Text Classification with the perceptron

February 18, 2019

Registration Number : **180128033**

In this report, Binary Perceptron has been used to calculate the accuracy and errors of Bag of words.*Bigram and Trigram* are the feature types beyond bag-of-words.

1 Standard Binary Perceptron

While running standard binary perceptron with and without shuffling data, we got below results for Unigram, Bigram and Trigram. Below given result are produced on seed=12345.

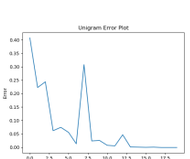
1.1 Result:

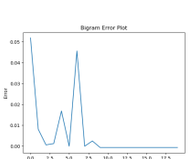
Ngram	**Accuracy w/o Shuffling**	**Accuracy w Shuffling**
Unigram	0.5025	0.7825
Bigram	0.5025	0.765
Trigram	0.5075	0.7525

2 Binary Perceptron with Iteration

To train algorithm on training data and test it over testing data, we have taken 1600 (positive and negative document) in x_train and left 400 in x_test.After dividing data, we have iterated it 20 times to train the model created using algorithm,below are the results on seed=12345.

2.1 Result:

[Unigram] 

[Bigram] 

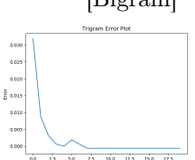
[Trigram] 

Figure 1: Learning rate for *Unigram*, *Bigram* and *Trigram*

- On the basis of result from above table, we can conclude that accuracy decrease in *Trigram* and *Bigram* in comparison to the *Unigram*.

Ngram	Average Accuracy	Length of weights
Unigram	0.8625	40300
Bigram	0.8175	432588
Trigram	0.75	843298

- Below are the top 10 features of each class:
- **Unigram:** Top 10 Positive words 'great','well', 'seen','quite','see','job', 'jackie', 'most','also','best'. Word like 'great', 'best'.Top 10 Negative word 'bad','worst','only','plot','any','script','even','supposed','boring','no'.
- **Bigram:** Top 10 Positive word 'black cauldron"but it','the black','the best','even if','sense of','a little','and for','is an','man who'. Top 10 Negative word 'have been','should have','supposed to','the worst','s the','is so','to work','but this','like a','to have'.
- **Trigram:**Top 10 Positive words 'the deep end','a man who','of the best','the most part','due to the','up to the','the voice of','isn t as','a lot more','s life is'. Top 10 Negative word 'of the worst','supposed to be','to work with','s a','wouldn t have','problem is that','to be an','the phantom menace','the movie and','could have been'.
- These features will not generalise well for laptop reviews or restaurant reviews.The better feature for new domain should more focused on qualities of food such as yummy,delicious etc and for laptop such as high speed, ram etc. .