THE REPORT OF FLIP01 FINAL PRESENTATION

ZHAOYANG WANG

Abstract. This report contains five parts.The first part will introduce the problem, describe the data and analyzes the problem. And the Second will do the statistic of the data, visualize the data . Third, this part will explain the method that i use. The Fourth will introduce the experiment and analysis of the algorithm and result.The last one is conclusion.

Contents

1. Introduction 2
1.1. Problem Statement 2
1.2. Data List 2
1.3. Problem Analysis 2
2. Exploratory Data Analysis 2
2.1. Data Information 2
2.2. Text Preprocessing 3
2.3. visualization 3
3. Methods 4
3.1. convert text to word vectors 4
3.2. model 4
4. Experiment and Analysis 6
5. Conclusion 6
References 7
List of Todos 7

Date: (None).
1991 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. Machine Learning, Data Mining, ...

1. Introduction

1.1. **Problem Statement.**

Some of our strongest geographic and cultural associations are tied to a region's local foods. This playground competitions asks you to predict the category of a dish's cuisine given a list of its ingredients. This is a natural language processing problem, so we need to use related methods to deal with it.

1.2. **Data List.** The data provided in this topic are country name, menu, ID. These three attributes have a country corresponding to each menu. Finally, it is required to establish a prediction model to which country the menu belongs.

cuisine: - The country of every ingredients.
ingredients: - Contains the ingredients needed for this dish.
ID: - Item ID.

1.3. **Problem Analysis.**

This problem is a text multi-classification problem in a typical natural language processing problem. First, the text data is cleaned. Second, the words in the text are converted into word vectors, which can then be processed by the computer. Finally put it into the classification model for classification.

2. Exploratory Data Analysis

2.1. **Data Information.**

From the table 1, we can clearly understand the situation of the train data set. and from the table 2, we can clearly understand the situation of the test data set. All we have to do is predict the cuisine of the ingredients in the test set.

Table 1. The head of the train data

	cuisine	id	ingredients
0	greek	10259	[romaine lettuce, black olives, grape tomatoes...
1	southern˙us	25693	[plain flour, ground pepper, salt, tomatoes, g...
2	filipino	20130	[eggs, pepper, salt, mayonaise, cooking oil, g...
3	indian	22213	[water, vegetable oil, wheat, salt]
4	indian	13162	[black pepper, shallots, cornflour, cayenne pe...

Table 2. The head of the test data

	id	ingredients
0	18009	[baking powder, eggs, all-purpose flour, raisi...
1	28583	[sugar, egg yolks, corn starch, cream of tarta...
2	41580	[sausage links, fennel bulb, fronds, olive oil...
3	29752	[meat cuts, file powder, smoked sausage, okra,...
4	35687	[ground black pepper, salt, sausage casings, l...

2.2. **Text Preprocessing.**

The text data needs to be pre-processed before the model prediction. The text we got contains a lot of information we don't need. Therefore, we need some algorithms and tools to pre-process the text. Then you can go to the next step.

- stopwords
- regularization
- convert to lowercase letters

2.2.1. *stopwords.*

In the text, there are many words that we don't need. They have no practical meaning. They are just words that complete the grammatical structure when people communicate. Therefore, these words can be removed by stop words, which makes the data required by our model simpler. These text words are also prevented from affecting model predictions.

2.2.2. *regularization.*

There are some punctuation marks in the text, or smiley faces, crying faces, and the like, which will also prevent our prediction of the text. Therefore, a regularization method is introduced to process the string and limit the characters of the text to the range we need.

2.2.3. *convert to lowercase letters.*

In English text, there are lowercase and uppercase. In our predictive models, we usually convert uppercase to lowercase. This makes the entire text lowercase. This can reduce the complexity of the text, so that you can get a better text to be processed.

2.3. **visualization.**

Draw word cloud for text by using word cloud technology. In this way, we can know which words appear more frequently in the text, which may provide some reference for our future model selection and model processing.

