What's Cooking

Siyu Chen

Xi'an Shiyou University, China

Introduction

Asks you to predict the category of a dish's cuisine given a list of its ingredients. Test and submit the results to see the experimental score.

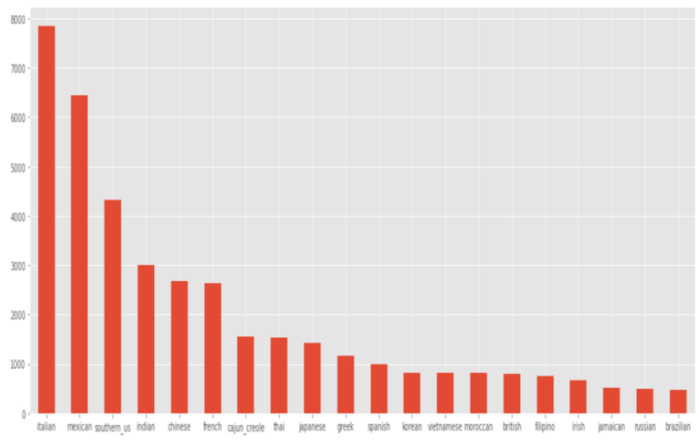
ID	Cuisine	Ingredients
10259	greek	romaine lettuce, black olives, grape tomatoes...
25693	southernus	plain flour, ground pepper, salt, tomatoes, g...
20130	filipino	eggs, pepper, salt, mayonaise, cooking oil, g...
22213	indian	water, vegetable oil, wheat, salt]
13162	indian	black pepper, shallots, cornflour,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

Dataframe Analysis

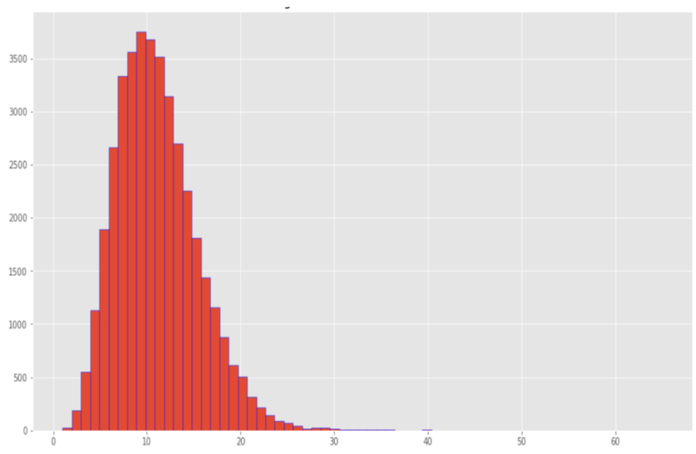
- To better process the data, we need to do the following:
  - Count the total data of training set and test set.
    train shape: 39774
    test shape: 9944
  - Maximum Number of Ingredients in a Dish: 65
  - Minimum Number of Ingredients in a Dish: 1
    train: 8529
    test: 3310

Data Visualization

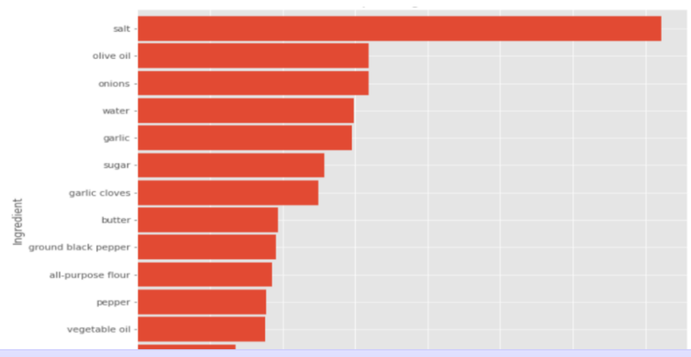
Number of receipes by Cuisine. Italian cuisine dominates all cuisine.



Ingredients in a dish distribution.



Salt is the largest share of all the Greek ingredients.



NLP Analysis

- Model of TF-IDF algorithm

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

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

  By comparing the accuracy of the two models, we found that the integrated accuracy is higher than that of a single classifier.

Create Submission

- Output Dataframe:

	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

- Disadvantages and optimizations:
  - 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.