COL341: Assignment 5

Nihar Patel

2017MT60780

I experimented with the following text pre-processing :

- Part-a:

Experiment	Accuracy
Just split the words on white spaces	83.965
Convert all to lowercase	83.94
Remove punctuations and special characters	83.695

- Part-b:

Experiment	Accuracy
Remove stop-words	85.005
Remove stop-words and Porter stemmer	84.485
I also tried Snowball stemmer and experiments from part-a but accuracy was not improved	

- Part-c:

Experiment	Accuracy
Feature engineering- use bigrams	85.09
Feature engineering- use trigrams	50.07
Feature engineering- use bigrams and unigrams	86.6975
Unigrams and bigrams without stemming	**86.98**
I tried unigrams, bigrams and trigrams together but it gave 'out of memory' error	

Remarks:

- The accuracy after stemming seems to reduce here. This may be because of less training data resulting in small vocabulary for Naïve Bayes classification.
- Removing stop words increased accuracy by >1%, further using bigrams alongside with unigrams resulted in an increase in accuracy by around 3%