What's Cooking

Siyu Chen

Xi'an Shiyou University

(None)

Overview

Problem Definition

Problem Description

Description

Asks you to predict the category of a dish's cuisine given a list of its ingredients.

- 1. Dish data loading;
- 2. Preprocess seasoning and visualization data set structure;
- 3. Loading Logistic Regression Model and Ensemble Model training;
- 4. Test and submit the results to see the experimental score.

Read Data

- train.tsv :The data used for training that contains ID, cuisine, and ingredients.
- test.tsv : After training the model with the training data set, use the test data set to generate a file similar to sample_submission.csv.
- sampleSubmission.csv : A submission that meets the purpose.

Infer the characteristics of its cuisine by the seasonings used:

- ID : Each data sequence number
- ingredients : The reason for classification is also the most important feature.
- cuisine : It’s the target of the classification,The total dish coefficient is 20,such as’brazilian’ ’british’ ’chinese’ ’filipino’.

Analysing Data

Data Statistics

- Statistic the data in the training set.

Table 1: Statistic the data in the training set

	ID	*Cuisine*	*Ingredients*
0	10259	greek	[romaine lettuce, black olives, grape tomatoes...
1	25693	southern_us	[plain flour, ground pepper, salt, tomatoes, g...
2	20130	filipino	[eggs, pepper, salt, mayonaise, cooking oil, g...
3	22213	indian	[water, vegetable oil, wheat, salt]
4	13162	indian	[black pepper, shallots, cornflour, cayenne pe...

Dataframe Analysis

- To better process the data, we need to do the following:
  - Count the total data of training set and test set.
    train shape: 39774
    test shape: 9944
  - Maximum Number of Ingredients in a Dish: 65
  - Minimum Number of Ingredients in a Dish: 1
    train: 8529
    test: 3310

Visualization and Data Preprocessing

Number of receipes by Cuisine

- Look at the graphs below
  - Italian cuisine dominates all cuisine.

