Group Outlying Aspects Mining

Jincai Ma

Xi'an Shiyou University, China

Introduction

The Data Source The data comes from Kaggle https://www.kaggle.com/c/bike-sharing-demand

Project Purpose Use recipe ingredients to categorize the cuisine.
Given the name of the condiment, predict the cuisine to which the dish belongs.

Related Field Name Interpretation

- 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

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']

Data Analysis

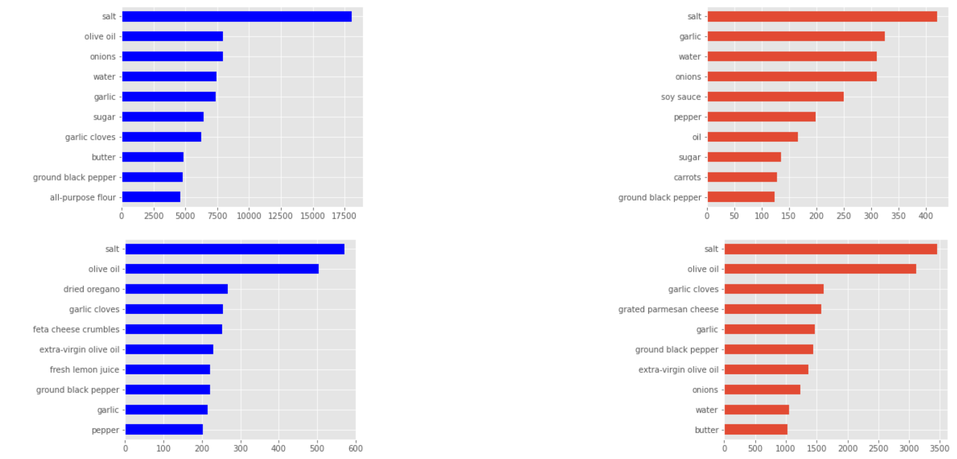
- The data set is divided into Features and Target Variables.
- Features:'ingredients', we were given the names of the ingredients contained in each dish. Target variable:'cuisine', is the classification of cuisines that we want to predict.
- Extract the Feature of training data set into train_integredients 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 crumbles]
1                    [plain flour, ground pepper, salt, tomatoes, ground black pepper, thyme, eggs, green tomatoes, yellow corn mea
l, milk, vegetable oil]
2          [eggs, pepper, salt, mayonaise, cooking oil, green chilies, grilled chicken breasts, garlic powder, yellow onion, soy sauce, b
utter, chicken livers]
3                                                                                                                [water, veget
able oil, wheat, salt]
4          [black pepper, shallots, cornflour, cayenne pepper, onions, garlic paste, milk, butter, salt, lemon juice, water, chili powder,
passata, oil, grou...
                                                                ...
39769    [light brown sugar, granulated sugar, butter, warm water, large eggs, all-purpose flour, whole wheat flour, cooking spray, boil
ing water, steel-cu...
39770                [KRAFT Zesty Italian Dressing, purple onion, broccoli florets, rotini, pitted black olives, Kraft Grated Parmes
an Cheese, red pepper]
39771                          [eggs, citrus fruit, raisins, sourdough starter, flour, hot tea, sugar, ground nutmeg, salt, ground c
innamon, milk, butter]
39772    [boneless chicken skinless thigh, minced garlic, steamed white rice, baking powder, corn starch, dark soy sauce, kosher salt, p
eanuts, flour, scal...
39773    [green chile, jalapeno chilies, onions, ground black pepper, salt, chopped cilantro fresh, green bell pepper, garlic, white sug
ar, roma tomatoes, ...
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
```

Data Analysis

- What are the top 10 most frequently used ingredients?
- What are the 10 most common ingredients in filipino,greek and Italian cuisine?



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
```

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 characteristics:['greek','southern_us','filipino','indian','indian', 'jamaican','spanish','italian','mexican','italian']

Build Model

1.Separate the training set and test set.
2.Remove unwanted eigenvalues:'casual','count','datetime','registered','date','atemp','mo
3.Cross validation is used to determine the optimal parameters.
4.View the selected optimal parameters:max depth: 20, n estimators: 150
5.Apply the optimal parameters to the model, it can be obtained
Accuracy on test set : 0.6945996275605214

Conclusion

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

Acknowledgement

- International Cooperation Project (Y7Z0511101) of IIE, Chinese Academy of Sciences