What Makes a Cuisine Unique?

Sunaya Shivakumar
sshivak2@illinois.edu

ABSTRACT

There are many different national and cultural cuisines from around the world, but what makes each of them unique? We try to answer that question by making use of multinomial logistic regression model to learn and predict unique cuisine styles. Furthermore, we attempt to identify fusion-cuisines that are borne of two or more distinct cuisine styles, using k-means clustering. Results demonstrate that cuisines are too diverse to predict with very high accuracies – which may have led to the creation of fusion-cuisines.

Keywords

Machine learning; text mining; cuisines; classification; logistic regression; k-means.

INTRODUCTION

We humans have always been very creative with the food that we eat – combining many distinct ingredients in ingenious combinations to create unique dishes and recipes. Historically speaking many food ingredients were native to only specific regions of the world. Recipes that make frequent use of these region-specific ingredients have come to characterize different national, regional, and cultural cuisine styles. For example, the Indian cuisine style recipes make heavy use of ingredients like *cumin*, *coriander*, and *turmeric*, while *olive oil*, *parmesan cheese*, and *basil* are more characteristic of the Italian cuisine.

There is an obvious relation between ingredients and cuisines – one that can be used to predict cuisines. However, recipes – which are unique combinations of these ingredients can be more useful in identifying a cuisine style. One goal of this project is to train a classification model that can predict and identify the cuisine style of any given recipe.

As we explore our dataset in the next section, we learn that some sets of ingredients are used in more than one cuisine type – an observation from which we can guess that these similar cuisines can be easily merged together to create fusion-cuisines. We try to solidify this train of thought by using k-means to cluster our dataset of recipes.

After inspecting the results section, we conclude this project with a discussion of future work and applications of classifying cuisines and fusion-cuisines.

DATASET

For the purposes of this project, we make use of the raw recipe data that is publicly available as part of recipe collection by Ahn *et al., 2011*[1]. The dataset used was obtained by crawling *epicurious.com*[2], a large, digital food platform, and comprises of 13408 recipes with 350 unique ingredients, and 26 different cuisines. Each recipe is represented as a cuisine style and a list of ingredients that it comprises of.

Preliminary analysis of this dataset shows us the distribution of cuisines and recipe ingredients. We can understand the cuisine distribution of the dataset, from Figure 1 – a bar plot of recipes per cuisine in the dataset, and the ingredient distribution across the recipe dataset, from Figure 2 – a plot of the top 15 ingredients that occur in the dataset we are using.

We find that the American cuisine clearly dominates the dataset with 4988 recipes, followed by Italian cuisine with 1715 recipes, and Asian cuisine with 1176 recipes. Examining the frequency distribution of the ingredients we find that there are some ingredients like *garlic*, *butter*, and *egg*, that are very ubiquitous in nature and occur in many recipes, and across cuisines, as evident in Table 1.

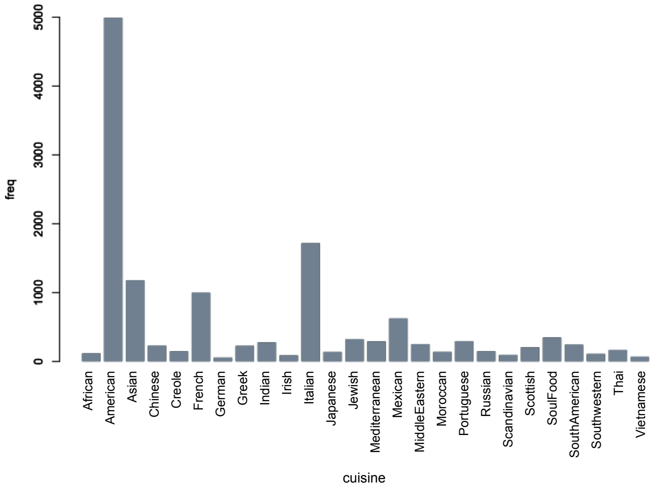

Figure 1. Bar plot showing the cuisine frequency distribution of the recipe dataset – American recipes make up for nearly 40% of the dataset.

