FLIP 01 PROJECT REPORT

Bing Liu[1]

[1] Jilin University, China

Introduction

This is a classification prediction problem. participants are asked to predict the category of a dish's cuisine given a list of its ingredients.

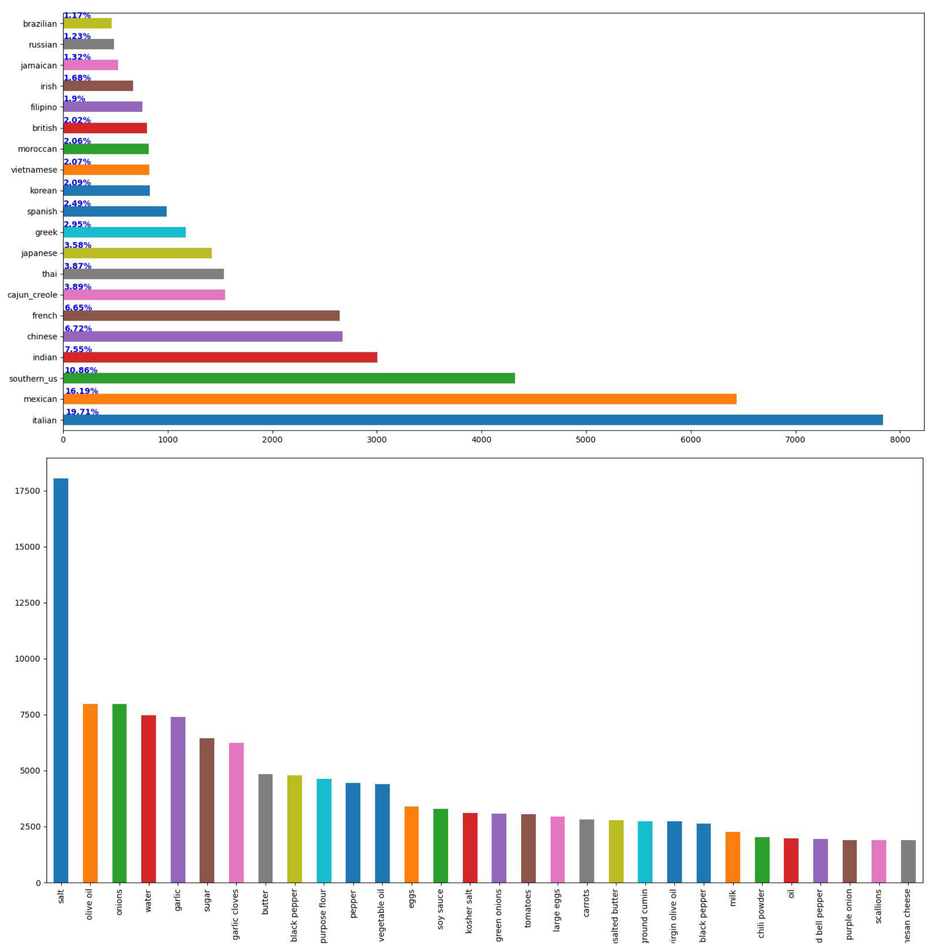
- **train.json** – the training set containing recipes id, type of cuisine, and list of ingredients. It contains 39774 entries.
- **test.json** – the test set containing recipes id, and list of ingredients. It contains 9944 entries.
- **sampleSubmission.csv** – a sample submission file in the correct format. It just contains 2 columns that is "id" and "cuisine".

Data processing

- **Missing values analysis** statistic missing value in the dataset.
- **Statistic Cuisine** There are a total of 20 types of cuisines, and the percentage of each type of cuisine is as follows.
- **Word Frequency Statistics** Count the number of occurrences of each word in the entire dataset 'ingredients', In order to analyze the importance of related vocabulary.

Data Visualization

Data visualization of the statistic of cuisine and word frequency statistics.



Feature Engineering

- **String Preprocess** 1:Use the WordNetLemmatizer().lemmatize() method to restore the part of speech
  2:remove the useless suffix of the word
  3:remove the non-letter symbols
  4:change the uppercase letters to lowercase
- **Count Vectorizer** Convert a document into a vector by counting to complete feature extraction, which get a word frequency matrix.
- **TFiDF Vectorizer** Input the word frequency matrix to get the TF-IDF weight matrix.
- **Cluster as Parameter** There are 20 different types of cuisine to classify. Certain groups of cuisine may have much more similarity than others. So we use the clustering information as part of the feature.

Model

I have choose the SVC as classification model.

```
from sklearn.svm import LinearSVC, SVC

C = 604.5300203551828
gamma = 0.9656489284085462

clf = SVC(C=float(C), gamma=float(gamma), kernel='rbf')
clf.fit(train, target)
y_pred = clf.predict(test)
```

Conclusion

Accuracy rate method The percentage of the number of correctly classified cuisines in the total number is used as the accuracy rate.

accuracy Value My accuracy rate is 81.06%