What’s Cooking: Cuisine Type Prediction

Guojing Wu[1] (gw2383)

[1]*Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA.*

Abstract

Nowadays, we often came into this question: what’s to cook tonight? Eating out costs too much and in the meantime, we have tons of ingredients in the kitchen, so we end up struggling about how to make a potential delicious combination out of them. Based on this simple but not easy question, we built up a recommendation system. We first run basic exploratory analysis on twelve thousand recipes with 20 cuisine types and thousands of ingredients. Then we applied several machine learning feature extraction methods to extract feature vectors from those recipes. Finally, we employed multiple classification methods including Linear Support Vector Machine, Logistic Regression, Decision Tree, and Random Forest to give out predicted cuisine type. In the end, we can recommend people what’s to cook based on what they have.

Keywords: cooking recipe; feature extraction; machine learning; SVM; logistic regression; decision tree; random forest.

Introduction

Nowadays, most of the people have spent some money on restaurants, and food delivery gradually blends into everyone’s life. Plenty of studies have shown that the percentage of eating outside had climbed approximately 50 percent from 1900 to 2010 [1]. Besides, fewer families could sit down together for a meal [2]. While eating outside gradually becomes a habit, this long-term bad eating habit could lead to lots of problems.

Why cooking at home? The top and most apparent reason is eating at home is healthier than dining in restaurants. But there are more potential problems. Research has shown that a family which regularly eating at home could have healthier and happier kids and teenagers who are less likely to use alcohol, drugs, or cigarettes [3]. For adults, eating at home on a regular basis could result in higher energy levels and better mental health [4]. Some researches show that it could even help people live longer.

The above benefits of cooking at home are still not enough. Researches have shown that home-cooked meals can benefit the environment [5]. As an example, if you cook at home, you could buy the ingredients that you need directly from local farmers, which in a way cutting down on packaging and reducing the transportation that required to get food to your plate.

Today, most of the people think cooking at home is inconvenient, and they usually have no idea of what's to cook even their refrigerators are full of ingredients. In order to encourage people to cook more at home and let people find that cooking could be easy. We built up a recommendation system based on the dataset posted on Kaggle. We first tokenize the input ingredients, then use feature extraction methods to extract feature vectors out of it. Finally, we trained several classification models for predicting cuisine type.

Data and Methods

Data. The original dataset is from Kaggle, with 39774 recipes in the training dataset and 9944 recipes in the testing dataset. The following are the column fields in the training dataset, testing dataset contains only id and ingredients column.

1. id: unique recipes ID for each recipe
2. cuisine: different 20 cuisine types which are Brazilian, British, cajun_creole, Chinese, Filipino, French, Greek, Indian, Irish, Italian, Jamaican, Japanese, Korean, Mexican, Moroccan, Russian, Southern_us, Spanish, Thai, and Vietnamese
3. ingredients: each recipe has its unique ingredients. The original data has ingredients number with the smallest number 1 to the largest number 110, and contains 6714 unique ingredients among all.

Feature Extraction. Since our ingredients for each recipe and cuisine is in a text form, which put all ingredients together as a list, so firstly, we use Tokenizer-a process of taking text and breaking it into individual terms-to break our list ingredients into individual words. Then we tried multiple feature extractors methods, and we found Word2Vec stands out from others and then the Term Frequency-Inverse Document Frequency (TF-IDF) performed as the second. The vector size of Word2Vec to set is 100 (*vecsize=100*) and feature number of TF-IDF is set to 20 (*numFeatures=20*).

Support Vector Machine. We tried linear support vector machines, but the SVM model in Pyspark cannot handle multiple class classifications, so we employed one-vs-rest strategy here to reduce multiclass classification to binary classification. For a multiclass classification with k classes, it trains k models (one per class). Each example is scored against all k models and the model with highest score is picked to label the example. The parameters we set in the models are maximum iteration equals to 1000 (*maxiter=1000*), convergence tolerance equal to 0.001 (*tol=0.001*), the aggregation depth equals to 3 (*aggregationDepth=3*), and the regularized parameter sets to 0.01 (*regParam=0.01*).

Logistic Regression. For the logistic model, we use multinomial distribution to handle multiclass classification. We set maximum iteration to 1000 (*maxiter=1000*), regularized parameter number equals to 0.01 (*regParame=0.01*).

Decision Tree. We set maximum depth to 30 (*maxdepth=30*).

Random Forest. We set maximum depth to 15 (*maxdepth=15*).

