Text Categorization Project

Bonny Wang

ECE467- Natural Language Processing

Spring 2021

Instructions

In order to run the program, nltk and sklearn library is needed besides the standard library from python3.The following command can be used to install this library:

```
pip install -U nltk
pip install -U sklearn
```

To run the program, simply type py/python command and enter the name of the training, testing and output file name. A sequence of sample commands are shown below:

```
>>py main.py
Please enter the name of the list of labeled training documents:corpus1_train.labels
Training Completed!
Please enter the name of the list testing documents:corpus1_test.list
Finished predicting the category for documents!
Please enter the name of the output file:corpus1_predictions.labels
```

Explanations

This project mainly used TF*IDF value to process the data. First, tokens are extracted from each document by the tokenizer and stemmer from nltk libaray. Then the TF*IDF value is calculated for each token in each document. By summing the TF*IDF from the all the documents in the same category, the weighted model is created. In addition, when calculating the term frequency, arbitrary weight of 1 is chosen, by experimenting, to be added for better performance.

The similarity between the predicting document and the weight matrix of tokens is measured by the cosine similarity metric, which is calculated by the dot product.

The second and third dataset is evaluated by separating part of the training data into the testing data.

The SVM with different kernels from the sklearn library is also experimented in this project. However, there is no significant advantage by using them from the results obtained. Therefore, they are not chosen for the ultimate version of the program.

There are further experiments can be performed for this project. Naive Bayes method can be alternatively used for this project. More experiment towards different tokenizer and tuning weights for specific tokens and categories may also improve the result. If the datasets get larger, machine learning methods include deep learning and neural networks might be cost effective.