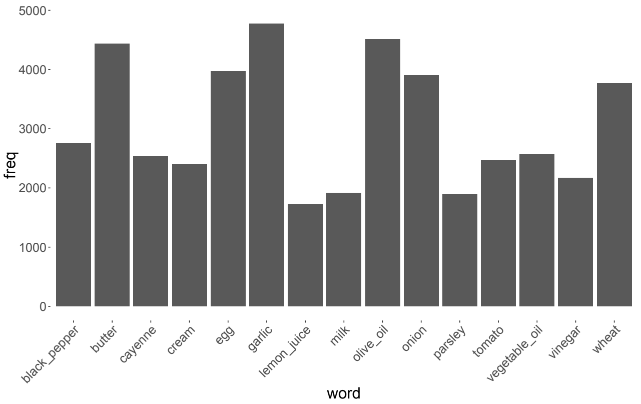

Figure 2. Top 15 ingredients and their frequency distribution in the recipe dataset – note how some ingredients dominate the ingredient space of the dataset.

Cuisine	Top 7 Ingredients
American	butter, egg, wheat, olive oil, garlic, onion, cream
Italian	olive oil, garlic, tomato, parmesan cheese, onion, butter, egg
Asian	soy sauce, ginger, garlic, rice, onion, vegetable oil, cayenne
French	butter, egg, wheat, olive oil, garlic, cream, onion
Mexican	cayenne, onion, garlic, tomato, cilantro, corn, olive oil.

Table 1. Top ingredients of the top 5 cuisines in the dataset – ubiquitous ingredients are frequently used in many cuisines.

To overcome the dominating effects of these commonly occurring ingredients, and to determine the ingredients that distinguish the different cuisines, we calculate the term-frequency – inverse-document-frequency (tf-idf)[2] weights of individual ingredients. Tf-idf weights of ingredients reflect how important or unique an ingredient is to a single recipe in a corpus of recipes, by calculating relative frequency of an ingredient in recipes of a particular cuisine in contrast to its frequency in recipes of other cuisines. This weighting model will prove to be relatively more helpful in predicting the cuisine style of a recipe.

$\mathrm{tf}_{(i,\,r)}$ = number of occurrences of ingredient i in a recipe r

$$\mathrm{idf}_{(i)} = log_2\left(\frac{total\ number\ of\ recipes}{number\ of\ recipes\ where\ ingredient\ i\ appears}\right)$$

term frequency – inverse document frequency = $\mathrm{tf}_{(i,\,r)} \cdot \mathrm{idf}_{(i)}$

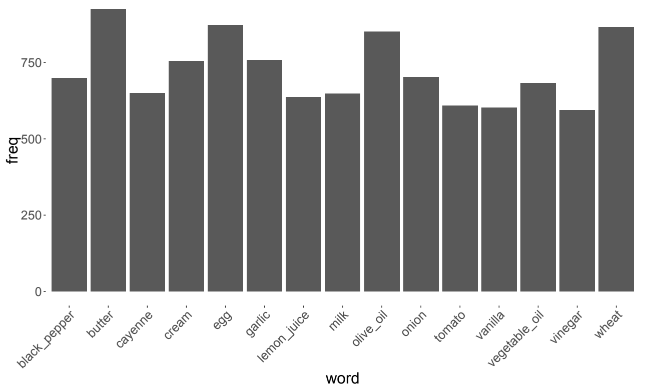

Figure 3. Top 15 ingredients and their frequency distribution in the recipe dataset, after calculating their tf-idf weights – this reduces existing bias towards any single ingredient.

In order to gain some insight to help us cluster the dataset, we try to visualize the structure of the recipe-ingredient space in our recipe corpus by using t-Distributed Stochastic Neighbor Embedding[3]. t-SNE makes use of dimensionality reduction and can be very helpful when trying to visualize high-dimensional data. Figure 4 – is a scatterplot mapping of recipe-ingredient space and cuisines in 2 dimensions.

As evident from the figure, American, Italian, French, and Mexican cuisines cover a wide variety of the ingredient space, and also overlap with other cuisines; Asian cuisines like Chinese, Japanese, Indian, Thai are seen in distinct groups that occur close to each other. These observations give us a clue as to what we can expect when we perform k-means clustering of our dataset.

PREDICTION TASK

For this task, we aim to predict the cuisine style of a recipe, given a list of ingredients used.