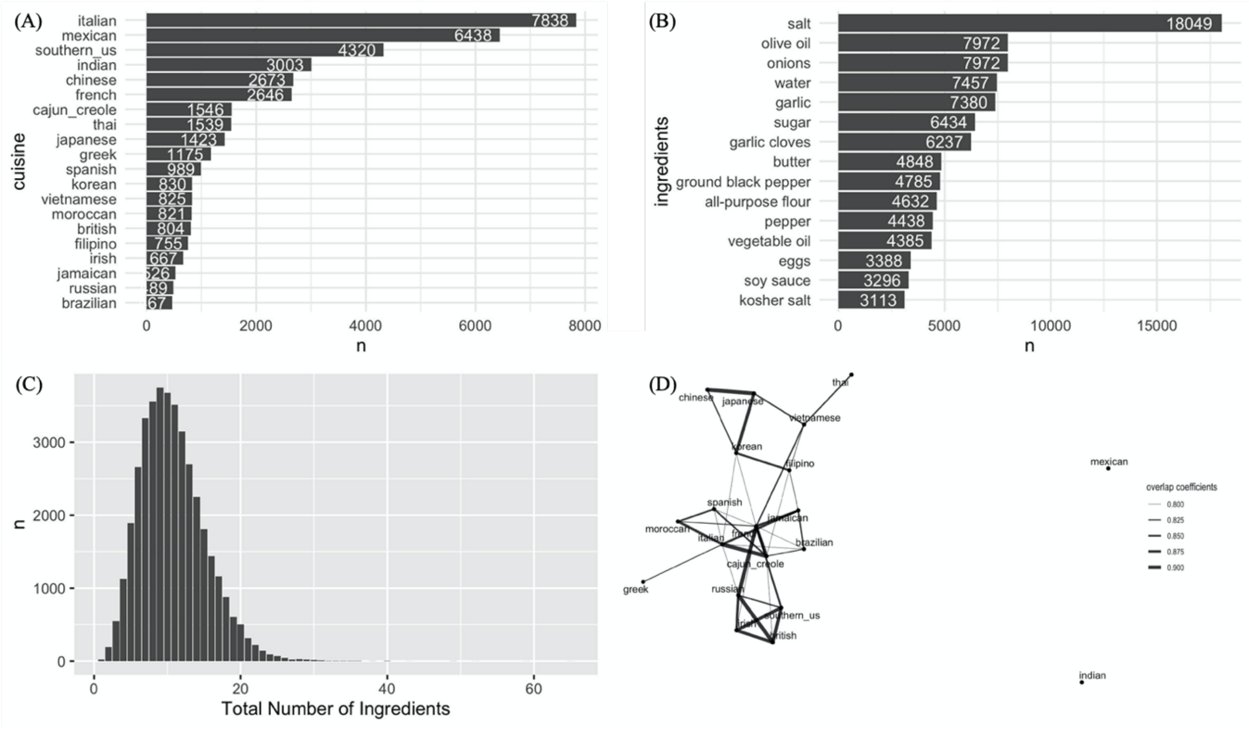

Figure 1. EDA result. (A) Total number of recipes for each cuisine (B) Most common ingredients (C) Recipe ingredients length distribution (D) Network graph

Results

Exploratory Data Analysis. We first ran some EDA to give us insight into the dataset. Fig.1A told us that among the recipes included in this dataset, Italian has the most while Brazilian has the least. Fig.1B told us that salt, olive oil and onions are the three most common ingredients. Fig.1C told us a recipe has around 10 ingredients in average. In Fig.1D, we first calculated the overlap coefficients between each cuisine type based on the ingredients they share, then map the result to a network graph, it showed that all the Asian cuisine type tends to gather together. We also drew word cloud plot for ingredients within each cuisine types (Fig.2), and found out that there exist some essential differences between different cuisine types.

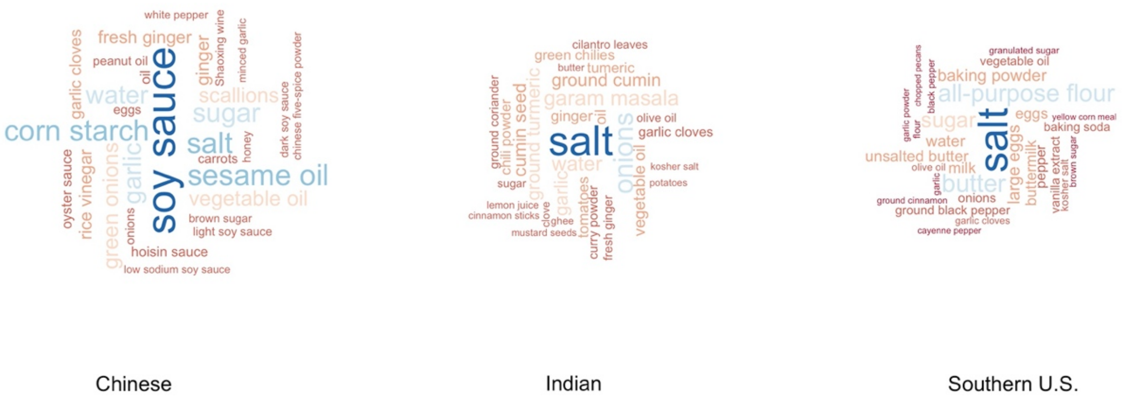

Figure 2. Word Cloud for Chinese, Indian and Southern U.S. cuisine.

Classification. Table 1 showed us the accuracies for each combination of feature extraction methods and classification models. We can see that, overall, Word2vec performs better than TF-IDF, and the combination of Word2vec and multinomial logistic regression models gives the best prediction accuracy.

		Classification			
		Decision Tree	Random Foest	OvR Linear SVM	Logistic
Feature Extraction	Word2Vec	0.54726	0.66522	0.64923	0.70233
	TF-IDF	0.31415	0.42226	0.32693	0.37982

Table 1. Prediction accuracies for testing dataset

Discussion

In our study, we assessed the performance of two different feature extraction methods: word2vec and TF-IDF, then we tested four different classification methods: SVM, logistic regression, decision tree and random forest. We found out that the combination of Word2vec and multinomial logistic regression models gives the best prediction accuracy, which can later be used for our cuisine recommendation system.

But there exists some limitation to our study. For example, we didn't have enough time and resources to conduct cross validation to tune some of the hyper parameters, maybe some models can perform better if giving a better parameter set. Also, since the feature vectors are calculated by machine learning methods, it'll be hard to interpret them, hence make the interpretation of the parameters difficult, which is not considered to be very 'biostatistical'.

Reference

[1] https://drhyman.com/blog/2011/01/07/how-eating-at-home-ca n-save-your-life/

[2] http://somethingnewfordinner.com/blog/new-years-resolution -cook-more-often/

[3] https://www.medicaldaily.com/health-benefits-home-cooked- meals-242919

[4] https://www.fix.com/blog/perks-of-home-cooked-meals/#Sources

[5] http://somethingnewfordinner.com/blog/new-years-resolution -cook-more-often/