kaggle competition report

蔡佳靜 (108070027)

outline

- preprocessing steps
- feature engineering steps
- model explaination
- something I tried to do but failed

processing step

First, I generate the features with BOW, TF-IDF and method provided in <helper> document.

Then, I trained the features I got by two different ways, Decision tree, Naive Bayes.

Latter, I train the feature with best accuracy by Deep learning.

After that, I increased the letters I trained and repeat the process above.

feature engineering step

First, I generate the features in basic way.

```
BOW_500 = CountVectorizer(max_features=500, tokenizer=nltk.word_tokenize, stop_words= "english")
```

Then I generate the features with stop words which provided in nltk libaray.

```
BOW_500 = CountVectorizer(max_features=500, tokenizer=nltk.word_tokenize, stop_words= "english")
```

I had tried to get rid of some specific comma, however, I failed to do this.

model explanation

when I train the model, I observe something. With the stop words, I get the better accuracy with TF-IDF features. However, after deep learning, the accuracy is not that perfect, the progress is less than 1%. Moreover, I observe that the accuracy we get by TF-IDF and naive bayes is better. With deep learning, we get the better accuracy, but the progress is small. The reason can be attribute to the missing of the answer. I have tried to increase the epochs, however the progress is limited.

something I have tried but fail

first, I tried to eliminate the punctuation marks, but when I tried to deal with the features, I faced some errors caused by the data type.

So I tried to work on different way to get the feature, but the word2vec model can’t get the prediction but the relationship between the words.

then I tried to do the bert analyze, but my computer can’t run the data.