# *Natural Language Processing*
# *CAP6640 – spring 2016*
# *Homework 1*

### *1.1.2 Text Classification*

***Part 1:***

After preprocessing the documents given, the number of unique words in the training dataset including all the categories was found to be ***23080***, the screenshot of the same has been shown below.

*Screenshot:*

```
In [161]: runfile('C:/Users/Syed/Google Drive/NLP/Assignment1/Q2/text_pre_pro.py',
wdir='C:/Users/Syed/Google Drive/NLP/Assignment1/Q2')
Reloaded modules: sentiment_reader, multinomial_naive_bayes, linear_classifier
The number of unique words in all 3 categories of training dataset is: 23080
```

***Part 2:***

Performance of the model was tested and the following macro-averaged f score was found:

***97.92%.*** The screenshot of the same has been provided below for your reference.

*Screenshot:*

```
In [163]: runfile('C:/Users/Syed/Google Drive/NLP/Assignment1/Q2/run_classifier.py',
wdir='C:/Users/Syed/Google Drive/NLP/Assignment1/Q2')
Unique words in training set: 23080
Macro-averaged f1 score for the testing dataset is: 0.979187
TIme taken to execute the program is: 36.000000 seconds
```

***Experimental setup:***

*Environment*

- Programming language used: Python
- Version : 2.7.11 64 bits

- IDE used: Spyder
- Version: 2.3.8

*Datasets*

The given newsgroup datasets consisted of the following:

| | Type of dataset | Category of files | | | Total |
|---|---|---|---|---|---|
| | | Graphics | Autos | Guns | |
| 1 | Training | 584 | 594 | 546 | 1724 |
| 2 | Testing | 389 | 396 | 364 | 1149 |

A multinomial naive Bayes classifier for text classification was designed and implemented. Before given the data to the classifier pipeline, the contents of these files were preprocessed. The preprocessing paradigm involved the following steps:

- Sentence segmentation: The content of each of the given files was broken into sentences and moved forward in the pipeline.
- Special characters like!,@,#,$, %,^,&,*,?,_,~,-,£,(,) etc., were removed.
- White spaces were removed.
- Word Tokenization: The sentences were divided into sequences of tokens such that each of these tokens corresponded to a word
- Lemmatization: The different inflected forms of the words resulting from the previous step were grouped together.
- All the letters were converted to lowercase
- Finally, the all the stop-words were removed.

The following libraries were used to implement the above mentioned steps:

- nltk.corpus
- nltk.tokenize → Natural Language Toolkit
- re → Regular expression
- PlaintextCorpusReader

It was found that there were addition stopwords in the given documents and hence the default stopword removal library was updated and then used.

The given files were preprocessed and stored in the data files in the following format. One data file for each category and dataset was created (eg. Graphics.test, guns.train etc., ) and each line in each of these files had words and the frequency of the respective word separated by " : "sign. The last token in each line indicated the category of the document.

Finally, the given started code was modified to handle three categories and the classifier was run. The split function used in the previous question was not incorporated here since the

training and testing datasets were already separated. I also implemented a function to calculate the F score for each category and the using that the Macro averaged F1 score was calculated.

*F1 score for each category:*

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

*Micro-averaged f1 score:*

Sum ($F_1$ for each category)/3. Since we have just three categories.

***Time taken for the process:***

Time taken to build the data files in the required format so as to use it with the given starter code:

***89 seconds.***

```
In [166]: runfile('C:/Users/Syed/Google Drive/NLP/Assignment1/Q2/text_pre_pro.py',
wdir='C:/Users/Syed/Google Drive/NLP/Assignment1/Q2')
Reloaded modules: sentiment_reader, multinomial_naive_bayes, linear_classifier
The number of unique words in all 3 categories of training dataset is: 23080
TIme taken to execute the program is: 89.000000 seconds
```

Time taken to run the classifier after modification:

***36 seconds.***

```
In [163]: runfile('C:/Users/Syed/Google Drive/NLP/Assignment1/Q2/run_classifier.py',
wdir='C:/Users/Syed/Google Drive/NLP/Assignment1/Q2')
Unique words in training set: 23080
Macro-averaged f1 score for the testing dataset is: 0.979187
TIme taken to execute the program is: 36.000000 seconds
```

Total running time: 125 seconds.