, I also ran an LDA model in order to do topic modeling. I was curious if it would be possible to separate out the different ingredients based on the cuisine that they typically belong to. I specified the number of topics to be 25, because I knew there were 25 different cuisines represented in my dataset. The results of LDA were a bit messy, however. In certain cases, the LDA topics were particular cuisines such as Italian or Thai. However, some of the topics were different categories of dishes such as desserts, sauces, or even cocktails. Although this result was not what I had intended, in retrospect it actually makes perfect sense. LDA is a machine learning technique that identifies groups of words that appear together frequently. So, in a corpus of over 12,000 recipes, there might be a stronger association of groups of words based on the type of dish (i.e. dessert, soup, salad, or sauce) versus the type of cuisine.

Things I'd do differently next time

I was actually quite pleased with the results of my project, so there is not a lot I would do differently next time. However, one thing in particular that probably would have improved my LDA results would have been to have more data and then filter out the recipes based on "Main Dishes" or "Lunch". That would have helped to mitigate the issues of grouping the words together based on the type of dish. If I had more time, I also would have tried out some supervised learning models to see if I could train a model to correctly categorize the cuisine based on ingredients.