Methods

The data is first converted to a document-term matrix, where each document is a recipe and each term is a distinct ingredient. Each row of this matrix represents a recipe and and each column - an ingredient, with values of 1 or 0, to indicate its presence or absence. We also calculate the tf-idf scores of ingredients in each recipe to create a new matrix.

To remove any existing bias, we then randomly shuffle these matrices by rows. The shuffled data is then divided into a training set (80%), and a validation set (20%). We use a multinomial logistic regression model[4] to classify our 26 classes/cuisines data, and to evaluate the performance of the model we measure the accuracy of the predictions made with our validation dataset.

Next, we test our model on feature vectors of each recipe in our validation set, and then determine the prediction accuracies.

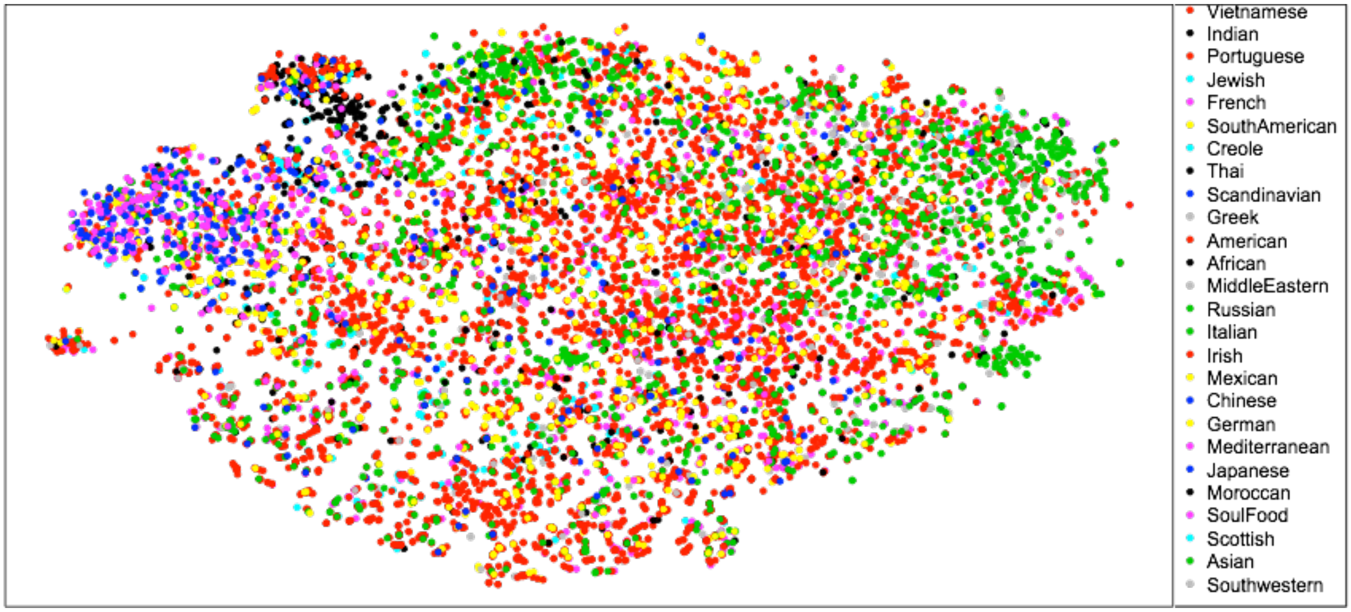

Figure 4. Visualizing recipe-ingredient space – a scatterplot mapping of recipes and cuisines in a 2-dimensional space using t-SNE. Each dot represents a single recipe and the different colors represent different cuisines.

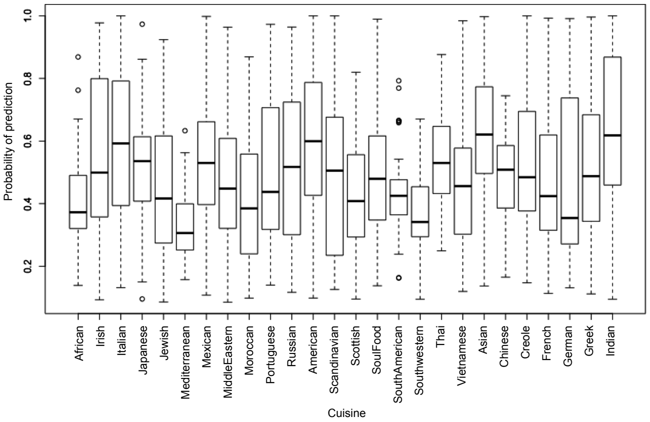

Figure 5. Box plot of the accuracies with which the cuisine type of each recipe was predicted – on average, Indian and Asian recipes are the most accurately identified cuisines.

