COMP814 Text Mining
Assignment 2 (60%)

Assignment

To be done in pairs. At least one member from the pair needs to have reasonable programming competency. You can work with the same person as assignment 1 or with a new person.

Objective

To be able to carry out sample of NLP pre-processing tasks, and use the results for a higher level information extraction. The results should then be analysed and presented as an appropriately formatted scientific paper.

Task Resources

You will be using models that you developed as part of labs in the python environment. You will use the dataset provided on Blackboard as a zipped file named Assignment2Data.7z. The data is from Reuters Corpus Volume 1 (RCV1). The corpus contains a huge collection of classified news articles, however we will use a subset of the corpus. More information at: http://www.daviddlewis.com/resources/testcollections/rcv1/

Task Specification

1. Your data is a set of 2500 text files which contain newswire articles chosen from the CCAT, corporate and industrial category.
2. The overall task for you to do is to determine the **top 5 most frequent organisations that appeared in the news and extract why they were in the news.**
3. To achieve the objective you are required to pre-process the texts in various ways in order to achieve the overall objective.
4. As a minimum you are required to do POS tagging and NER.
5. You can use either the pre-trained model from nltk or other python libraries or train your own model.
6. Summarise your results into an appropriate format so that it can be consumed by a journalist looking to write an article on the activities of top 5 organizations.

Write up

1. You need to document the research project as a scientific paper using latex double column IEEE conference format. The latex template can be downloaded from Blackboard.
2. You should use a minimum of 12 and a maximum of 15 pages excluding the references and appendices.

3. You should also submit a well commented and formatted python code as part of the appendix.

4. The paper should describe:

   a. The task you set out to solve.
   b. A literature review of same or similar tasks attempted by other researchers.
   c. The steps you took and the experiments you did to solve the problem and the rationale for them.
   d. How you ensured the accuracy of your results.
   e. The conclusion and how you would do the task differently if you were to do it again. In this section you should also summarize the contribution made by each member of the team (maximum of half a page each).

Assessment

This assignment contributes **60% towards your course grade**.

You are required to demonstrate a collaborative strategy to solve the problem. The evidence for this will come from your online discussion in the form of blogs on the topics and strategies developed to solve the problem. Use the discussion forum on Blackboard to look for a partner to work with or look for one in the class. The group should then enrol in one of the groups created for assignment 2 collaboration. You can then use any of the tools (such as file sharing) in the forum but use the **discussion tool** to collaborate and collect evidence for the purpose of the assignment.

Note the following about the discussion:

a. It needs to be sustained over the whole assignment period.
b. It should be constructively collaborative. Each partner should reference the other partners work and reply to questions/assertions made by presenting material/ideas that support or refute what the other partner has said.
c. You should consult multiple sources to support your assertions.
d. Avoid pasting large chunks of text in the blogs. It should mainly be your own words with links as references. The purpose of the blog is to show how you progressively solve the problem by doing research and developing strategies based on sound rationale. Your blog should be in conversational mode with your partner as you affirm and refute ideas and strategies developed to solve the problem.
e. Approximate marking scheme.

Part of Assignment	**Mark**
Research question and rationale description	10
Data description and analysis	15
Research Design	20
Implementation (code) submitted as appendix	15
Analysis and Evaluation	20
Conclusion and formatting and references	10
Collaboration blog	10
Total	100

Due Date

The assignment is due in week 14, Friday 7 June at midnight.
The assignment should be submitted via the Assignment 2 submission Link under Assignment 2 folder on AutOnline. Only one partner needs to submit the assignment.

Treat this as a learning experience rather than an assessment exercise.

******************************* Good Luck *******************************