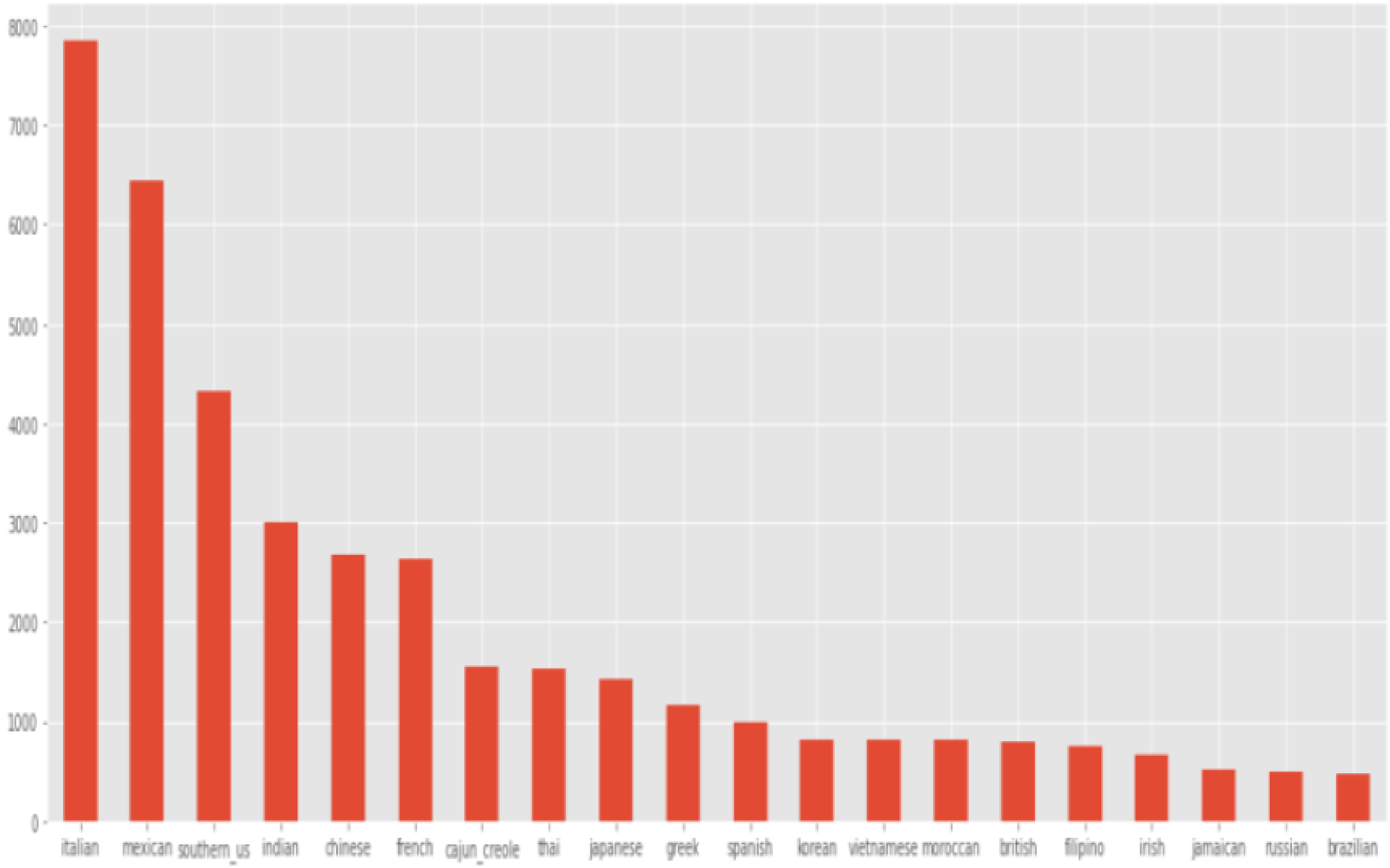

Figure 1: number of receipes by cuisine

Ingredients in a Dish

- Look at the graphs below
  - Ingredients in a dish distribution.

