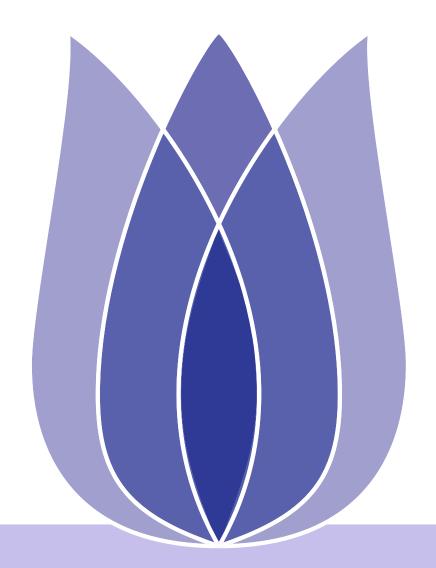
What's Cooking?

Yuhui Mou

Xi'an Shiyou University

October 16, 2020





Overview

Introduction

Data

Data Processing

Conclusion

Introduction

Data

Data Processing

Step One

Step Two





Data

Data Processing

Conclusion

Introduction





Data
Data Processing
Conclusion

Summary: Use recipe ingredients to categorize the cuisine

There are many countries on the earth, each country has its own characteristic food culture.

China's hot pot, American pizza, Japanese sushi... These are typical food of every country, while all these delicious food with different ingredients and different spices. Different food has its own characteristics. Chinese food is rarely used black pepper, while American food love black pepper. So with this project. According to the ingredient to distinguish food comes from which country.



Data

Data Processing

Conclusion

Data





Data

Introduction

Data

Data Processing

Conclusion

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.

ata

- train.json the training set containing recipes id, type of cuisine, and list of ingredients
- test.json the test set containing recipes id, and list of ingredients

In the test file test.json, the format of a recipe is the same as train.json, only the cuisine type is removed, as it is the target variable you are going to predict.



Introduction

Data

Data Processing

```
"id": 20130,
"cuisine": "filipino",
"ingredients": [
  "eggs",
  "pepper",
  "salt",
  "mayonaise",
  "cooking oil",
  "green chilies",
  "grilled chicken breasts",
  "garlic powder",
  "yellow onion",
  "soy sauce",
  "butter",
  "chicken livers"
```

Figure 1: Train.json





Data

Data Processing

Step One

Step Two

Conclusion

Data Processing





Step One

Introduction

Data

Data Processing

Step One
Step Two

- Read Data
 - Read train.json
 - Read test.json
- Processing Model:Bag-of-words model (BoW model)
 - ◆ BoW early in Natural Language Processing and Information Retrieval This model ignored the grammar and word order elements such as text, just as it is a collection of several words, the emergence of each word in the document are independent of each otherBoW to use an unordered list of words to express a text or a document.
 - ◆ CountVectorizer is a characteristic class of common numerical calculation, Is a text feature extraction method. For each training text, it only considers each of these words in the frequency of the training in the text. CountVectorizer Converts text of the words in the word frequency matrix. It does this by fit_transform function calculating the number of occurrences of all words.



Step Two

<u>Introduction</u>

Data

Data Processing

Step One
Step Two

Conclusion

After processing of data and extract the feature, It's time to choose a classifier.

Random Forest Classifier

- The first step is using feature and target training classifier.
- The second step is to input data to classifier which has trained before.
- Export data and Store them in a document.

```
读取训练集

39774

构造词袋

(39774, 1000)

[[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]
```

Figure 2: fileConstruct word bag

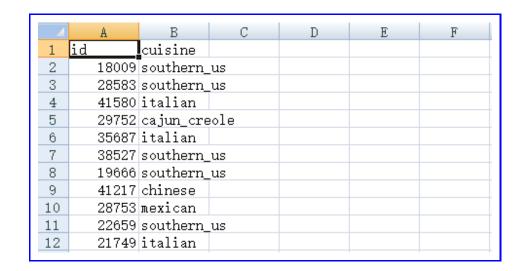


Figure 3: Output



Data

Data Processing

Conclusion





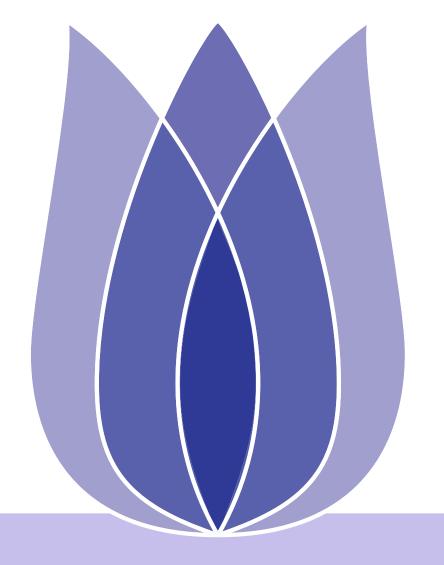
Introduction		
Data		
Data Processing		
~ 1 .		

- I can learn about theeating habits of different countries through this competition;
- I can practice professional skills;
- We can practice our competition tools, such as Latex Smartgit Github Python Data mining...;
- Know how to deal with large text class data set—-BoW model.
- Learn a classifier-RandomForestClassifier:can predict the test data and classify test data.
- But this method still has some drawbacks, such as test data just have txt, it's a single data. And, I just used one method to solve this problem. When I extracted the feature of the data and classified the train date, I just used CountVectorizer and RandomForestClassifier. This is the disadvantages of my project.



Contact Information

Thanks for watching!



Yuhui Mou Xi'an Shiyou University