WHAT'S COOKING

JINCAI MA

ABSTRACT. The project will use the data set provided by Yummly to train and test a model and test its performance and predictive power.A good model trained by this data can be used to predict the cuisine.The dataset for this project is from the Kaggle What's Cooking competition.A total of 39,774/9,944 training and test data points, covering information on Chinese, Vietnamese, French, etc.

CONTENTS

Date: (None).
1991 *Mathematics Subject Classification.* What's Cooking.
Key words and phrases. Application software learning, Kaggle topic selection.

1. Introduction

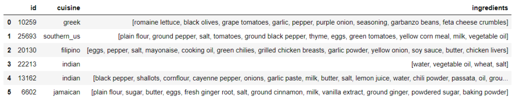
- Use recipe ingredients to categorize the cuisine. Given the name of the condiment, predict the cuisine to which the dish belongs.
- In the dataset,including the recipe ID, the dish, and the list of ingredients for each recipe (variable length).The data is stored in JSON format.
  1.train.json- A training set that contains the recipe ID, dish type, and ingredient list
  2.test.json- A test set containing a recipe ID and a list of ingredients
  3.sample_submission.csv-Properly formatted sample submission document

2. Data Analysis

First of all, our work can be divided into the following steps:

(1) Data Import And Introduction

- Import the JSON file with Pandas:We can get the data set of dish names, including 39774 training data and 9944 test samples.To see the distribution of our data set and the total variety of dishes, we printed out some of the data samples.

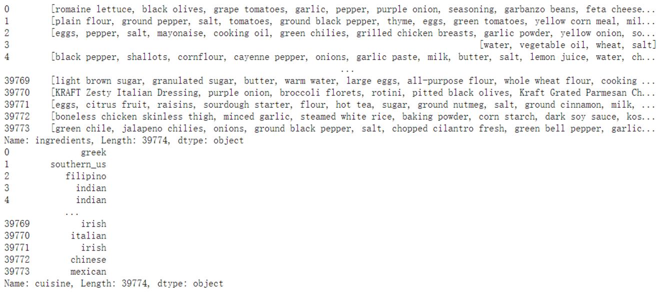


	id	cuisine	ingredients
0	10259	greek	[romaine lettuce, black olives, grape tomatoes, garlic, pepper, purple onion, seasoning, garbanzo beans, feta cheese crumbles]
1	25693	southern_us	[plain flour, ground pepper, salt, tomatoes, ground black pepper, thyme, eggs, green tomatoes, yellow corn meal, milk, vegetable oil]
2	20130	filipino	[eggs, pepper, salt, mayonaise, cooking oil, green chilies, grilled chicken breasts, garlic powder, yellow onion, soy sauce, butter, chicken livers]
3	22213	indian	[water, vegetable oil, wheat, salt]
4	13162	indian	[black pepper, shallots, cornflour, cayenne pepper, onions, garlic paste, milk, butter, salt, lemon juice, water, chili powder, passata, oil, grou...
5	6602	jamaican	[plain flour, sugar, butter, eggs, fresh ginger root, salt, ground cinnamon, milk, vanilla extract, ground ginger, powdered sugar, baking powder]

- Total dish classification
  There are 20 dishes in total, which are: ['brazilian' 'british' 'cajun_creole' 'chinese' 'filipino' 'french' 'greek' 'indian' 'irish' 'italian' 'jamaican' 'japanese' 'korean' 'mexican' 'moroccan' 'russian' 'southern_us' 'spanish' 'thai' 'vietnamese']

(2)Analyze Data

- The data set is divided into Features and Target Variables.
- Features:'ingredients', we were given the names of the ingredients contained in each dish; Target variable:'cuisine', is the classification of cuisines that we want to predict.
- Extract the Feature of training data set into train_integredients variable Extract the Target Variables into the train_Targets variable.

