# What's cooking?

Baojie Zhang

Xi'an Shiyou University

2020-11-17

# Overview



## Problem Description

Problem Description

## Data Analysis

data analysis

## Data Preprocessing

TFiDF Vectorizer

similar dishes

## Model Development



# Problem Description

# Problem Description



**Problem**

This time The Kaggle topic is What's cooking?
The subject requirements are asks to predict the category of a dish's cuisine given a list of its ingredients. There are three data sets:train.json,test.json,sample_submission.csv

Table 1: Data description

| 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_submission.csv | a sample submission file in the format | Data:id,cuisine |



# Data Analysis

# data analysis



Table 2: Data

| | id | cuisine | ingredients |
|---|---|---|---|
| 0 | 10259 | greek | ’romaine lettuce’, ’black olives’, ’grape tom... |
| 1 | 25693 | southern_us | ’plain flour’, ’ground pepper’, ’salt’, ’toma... |
| 2 | 20130 | filipino | ’eggs’, ’pepper’, ’salt’, ’mayonaise’, ’cooki... |
| 3 | 22213 | indian | ’water’, ’vegetable oil’, ’wheat’, ’salt’ |
| 4 | 13162 | indian | ’black pepper’, ’shallots’, ’cornflour’, ’cay... |

- 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.

# data analysis



- The percentage of dishes from each country in the total training set:

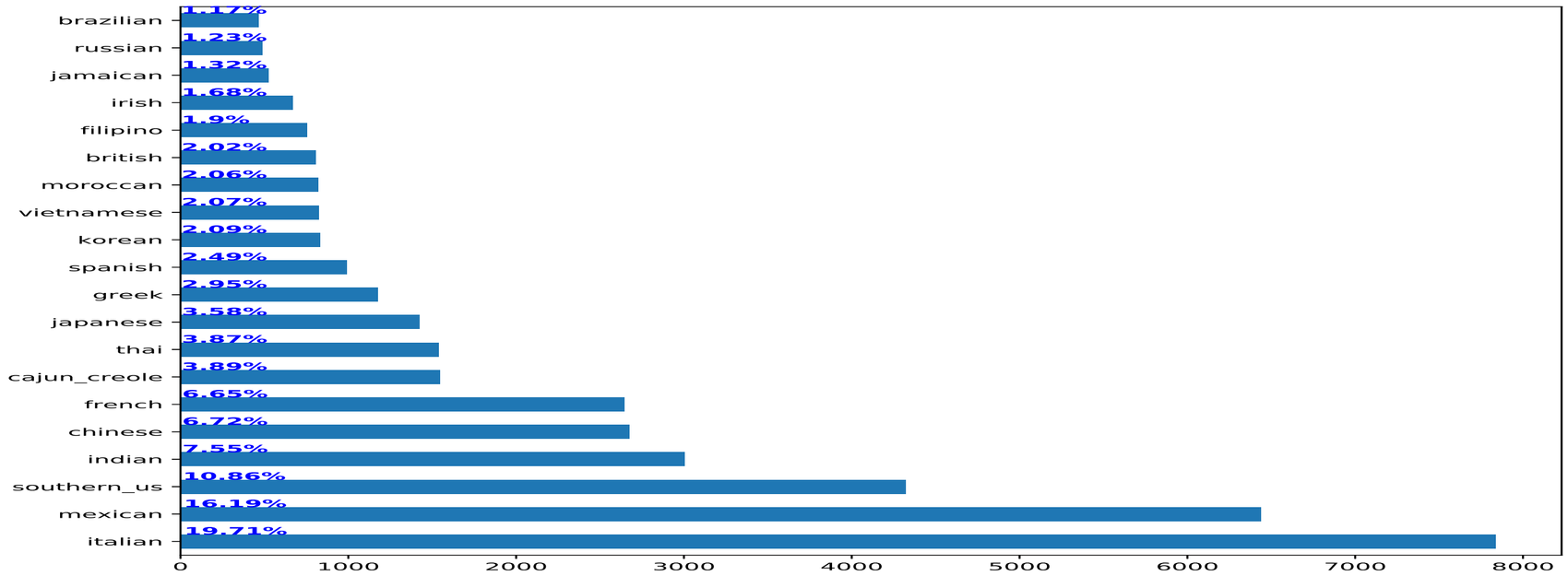


image1:The percentage of dishes



- The following results are obtained by sorting and sorting the data of ingredients:

Table 3: Data

| id | cuisine | id | cuisine | id | cuisine |
|---|---|---|---|---|---|
| salt | 18049 | sugar | 6434 | pepper | 4438 |
| onions | 797 | garlic cloves | 6237 | vegetable oil | 4385 |
| olive oil | 7972 | butter | 4848 | eggs | 3388 |
| water | 7457 | ground black pepper | 4785 | soy sauce | 3296 |
| garlic | 7380 | all-purpose flour | 4632 | kosher salt | 3113 |

- We can also get the dish label from the training set,as follows:
'greek', 'southern_us', 'filipino', 'indian', 'jamaican', 'spanish', 'italian', 'mexican', 'chinese', 'british', 'thai', 'vietnamese', 'cajun_creole', 'brazilian', 'french', 'japanese', 'irish', 'korean', 'moroccan', 'russian'



# Data Preprocessing

# Data Preprocessing



First process the string:

- To remove everything except a-z and A-Z and to make list element a string element.