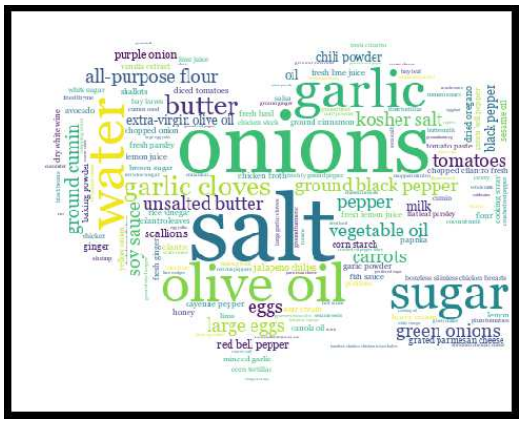

FIGURE 1. Displaying the words in text

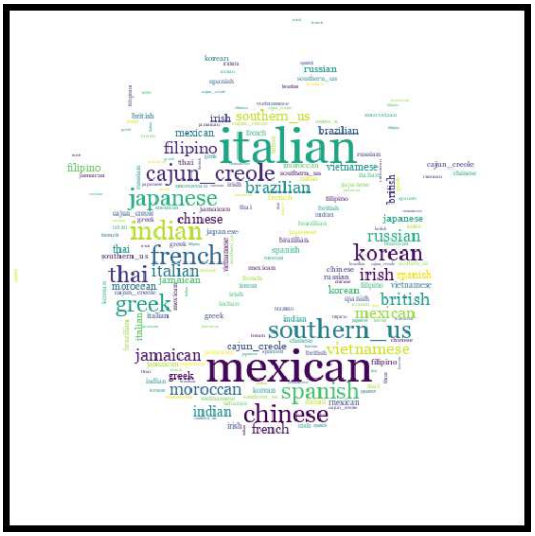

FIGURE 2. Displaying the words in text

From Figure 1 and Figure 2, we can see that Mexican is the most frequently occurring in Cuisine. And salt onions and other words are the most frequent words in the recipe.

3. METHODS

There are many machine learning methods for text classification. We have selected the following five methods:

- Logistic Regression
- KNN
- Random forest
- SVM
- CNN

3.1. **convert text to word vectors.**

By using word2vec, the words in the text are converted into word vectors, and then the word vectors are converted into sentence vectors. Make each label correspond to a sentence vector. In this way, labels and sentence vectors can be put into the model for classification.Use functions to convert text labels into numeric labels, which can eliminate the inconvenience caused by text labels.

3.2. **model.**

Use the following five models to classify the preprocessed data

3.2.1. *model predict.*

Text classification was performed using five methods of logistic regression, random forest, KNN, support vector machine, and convolutional neural network.

Among them, in KNN, random forest, support vector machine, a grid search method is used to adjust the parameters. Among them, the kernel function of the support vector machine is LinearSVC.

3.2.2. *CNN.*

In the convolutional neural network model, it is constructed as an embading layer, two convolutional base layers, and an output layer. It uses dropout technology. and batch normallization technology.

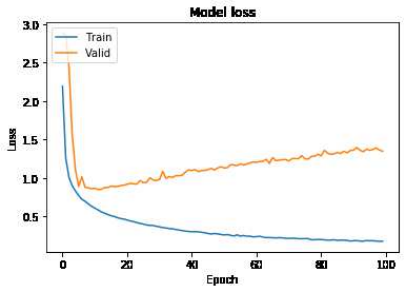

FIGURE 3. Displaying the relationship between the loss and epoch

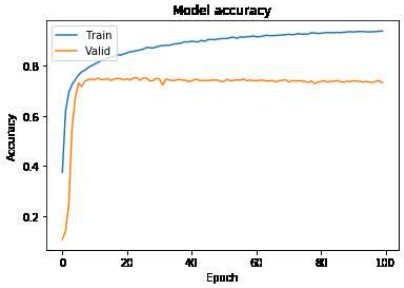

FIGURE 4. Displaying the relationship between the accuracy and epoch

It can be seen that the model gradually started to stabilize when iterating about 10 times.

3.2.3. *model score.*

The accuracy of each model is obtained through model training. Show in the table below.

TABLE 3. The score of models

	model	score
1	Logistic Regression	0.729
2	KNN	0.740
3	Random forest	0.739
4	SVM	0.736
5	CNN	0.753

4. Experiment and Analysis

This time the accuracy is slightly lower. There may be two reasons. The first is to use the average method when converting word vectors into sentence vectors. The other is that word vectors are trained with their own words, and the distance between word vectors is relatively close. So there is no distinction.

5. Conclusion

1.Using the Word2vec to help us process the textdata.If the text data is Chinese, we can use jieba for word segmentation.

2.There are many ways to deal with text classification in machine learning .we can select suitable ways on combination with the problem

3.In this problem, i use the mean of each words vector to caculate the sentence vector. Maybe this is the question why accuracy is lower than my espect.

References

List of Todos

(A. 1) School of Computer Science,, Xi'an Shiyou University, Shaanxi 710065, China
Email address, A. 1: xxx@tulip.academy