Sentiment analysis with the perceptron report

Ziyang Li

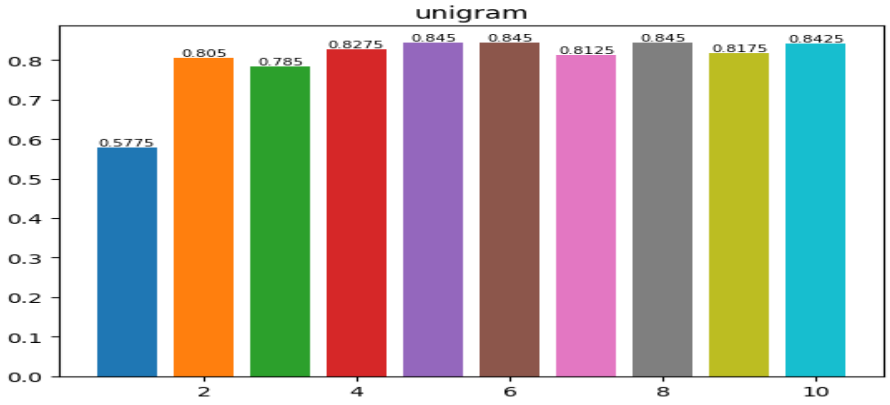
1. Feature choice

In this lab, I select bigram and tf-idf beyond bag-of-words as features. Also, the training examples have been updated so I don't need to divide them into train and test sets.

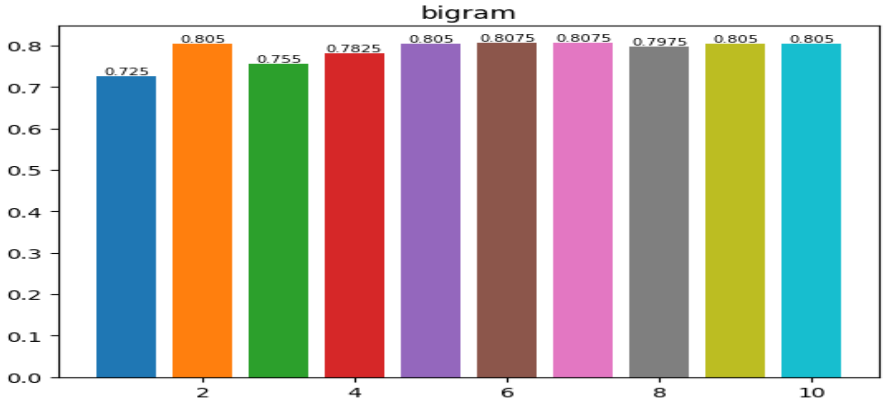
2. Accuracies of 3 features

In the process, the iteration runs ten times on each feather. Furthermore, I found that the value of seed in random function can affect accuracies which means that the order of all the documents has influence on the training results.

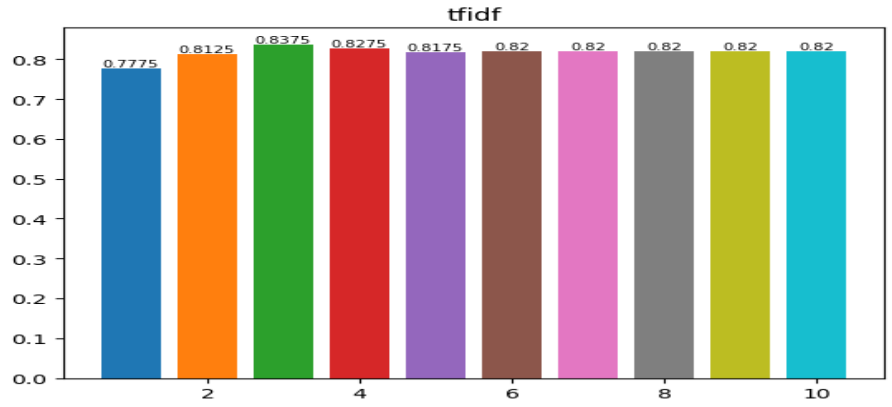
Different accuracies on each iteration are as follow:



Accuracy of average weight vectors of unigram is 0.84.



Accuracy of average weight vectors of bigram is 0.8125.



Accuracy of average weight vectors of tf-idf is 0.8225.

The top 10 most positive features in bag-of-words

Positive	Weight	Negative	Weight
'trek'	111.2	'if'	-117.9
'life'	113.4	'director'	-125.1
'movies'	114.8	'boring'	-126.8
'also'	118.1	'no'	-134.7
'jackie'	118.9	'why'	-153.3
'see'	127.1	'script'	-160.8
'well'	133.8	'worst'	-163.8
‘seen'	135.8	'plot'	-196.8
'great'	139.2	'only'	-199.0
'most’	146.3	'bad'	-309.3

Table 1

3. Discussion

The reason why I select others 2 features is that bigram contains more information of data and tf-idf is useful in information retrieval. Tough they were used in the model, but both of them doesn’t improve the accuracy compared with bag-of-words.

According to table 1, I do not think all of the weights make sense. In the positive side, ‘trek’, ‘life’, ‘movies’, ‘jackie’ are not useful for judging positive sentence, also, ‘director’, ‘script’ do not make sense in negative sentence. There are lot of weighted word related to the topic of movie, so, I do not think the classifier we learnt would generalize well if we apply it to a different domain.

The better feature for the task is word with label, so the part of words can be recognized by computer, such as noun, adjective or adverb. It is common that the sentiment of a sentence is determined by people based on its adjective, adverb etc. If we can use this feature to implement more effective model, it could be useful and effective.