Logistic Regression Model	Accuracy (Validation Set)
No tf-idf	0.5706
With tf-idf	0.5878

Table 2. Performance accuracies of our multinomial logistic regression model for 26 classes (cuisines).

Results

The performance of our classification model, is recorded in Table 2. We can observe that the model performs better when trained on the recipe-ingredient matrix with tf-idf weights, but only slightly.

To better understand the cuisine prediction performance of our model, we plot the accuracies with which our logistic regression model predicts the cuisine type of each recipe in the validation set – Figure 5. We can observe that Indian and Asian recipes were the most correctly identified cuisines, on average – which can be explained by the vast number of these cuisines' recipes in the dataset, and how distinct Asian and Indian food ingredients can be, when compared to the rest.

CLUSTERING TASK

The objective of this task is to identify and understand cuisine and fusion-cuisine clusters.

Methods

We use the same recipe-ingredient matrix that was computed for the previous task. However, we do strip away the cuisine labels associated with each of the recipes. Then we cluster our data using the k-means clustering algorithm that uses Euclidean distances between the feature vectors of every recipe in our data.

Having prior knowledge that we have 26 different cuisine types in our dataset, a naïve-way to cluster data would be to run k-means using 26 centers. We then do a frequency analysis of every cluster to obtain information about top common ingredients used. Mapping the recipes in each cluster back to our original data, we can also assess the top cuisines and/or fusion-cuisines that can be seen in a cluster of recipes.

Subsequently, we try to determine what the ideal number of clusters should be, by using the Elbow method – Figure 6.

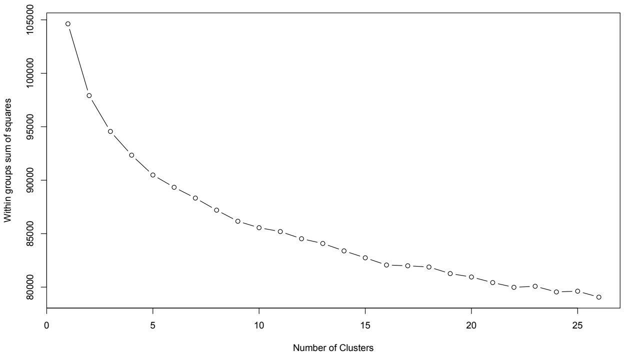

Figure 6. Elbow plot to determine the ideal number of clusters for our dataset – "elbow" occurs when number of clusters = 7

Cuisine distribution	Top ingredients
American (64%) French (8%) Italian (5%)	Egg, wheat, butter, milk, cream, vanilla, cinnamon, cane molasses, milk fat, lemon juice, cocoa, vegetable oil, nutmeg, ginger, lard
Asian (44%) American (14%) Indian (10%)	Garlic, ginger, vegetable oil, cayenne, soy sauce, rice, scallion, vinegar, pepper, cilantro, coriander, onion, sesame oil, fish
Mexican (26%) American (26%) SouthAmerican (8%)	Cayenne, onion, garlic, tomato, cilantro, cumin, olive oil, corn, bell pepper, lime juice, pepper, oregano, vinegar, scallion, beef

Table 3. A sample of 3 clusters from the first run of k-means clustering using 26 centers – note the distinctiveness of the ingredient space.

We can notice a slight "elbow" in the plot, indicating that k-means clustering using 7 clusters will give us the best results. Therefore, we run k-means using 7 centers to cluster our data.

Results

A sample clusters are tabulated in Tables 3 and 4.