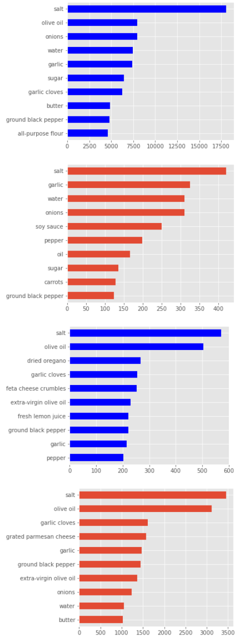


```
0        [romaine lettuce, black olives, grape tomatoes, garlic, pepper, purple onion, seasoning, garbanzo beans, feta cheese...
1        [plain flour, ground pepper, salt, tomatoes, ground black pepper, thyme, eggs, green tomatoes, yellow corn meal, mil...
2        [eggs, pepper, salt, mayonaise, cooking oil, green chilies, grilled chicken breasts, garlic powder, yellow onion, so...
3                                                                                            [water, vegetable oil, wheat, salt]
4        [black pepper, shallots, cornflour, cayenne pepper, onions, garlic paste, milk, butter, salt, lemon juice, water, ch...
                                                                ...
39769    [light brown sugar, granulated sugar, butter, warm water, large eggs, all-purpose flour, whole wheat flour, cooking ...
39770    [KRAFT Zesty Italian Dressing, purple onion, broccoli florets, rotini, pitted black olives, Kraft Grated Parmesan Ch...
39771    [eggs, citrus fruit, raisins, sourdough starter, flour, hot tea, sugar, ground nutmeg, salt, ground cinnamon, milk, ...
39772    [boneless chicken skinless thigh, minced garlic, steamed white rice, baking powder, corn starch, dark soy sauce, kos...
39773    [green chile, jalapeno chilies, onions, ground black pepper, salt, chopped cilantro fresh, green bell pepper, garlic...
Name: ingredients, Length: 39774, dtype: object
0              greek
1        southern_us
2           filipino
3             indian
4             indian
            ...
39769          irish
39770        italian
39771          irish
39772        chinese
39773        mexican
Name: cuisine, Length: 39774, dtype: object
```

(3) Data Visualization

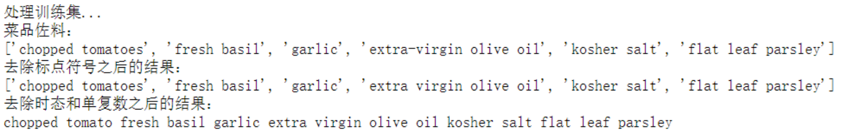
What are the top 10 most frequently used ingredients?

What are the 10 most common ingredients in filipino,greek and Italian cuisine?

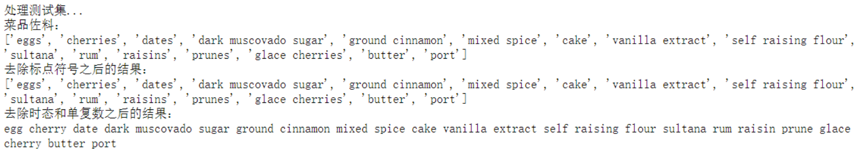


3. Build Model

(1) Data Cleaning:Since dishes contain a large number of ingredients, and since the same ingredients can vary in numbers, tenses, and so on, we considered sifting through a potatos to remove any such differences.



```
处理训练集...
菜品佐料:
['chopped tomatoes', 'fresh basil', 'garlic', 'extra-virgin olive oil', 'kosher salt', 'flat leaf parsley']
去除标点符号之后的结果:
['chopped tomatoes', 'fresh basil', 'garlic', 'extra virgin olive oil', 'kosher salt', 'flat leaf parsley']
去除时态和单复数之后的结果:
chopped tomato fresh basil garlic extra virgin olive oil kosher salt flat leaf parsley
```



```
处理测试集...
菜品佐料:
['eggs', 'cherries', 'dates', 'dark muscovado sugar', 'ground cinnamon', 'mixed spice', 'cake', 'vanilla extract', 'self raising flour',
'sultana', 'rum', 'raisins', 'prunes', 'glace cherries', 'butter', 'port']
去除标点符号之后的结果:
['eggs', 'cherries', 'dates', 'dark muscovado sugar', 'ground cinnamon', 'mixed spice', 'cake', 'vanilla extract', 'self raising flour',
'sultana', 'rum', 'raisins', 'prunes', 'glace cherries', 'butter', 'port']
去除时态和单复数之后的结果:
egg cherry date dark muscovado sugar ground cinnamon mixed spice cake vanilla extract self raising flour sultana rum raisin prune glace
cherry butter port
```

