**

COMP814 – Text Mining

Pre-processing I Lab

# Objective

1. To be able to use tokenisation and other simple nltk libraries to do basic pre-processing.

# Task

1. You will need the sample code snippets from lectures for the following tasks.
2. Run the demo code to ensure that it runs error free.
3. Google “*python read all text files in directory*” and figure out how to read in multiple text files from a directory and store the texts in a variable.
4. Read in all the files from the data file, terrorism data.7z, from Canvas.
5. Use the demo code to find bigrams.
6. Again google how to “*count and sort objects”* in a list in python.
7. **Task description:**
   1. Clean the text using the techniques from lectures. You should remove punctuations, stop words, use stemming or lemmatization etc. Leave the numbers.
   2. Find trigrams of tokens for the all the articles in the dataset.
   3. Determine the **most common** trigram.
   4. See if you can figure out what if any useful information this gives you. How can you modify the strategy to extract the “most common topic” in the corpus. Discuss your strategy with others in the class to see the differences and similarity with your one.
8. **Challenge Task** (not required to finish): Extract the two most common bigram and then extract the 4 grams containing the two most common bigrams. See if you can cluster them into clusters to extract the most common topic.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*End of Lab\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*