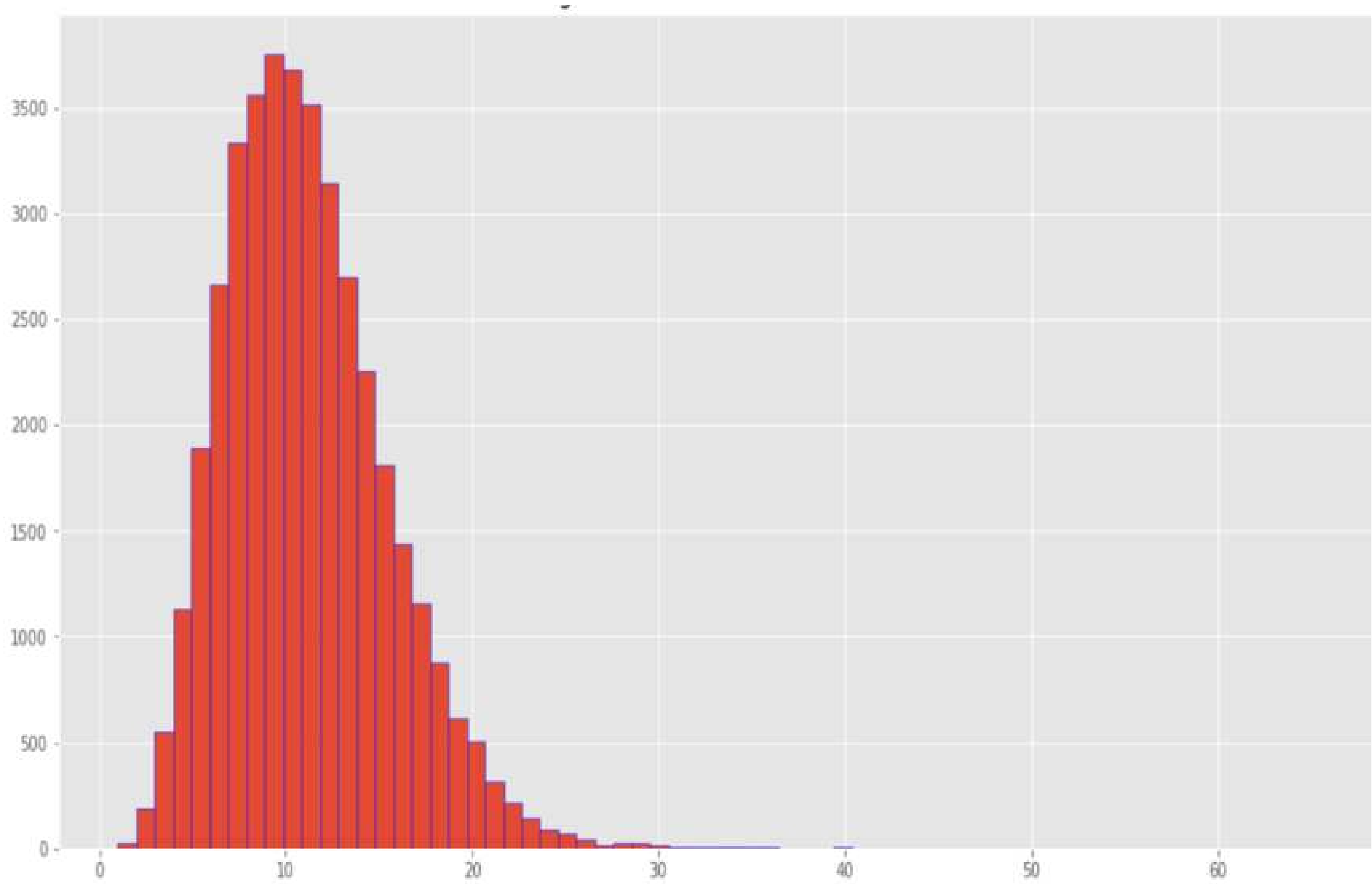

Figure 2: distribution

Main Ingredients

- Look at the graphs below
  - Salt is the largest share of all the Greek ingredients.

