**Problem Set 2 – Write-up**

The following is a short summary of the approach used for the NLP assignment:

* The 10 text documents were first downloaded using the URLs provided.
* The NLTK package [1] in python was used to perform most NLP tasks.
* The stop words were removed from each file using the “stopwords” library inside nltk.corpus [2]. The following unique stop words were removed:

'and', 'all', 'being', 'is', 'both', 'it', 'an', 'as', 'through', 'at', 'during', 'in', 'further', 'its', 'from', 'for', 'their', 'there', 'when', 'same', 'any', 'to', 'other', 'which', 'between', 'has', 'into', 'more', 'be', 'over', 'above', 'than', 'that', 'after', 'they', 'such', 'with', 'by', 'a', 'on', 'about', 'off', 'of', 'against', 'or', 'will', 'while', 'so', 'can', 'each', 'the', 'where', 'having', 'are'

* The stemming was performed using the implementation of the Porter Stemming algorithm made available through the nltk.stem.porter library [3].
* The n-grams were extracted using the “ngrams” method under the nltk.util library [4].
* All n-gram frequencies were then stored as dictionaries [5].
* The csv library in Python was used to export these dictionaries to csv files [6].
* The “cp” command was used from AWS CLI as follows to upload the files to the S3 bucket [7].

C:\Users\Anupama>aws s3 cp C:\Users\Anupama\OneDrive\INFX575\Assignment2\counts

s3://anupama-garimella-bucket –recursive

References:

1. <http://www.nltk.org/>
2. <http://www.nltk.org/_modules/nltk/corpus.html>
3. <http://www.nltk.org/howto/stem.html>
4. <http://www.nltk.org/_modules/nltk/util.html>
5. <http://stackoverflow.com/questions/2161752/how-to-count-the-frequency-of-the-elements-in-a-list>
6. <http://stackoverflow.com/questions/8685809/python-writing-a-dictionary-to-a-csv-file-with-one-line-for-every-key-value>
7. <http://docs.aws.amazon.com/cli/latest/reference/s3/cp.html>