Table 4: Data

| id | cuisine | ingredients |
|---|---|---|
| 0 | greek | romaine lettuce black olive grape tomato garli... |
| 1 | southern_us | plain flour pepper salt tomato black pepper ... |
| 2 | filipino | egg pepper salt mayonaise cooking oil green ch... |
| 3 | indian | water vegetable oil wheat salt |
| 4 | indian | black pepper shallot cornflour cayenne pepper ... |

# TFiDF Vectorizer



TFiDF Vectorizer on the processed data:

- Use TFiDF Vectorizer to evaluate the importance of each dish of vegetable raw materials

# similar dishes



More similar dishes and visualize dishes:

From this picture, We can notice there are 3 clusters of cuisines.

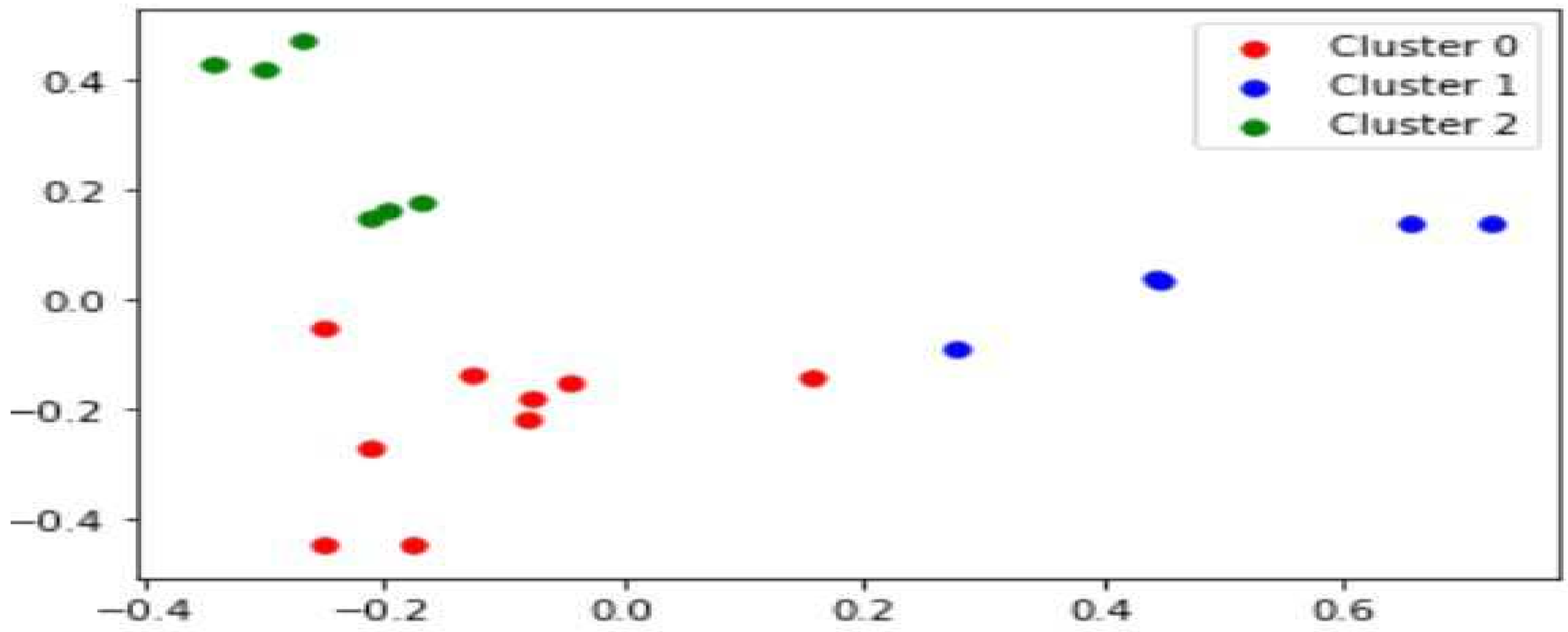


image2:More similar dishes



# Models and predictions

# Model



Use Linear SVC model for prediction:

- best_score:0.785789806155981 5