FLIP 01 PROJECT REPORT

BAOJEI ZHANG

Abstract. In this report, the topic of kaggle was mainly discussed: what's cooking? After thinking and solving, finally make a summary of this practice process and the problems encountered.

Contents

Date: 2020-11-17.
Key words and phrases. Kaggle topic selection, nlp, TFiDF Vectorizer.

1. Introduction

Picture yourself strolling through your local, open-air market... What do you see? What do you smell? What will you make for dinner tonight? If you're in Northern California, you'll be walking past the inevitable bushels of leafy greens, spiked with dark purple kale and the bright pinks and yellows of chard. Across the world in South Korea, mounds of bright red kimchi greet you, while the smell of the sea draws your attention to squids squirming nearby. India's market is perhaps the most colorful, awash in the rich hues and aromas of dozens of spices: turmeric, star anise, poppy seeds, and garam masala as far as the eye can see. The subject requirements are asks to predict the category of a dish's cuisine given a list of its ingredients.
In the dataset, we include the recipe id, the type of cuisine, and the list of ingredients of each recipe (of variable length). The data is stored in JSON format.

Name	Description	Attribute
train.json	training set(the type of cuisine, and the list of ingredients of each recipe)	Data: id, cuisine, ingredients
test.json	Test set(predict the cuisine type of the list ingredients)	Data:id,ingredients
sample.csv	a sample submission file	Data:id,ingredients

2. Data Analysis

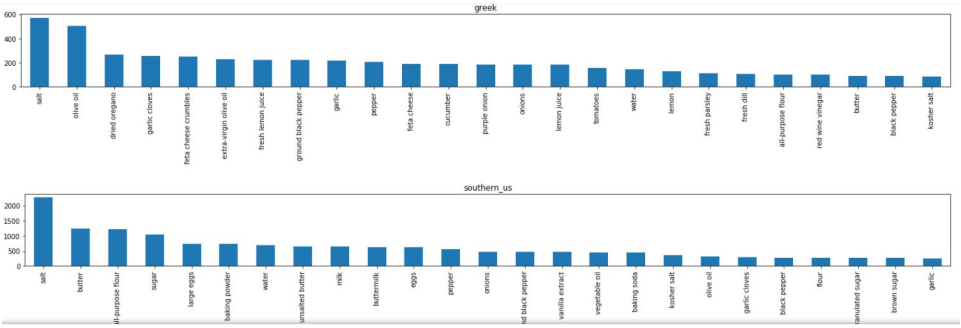
- There are 39774 data in the training set.
  There are 9944 data in the test set.
- Data is imported as DataFrame object, each recipe is a separate line.
- There are no missing values in the training set.
- The following results are obtained by sorting and sorting the data of ingredients:

id	cuisine
salt	18049
onions	797
olive oil	7972
water	7457
garlic	7380

From here we can see that in the dishes in the training set, salt, water, oil and other ingredients are used more, which is not very useful for our future learning.

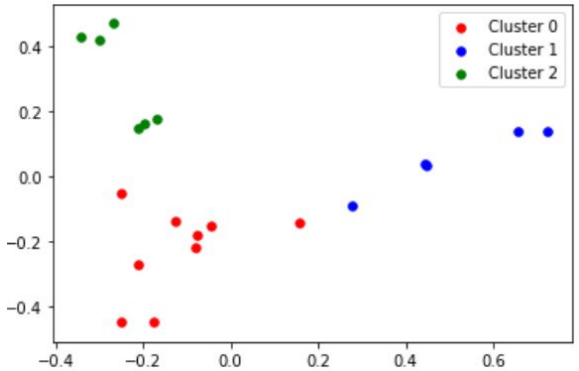
3. Data Description and Processing

We can find the percentage of each country's dishes in the total test set. We can find the percentage of each country's dishes in the total test set. At this point, we will find that in the test set, Italian dishes account for the largest percentage, more tahn Nineteen percent in the training set, followed by Mexican dishes, which more than sixteen percent in the training set, and the following order is:
Itian, Indian,Chinese, French,Cajun creole, thai, japanese, greek, spanish, korean, vietnamese, moroccan,british,fillplno, ifish, jamaican, russian, brazilian.
We can find the ingredients that account for the most ingredients in each country's dishes:



In the same way, we can find the 25 cooking ingredients or cooking essences with the least amount of each dish.

In the data processing stage, we first need to process the text data. Since the initial data type we get is dataFrame type data, we first need to process the data and convert the original list type data to string type.
After the data is converted, we need to use TF-IDF to calculate the degree of similarity between each ingredient and the dish, on this basis, we can find dishes with more similarities:



Next, we can use Linear SVC to predict the dishes: For the first modeling synthesis, its best params are'C': 1,'loss':'hinge','penalty':'l2'
The best result we can get at this time is 0.7857646647354912.

Conclusion

In this practice, I learned how to process text-related materials. When processing text, it is roughly divided into the following steps: 1. Text preprocessing; 2.

Text representation; 3. Spatial dimensionality reduction; 4. Classification model training; 5. Classification performance evaluation.
In the above steps, I still have deficiencies in the two aspects of spatial dimensionality reduction and classification performance evaluation. In the subsequent learning, I will pay attention to strengthening the learning of these two aspects.In addition, in the process of processing the data, I also tried to use the bag of words model, and here is a word cloud, which is a very interesting attempt.

In general, I have learned a lot during this practice, but there are also many shortcomings, which need to be filled in the next time.