FLIP01 FINAL PRESENTATION

Zhaoyang Wang
Xi'an Shiyou University

February 25, 2020

Overview

Problem Statement

Text Preprocessing

Text feature extraction

Modeling

Conclusion

Thanks for watching

Problem Statement

Problem Definition

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.

Data Set

- train data

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...

- Display the data set

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...

Problem Statement

Text Preprocessing

Preprocessing visualization

Text feature extraction

Modeling

Conclusion

Thanks for watching

Text Preprocessing

Preprocessing

- stopwords
- regularization
- convert to lowercase letters

Since my text data is relatively clean, I only used the stopwords method

visualization

By using wordcloud to discribe the frequency of text data.

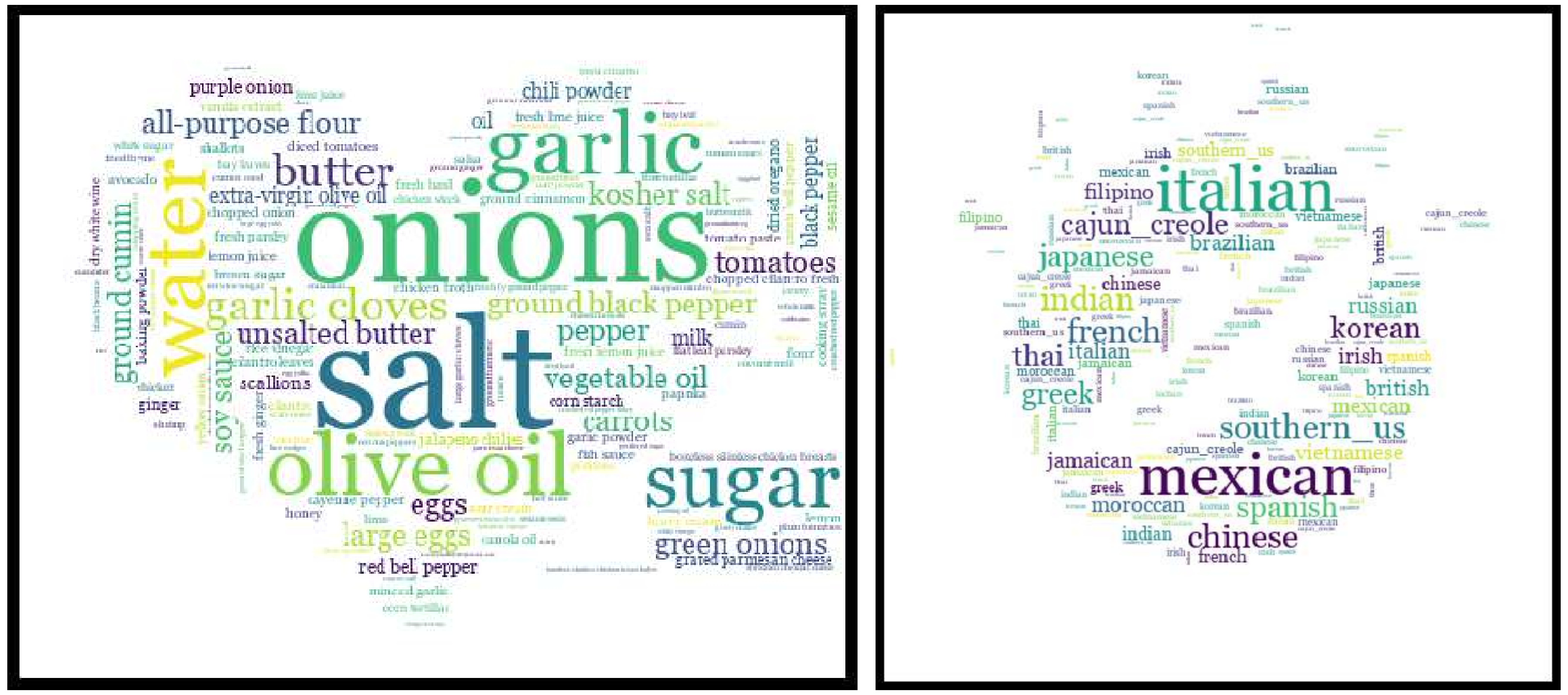

Figure 1: Displaying the words in text

Problem Statement
Text Preprocessing
Text feature extraction
Replace text labels
word2vec
Modeling
Conclusion
Thanks for watching

Text feature extraction

Replace text labels in the data set with numbers. Transform text in feature values into word vectors.

- unique() and apply()
- word2vec

Replace text labels

By using the unique() and apply() can replace the texe label into figures.

Table 3: Replace the text label

	cuisine	label
0	irish	16
1	italian	6
2	irish	16
3	chinese	8
4	mexican	7

word2vec

Problem Statement
Text Preprocessing
Text feature extraction
Replace text labels
word2vec
Modeling
Conclusion
Thanks for watching

Use word2vec to convert text to word vectors. And convert word vectors to sentence vector.and then for each sentence vector we have one label for it.

- vector size 300
- mean

Modeling

Modeling

- problem analysis

  This is a text classification problem.we can use many ways to solve text classification problems.

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

Step

- Divide training data and test data
- Do a model training
- Model evaluation
- Model prediction

The score of models

Table 4: 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

Conclusion

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

Problem Statement
Text Preprocessing
Text feature extraction
Modeling
Conclusion
Thanks for watching

Thanks for watching