(2) Feature extraction

- We convert the ingredients of the dish into a numerical feature vector.Consider that most dishes include salt, water, sugar, butter, etc,We will consider weighting the seasonings according to the occurrence times of the seasonings, that is, the more the occurrence times of the condiments, the lower the discriminability of the condiments.The feature we adopt is TF-IDF.
- We can get the top 5 characteristics:['greek','southern_us','filipino','indian','indian']
- The top five data in train_tfidf:
  [[0. 0. 0. ... 0. 0. 0.]
  [0. 0. 0. ... 0. 0. 0.]
  [0. 0. 0. ... 0. 0. 0.]
  [0. 0. 0. ... 0. 0. 0.]
  [0. 0. 0. ... 0. 0. 0.]]

(3) Validation set partitioning

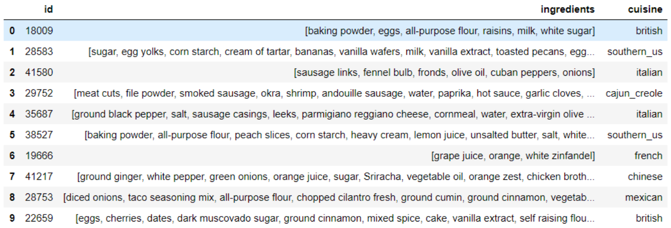
- The training set is divided into a new training set and a validation set by calling train_test_split function, which is convenient for the subsequent accuracy observation of the model.
- 1.Import train_test_split from sklear.model_selection
  2.Use train_tfidf and train_targets as input variables for train_test_split
  3.The test_size is set to 0.2, 20% of the validation set is divided, and 80% of the data is reserved for the new training set.
  4.Set the random_state random seed to ensure that the same partition results are obtained every time you run it.

(4) Training Model

- Invoke the logistic regression model in sklearn.
  1.Import the LogisticRegression from sklear.linear_model.
  2.GridSearchCV is imported from sklearn.model_selection, and the parameters are automatically searched. As long as the parameters are typed in, the best results and parameters can be given.
  3.Define the parameters variable: Create a dictionary for the C parameters, whose values are an array from 1 to 10;
  4.Define the classifier variable: Create a classification function using the imported LogisticRegression;
  5.Define Grid variables: Create a grid search object using the imported GridSearchCV;Pass the variables 'classifier', 'parameters' as arguments to the object constructor;
- After the model training, we calculated the prediction results of the model on the validation set X_VALID, and calculated the prediction accuracy of the model. The score on the validation set is: 0.7958516656191075.

(5) Predictive test set

Test set test_tfidf is predicted by the model grid, and then the predicted results are viewed.
The predicted number of test sets is 9944.



	id	ingredients	cuisine
0	18009	[baking powder, eggs, all-purpose flour, raisins, milk, white sugar]	british
1	28583	[sugar, egg yolks, corn starch, cream of tartar, bananas, vanilla wafers, milk, vanilla extract, toasted pecans, egg...	southern_us
2	41580	[sausage links, fennel bulb, fronds, olive oil, cuban peppers, onions]	italian
3	29752	[meat cuts, file powder, smoked sausage, okra, shrimp, andouille sausage, water, paprika, hot sauce, garlic cloves, ...	cajun_creole
4	35687	[ground black pepper, salt, sausage casings, leeks, parmigiano reggiano cheese, cornmeal, water, extra-virgin olive ...	italian
5	38527	[baking powder, all-purpose flour, peach slices, corn starch, heavy cream, lemon juice, unsalted butter, salt, white...	southern_us
6	19666	[grape juice, orange, white zinfandel]	french
7	41217	[ground ginger, white pepper, green onions, orange juice, sugar, Sriracha, vegetable oil, orange zest, chicken broth...	chinese
8	28753	[diced onions, taco seasoning mix, all-purpose flour, chopped cilantro fresh, ground cumin, ground cinnamon, vegetab...	mexican
9	22659	[eggs, cherries, dates, dark muscovado sugar, ground cinnamon, mixed spice, cake, vanilla extract, self raising flou...	british

4. Conclusions

Through this Kaggle project, I practiced by myself to have a deeper understanding of data visualization and to explore the structure and rules of data by means of drawing and tabulating.

List of Todos

Jupyter Notebook,Visual Studio Code,Latex,Git.