Zhihan Li
DS_GA 1011 HW1
10/09/2018

# Analysis of the Results for Bag of N-Gram Document Classification

The public github repo can be found at:
https://github.com/Evaaaaaaa/NLP_HW1/blob/master/ZHIHAN%20LI_HW1_NLP.ipynb

## 1. Methods

Under each of the recommended hyper-parameter categories, I have tested

1. Tokenization schemes of the dataset:
   1.1 Convert all words to lower cases and ignore punctuation.
   1.2 Pre-process by deleting line break '< br />' and parentheses.
   1.3 Remove web URLs using regular expression.
2. Model hyper-parameters:
   2.1 Unigram
   2.2 Bigram
   2.3 Trigram
   2.4 Quadrigram
   2.5 Vocabulary size
   2.6 Embedding size
3. Optimization hyper-parameters:
   3.1 SGD optimizer
   3.2 Adam optimizer
   3.3 Change learning rate
   3.4 Use linear annealing of learning rate with original learning rate to be 0.01

# 2. Result and Analysis

## a. Accuracy Table

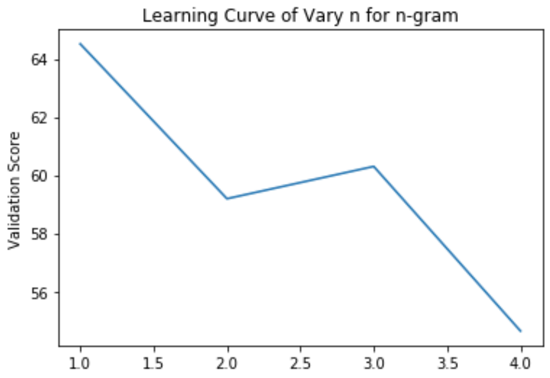
In the ablation study, I set all other variables to constant while allowing one to vary. The values for constant parameters I used are

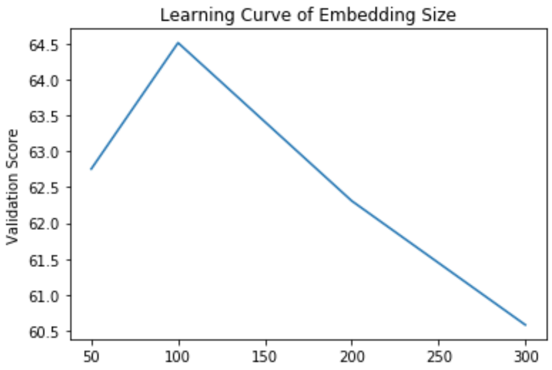
Tokenization schemes = the combination of all three pre-processing methods
n for n-gram = 1
Vocabulary size = 10000
Embedding size = 100
Optimizer = Adam
Learning rate = linear annealing

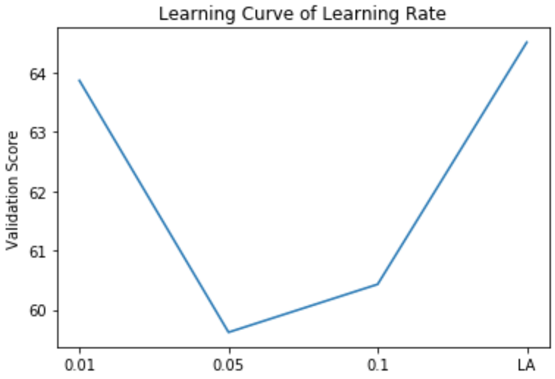
The validation accuracy of each parameters with selected values is shown below:

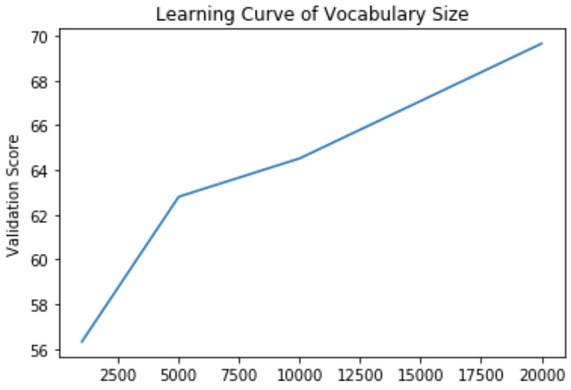
| Category | Comparable attribute | Validation Accuracy |
|---|---|---|
| Vary n for n-gram | N=1 | 64.515 |
| | N=2 | 59.19 |
| | N=3 | 60.3 |
| | N=4 | 54.64 |
| Vocabulary size | max_vocab_size = 1000 | 56.32 |
| | max_vocab_size = 5000 | 62.8 |
| | max_vocab_size = 10000 | 64.515 |
| | max_vocab_size = 20000 | 69.65 |
| Embedding size | emb_dim = 50 | 62.755 |
| | emb_dim = 100 | 64.515 |
| | emb_dim = 200 | 62.31 |
| | emb_dim = 300 | 60.58 |
| Optimizer | SGD | 65.91 |
| | Adam | 64.515 |
| Learning rate | learning_rate = 0.01 | 63.87 |
| | learning_rate = 0.05 | 59.62 |
| | learning_rate = 0.1 | 60.43 |
| | Linear annealing | 64.515 |

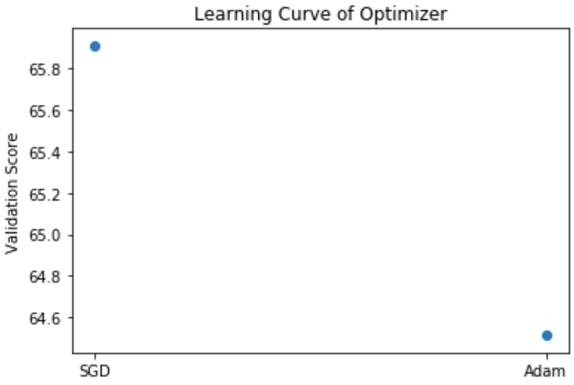
## b. Training Curves


Learning Curve of Vary n for n-gram


Learning Curve of Embedding Size


Learning Curve of Learning Rate


Learning Curve of Vocabulary Size


Learning Curve of Optimizer

## 3. Discussion

Validation set samples:
3 correct: 8367_10.txt, 11778_3.txt, 9006_7.txt
3 incorrect: 42107_0.txt, 2582_2.txt, 12249_8.txt

## 4. conclusion

The best set of parameters I have found is:
Tokenization schemes: the combination of all three pre-processing methods
n for n-gram: n = 1
learning rate: Linear annealing
Vocabulary Size: 20000
Embedding Size: 100
Optimizer: SGD

The validation accuracy is 69.65
The test accuracy is 67.58