The first run of k-means yielded 26 clusters that were mostly groupings of a very high number of recipes from a single cuisine – American, Asian, Italian, etc. There were very few clusters that had nearly equal distributions of recipes from two cuisines – Mexican-American, etc. This clustering behavior can tell us the most recipes in a single cuisine make heavy use of the same ingredients, maybe in different combinations and measures. We can also observe how some clusters are very sweet, some savory, and some spicy – telling us that cuisines around the world use the same basic ingredients to create meals and desserts

There were 7 clusters as a result of the second run of k-means clustering. These clusters had more equal distributions of the top two cuisines – which we take as an indication of a possible fusion cuisine style. Of the 7 clusters, 4 of them had distributions like, American-Asian, American-French, American- Italian, American-Mexican – which is a strong indication of how diverse American cuisine can be, and how it has been influenced by other major cuisine styles. This can also be confirmed – our dataset of recipes from *Epicurious* has an American-centric view of world cuisines, as these recipes are collected from American publications like *Bon Appétit*. The other 3 clusters had interesting and exotic distributions like, Mexican-Italian, Asian-Chinese, Japanese-Mediterranean. These were great examples of how well two distinct cuisines can work together to create appealing fusion-cuisines.

Cuisine distribution	Top ingredients	Sample recipes in cluster
Mexican (36%) Italian (31%)	Garlic, olive oil, onion, tomato, cayenne, wheat, egg, butter, vegetable oil, cilantro, black pepper, corn, parsley, cream, milk	[1] "tomato olive_oil lemon cayenne garlic bell_pepper olive" [2] "olive_oil wheat cheese corn cayenne oregano" [3] "coriander tomato shallot avocado lime_juice garlic"
American (32%) Italian (24%)	Garlic, olive oil, onion, tomato, black pepper, vinegar, parsley, butter, wheat, egg, bread, chicken, basil, pepper, beef, thyme, cheese	[1] "olive_oil wheat yeast fish bell_pepper oregano" [2] "butter cheese goat_cheese macaroni black_pepper" [3] "tomato olive_oil onion chicken_broth garlic bread"
Japanese (43%) Mediterranean (40%)	Olive oil, garlic, soy sauce, onion, rice, vinegar, vegetable oil, scallion, tomato, wine, ginger, wheat, bell pepper, egg, lemon juice, parsley, fish, cayenne, sake, barley, honey, beef, potato	[1] "beef sake soy_sauce scallion chive vegetable_oil wine" [2] "kohlrabi mandarin_peel olive_oil pepper sesame_seed potato pea wine" [3] "olive_oil onion vinegar lamb red_wine fennel garlic lemon wheat"

Table 4. A sample of 3 clusters from the second run of k-means clustering using 7 centers.

FUTURE WORK AND APPLICATIONS

We classify recipes by cuisine style, using a multinomial logistic regression model. It would be helpful to run this classification experiment with different models to find the best one. Classifying recipes can be used to learn more about the cuisine style of a dish, just by looking at the ingredients used, or even by looking at a picture of the dish. Recipe and restaurant suggestions can be made by using user preferences observed over time.

The data set that we use for this project lists recipes as a set of ingredients accompanied by a cuisine label. This dataset can be enhanced by adding another feature to the recipes – the name of the recipe. Equipped with this additional information, we can suggest recipes, based on user input that is set of ingredients. This can be very useful when trying to decide what to cook with given a restricted set of ingredients, or even trying to decide what to eat a restaurant.

Topic modelling can be used to learn generative models of recipe and ingredient distributions, to later produce new recipes to try, or even recipes similar to existing ones, so that users can enjoy a new dish that has their favorite ingredients.

CONCLUSION

In this project we use machine learning approaches to classify recipes into distinct cuisines. We also demonstrate initial attempts that aim to understand the recipe-ingredient space and learn more about fusion-cuisines.

Multinomial Logistic Regression classifier is used to classify a recipe or a list of ingredients into a cuisine type, with fairly good accuracies. This is followed by k-means clustering and analysis to learn more about fusion-cuisines. Results from this approach show some promise – clusters containing nearly even distributions of two different cuisines inform us about possible fusion-cuisine styles. Future work on this subject can show us if this can be observed better with vast datasets.

REFERENCES

1. Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási. Flavor network and the principles of food pairing. *Scientific Reports*, 1(196), 2011. http://www.nature.com/articles/srep00196
2. http://www.epicurious.com/
3. https://cran.r-project.org/web/packages/tm/tm.pdf
4. L. J. P. van der Maaten and G. E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579 – 2605, Nov 2008. http://lvdmaaten.github.io/tsne/
5. https://cran.r-project.org/web/packages/maxent/maxent.pdf