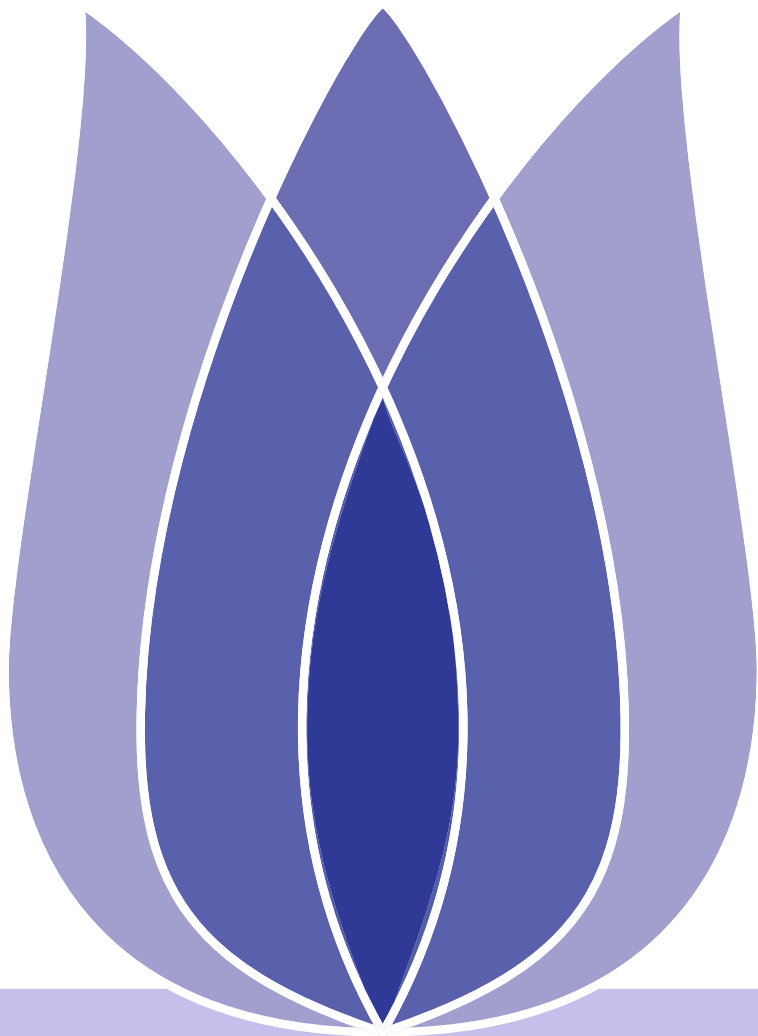


FLIP01 MIDTERM PRESENTATION

Zhaoyang Wang
Xi'an Shiyou University

January 12, 2020





Overview

- [Problem Statement](#)
- [Text feature extraction](#)
- [Modeling](#)
- [Conclusion](#)
- [Thanks for watching](#)

Problem Statement

Text feature extraction

Modeling

Conclusion

Thanks for watching



- Problem Statement
- Problem Definition
- Data Set
- Text feature extraction
- Modeling
- Conclusion
- Thanks for watching

Problem Statement



Problem Definition

Problem Statement
Problem Definition
Data Set
Text feature extraction
Modeling
Conclusion
Thanks for watching

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

- Problem Statement
- Problem Definition
- Data Set
- Text feature extraction
- Modeling
- Conclusion
- Thanks for watching

■ train data

Table 1: The head of the train data

	cuisine	id	igredients
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	igredients
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 feature extraction](#)

[CountVectorizer](#)

[Modeling](#)

[Conclusion](#)

[Thanks for watching](#)

Text feature extraction



CountVectorizer

Problem Statement
Text feature extraction
CountVectorizer
Modeling
Conclusion
Thanks for watching

By using CountVectorizer to transform the text data to be the word frequency matrix. And we can use `toarray` to help us.

The output example:

```
array([[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],
       [0,0,0,...0,0,0]],
      )
```





[Problem Statement](#)

[Text feature extraction](#)

[Modeling](#)

[Conclusion](#)

[Thanks for watching](#)

Modeling



Modeling

- [Problem Statement](#)
- [Text feature extraction](#)
- [Modeling](#)
- [Conclusion](#)
- [Thanks for watching](#)

- problem analysis

This is a text classification problem.Machine learning has many ways to solve text classification problems.

- Random forest

- SVM



Random forest

Problem Statement
Text feature extraction
Modeling
Conclusion
Thanks for watching

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





SVM

Problem Statement
Text feature extraction
Modeling
Conclusion
Thanks for watching

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





The score of Random forest and SVM

[Problem Statement](#)

[Text feature extraction](#)

[Modeling](#)

[Conclusion](#)

[Thanks for watching](#)

Table 3: The score of 2 model

	model	score
1	Random forest	0.71
2	SVM	0.79



Forecasting

[Problem Statement](#)

[Text feature extraction](#)

[Modeling](#)

[Conclusion](#)

[Thanks for watching](#)

By using the above model, the prediction classification results can be obtained.

The output is:

```
array(['southern_us', 'southern_us', 'italian', ..., 'italian', 'southern_us', 'mexican'], dtype=object)
```



[Problem Statement](#)

[Text feature extraction](#)

[Modeling](#)

[Conclusion](#)

[Thanks for watching](#)

Conclusion



Conclusion

Problem Statement
Text feature extraction
Modeling
Conclusion
Thanks for watching

Text feature extraction Using the CountVectorizer and TfidfVectorizer to help us process the text data.If the text data is Chinese, you can use jieba for word segmentation.

Modeling There are many ways to deal with text classification in machine learning Can be appropriately selected on combination with the problem.

Prospecting I woule like to select multiple models for comparison later.For example, Naive Bayes, there are some deep learning methods (RNN, CNN)





[Problem Statement](#)

[Text feature extraction](#)

[Modeling](#)

[Conclusion](#)

[Thanks for watching](#)

Thanks for watching