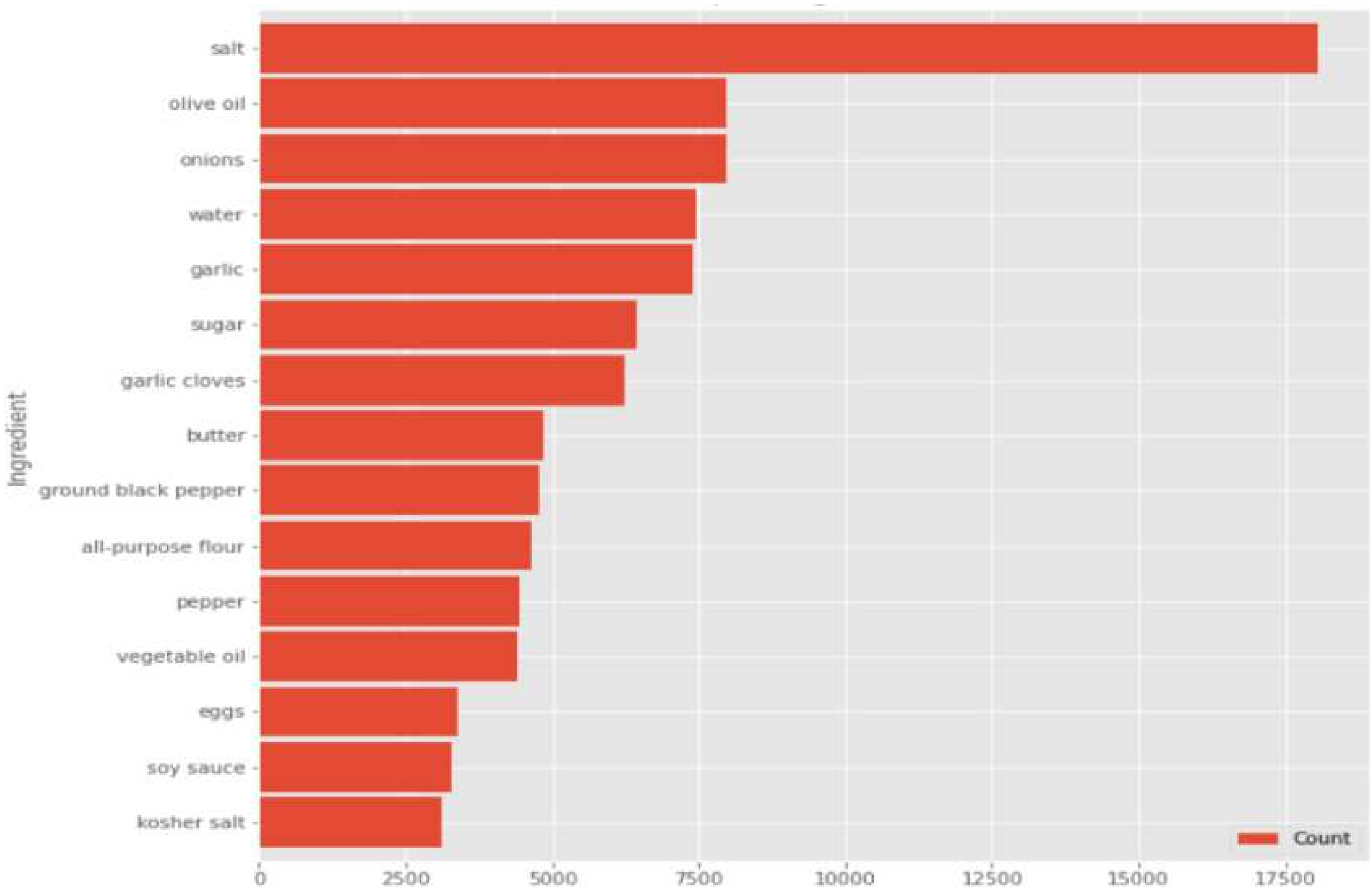

Figure 3: top 15 ingredients

Ingredients in Each Cuisine

- The proportion of ingredients in Mexican cuision.

- The proportion of ingredients in Chinese cuision.

