CENG 563 Assignment 1 Report

Burak Tokak - 2533610

Implementation Details

I was supplied with a dataset containing book titles and their description from Goodreads website, which was divided into their respective book categories. I have first read and parsed these text files into a megadoc coupled with the category slug.

I have done some preprocessing on these tuples and using the train set I have trained a Naive Bayes Classifier and a Support Vector Classifier, and using the (also preprocessed) test dataset I have calculated accuracy of both of these classifiers. This report contains details on the methods I used while preprocessing text, feature selection and accuracy performance of the classifiers.

Preprocessing

Explain basic text processing operations you have applied. Why do you think they are required/effective/beneficial for the given task?

- First of all I combined the book title and description on the same training and test elements. Since the title and description will be fed to the classifier together. This will make the future task easier. (Concatenated title and description into same string)
- Striped all extra white spaces for better accuracy
- Remove all numbers in the strings, since the numbers have no correlation for the book's category
- I turn all letters to lowercase since our goal is to classify depending on topic
- Expand contracted words. (can't -> can not etc)
- Remove punctuations
- Remove the stop words (on, then, is, am) since they don't have any contextual information
- Instead of stemming words, I lemmatized them with wordnet article database
- Remove single letter words in the text like: j. f. kennedy => kennedy

Feature Selection

Explain your feature selection. Why do you think your features are required/effective/beneficial for the given task?

I decided to use "bag of words model" for feature selection, which will extract contextual information with the frequencies of the words used in the book title and description combination.

I opt-out of using 2-gram or n-gram bag of words model instead of 1-gram since the description paragraphs are considerably short and using even 2-gram model would be hard to have frequent word groups.

Though I would imagine, if I had more time and computing power, using 1-gram and 2-gram bag of words model together would be a better feature selection and yield better results in terms of accuracy.

Classifier Performance

Evaluate performance of your classifiers. Give their accuracies, recalls, precisions and F1-scores on each genre and on the overall test set. Comment on those.

```
Here is the output of the
Train Megadoc length:  6536
Test Megadoc length:  1865
Training...
Testing...
NaiveBayes Classifier Accuracy:
0.7152815013404826%

Training...
Testing...
SVC Classifier Accuracy:
0.710455764075067%
```

Compare performances of NB and SVC classifiers. Did one of them significantly outperform the other? Why? Is this result expected?

In my case NB and SVC classifier accuracies were very similar, Naive Bayes giving just a little better results.

Additional Notes

Training and testing for both classifiers takes about 20 minutes on Macbook M1 Pro. It might seem like its hanging when the console prints the megadoc length and then it prints “Training…” and “Testing..”.

Turn `load_from_pickle` global variable into 1 in order to load the trained classifiers from drive.