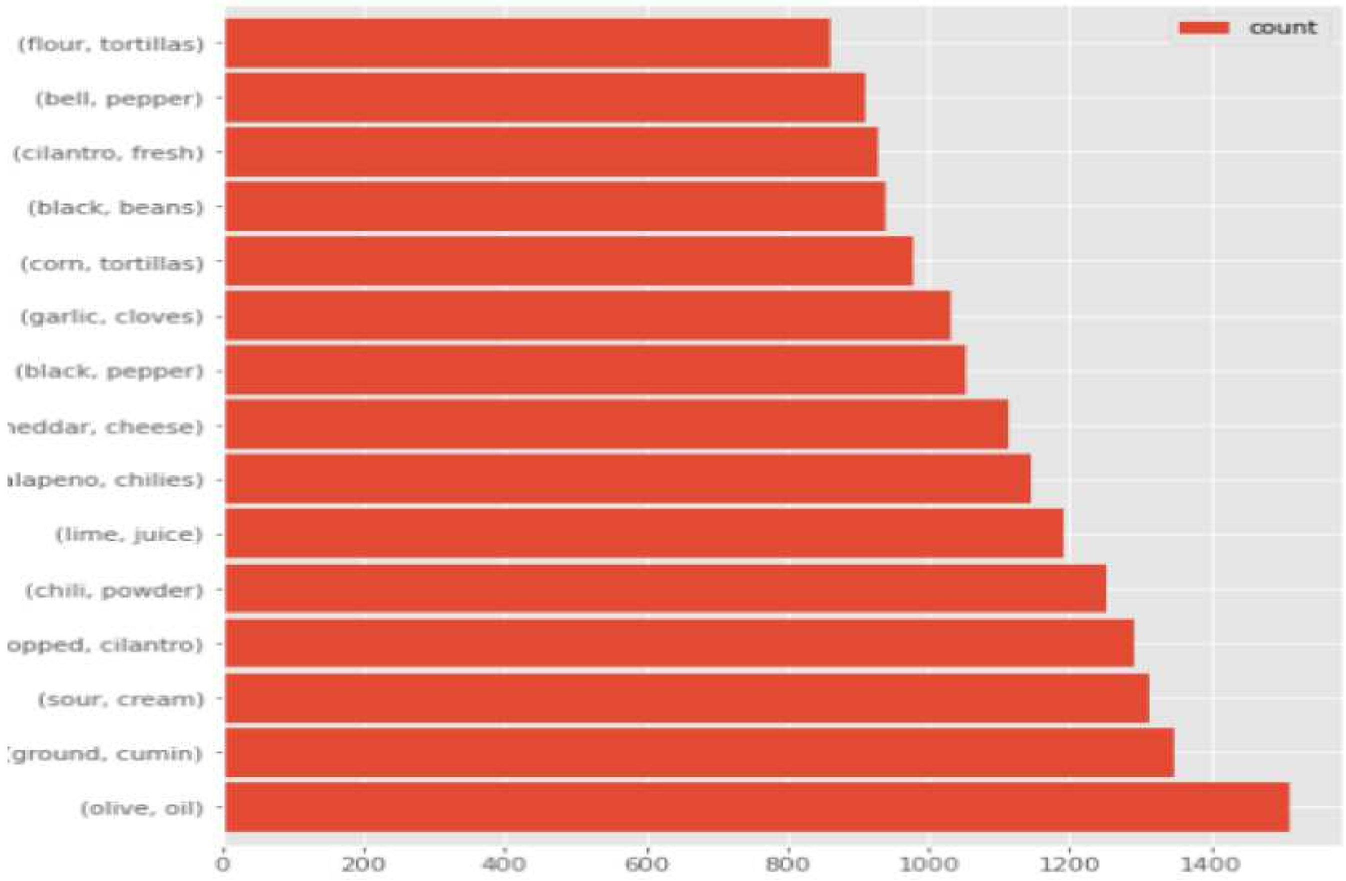

Figure 4: Mexican cuisine

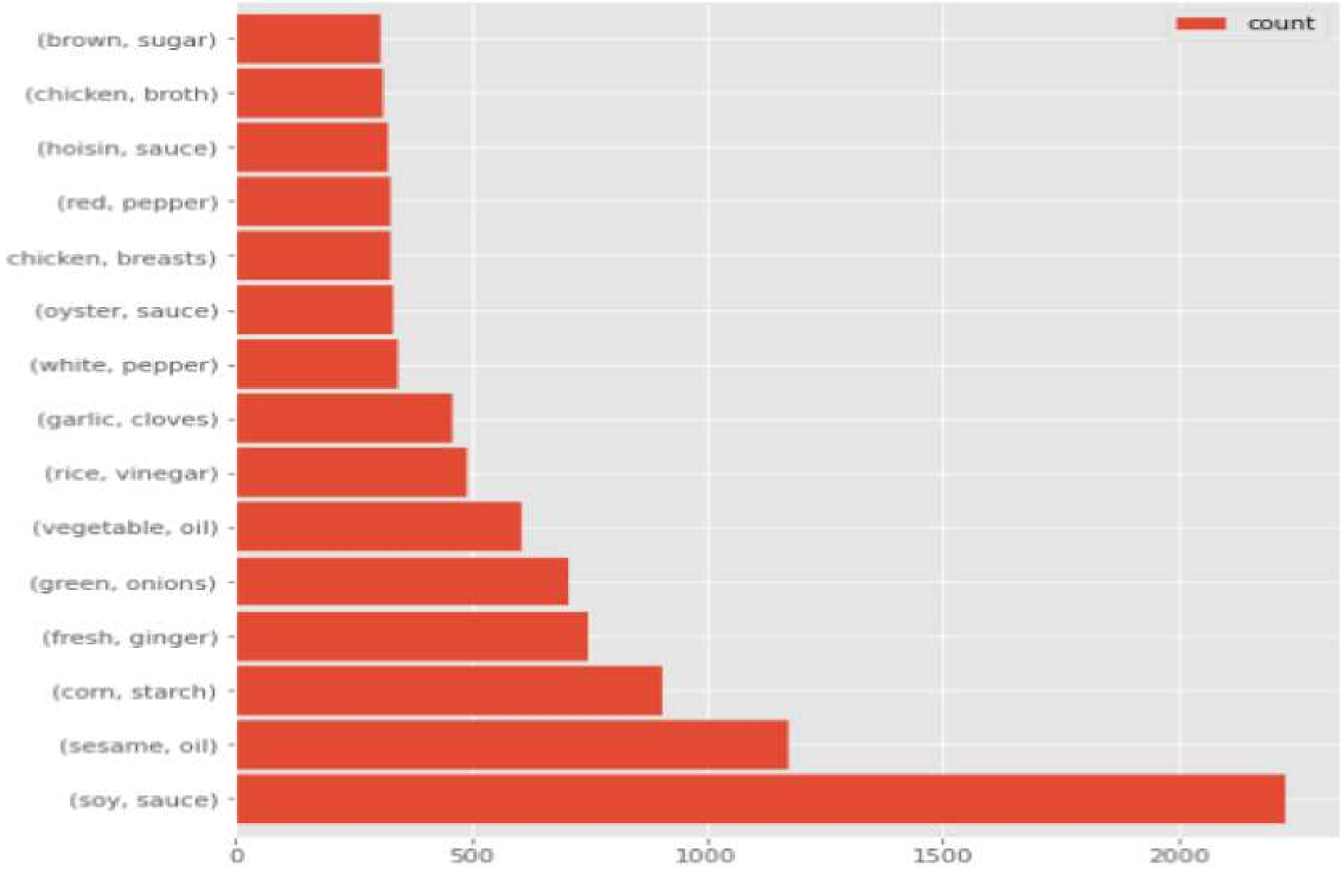

Figure 5: Chinese cuisine

NLP Analysis

TF-IDF Algorithm

- Model of TF-IDF algorithm

$$TF-IDF(d,w) = TF(d,w) * IDF(w)$$

- $TF(d,w)$ ⇔ Frequency of occurrence of w in document d.
- $IDF(w) = log\frac{N}{N(w)}$
- N ⇔ The total number of documents in a corpuss.
- N(w)⇔ How many documents does the w appear in.

TfidfVectorizer Grammar

- Steps.
  - The counting matrix of words is converted to TF-IDF representation, and then normalized.
  - Scikit-learn provides a TfidfVectorizer class, which has the ability to remove common stop words (like a, the, and, or).
  - TF-IDF tends to filter out common words and retain important words.

Modeling

Logistic Regression and Ensemble Model

- Logistic Regression
  - Random seeds are not fixed and generate random sequences.
  - Use the logistic regression model in sklearn.
  - Score:0.787711182622687.
- Ensemble Model
  - Ensemble in Sklearn is called to integrate the two classifiers, logistic regression and SVM.
  - in the way of soft voting, to show the accuracy
  - Score:0.8119469026548672.

Create Submission

Table 2: Predictions from first level models

	ID	Cuisine
0	18009	british
1	28583	southern_us
2	41580	italian
3	29752	cajun_creole
4	35687	italian
5	38527	southern_us

Reflection and Summary

- Dishes can contain a variety of ingredients, and the same ingredients may vary in number and number, so the integredients need to be filtered.
- KNN mainly depends on the surrounding limited adjacent samples, rather than on the method of discriminating class domain to determine the category.
- KNN basically does not learn, resulting in a slower prediction speed than logistic regression and other algorithms.

Contact Information

Thank you for listening!

Siyu Chen

Xi'an Shiyou University

✉ 785987165@qq.com