COMP814 Text Mining
Assignment 2 (60%)

Due Date

The assignment is due 26 June at midnight.
The assignment should be submitted via the Assignment 2 submission Link under Assignment 2 folder on Blackboard.

Assignment

This is assignment can be done either individually or in pairs. If you are new to coding then it is recommended that you pair up with someone with some coding skills. Only one person from the pair needs to submit the assignment.

If you are doing the assignment as a pair you should write approximately 12 pages and if you are doing it individually you should write approximately 6 pages, both excluding the references and appendices.

Objective

1. To be able to carry out a typical text mining task based on an objective.
2. To document the methodology and the findings in an appropriately formatted scientific paper suitable for publication in a conference.

Task Resources

You will be using models and code snippets that you developed as part of labs in the python environment. You will use the dataset provided on Blackboard as a zipped file named Assignment2BlogData.7z.

Your dataset consists of a set of 19,320 xml formatted text files. These files contain blogs collected from an anonymous blogging site which have been annotated with various types of anonymised metadata. The metadata has been integrated into the filenames. The text in each of the files contains the blogs corresponding to a blogger (as described in the metadata) with blog dates ranging from approximately 2001 to 2004.

Task Brief

You are employed by an innovation company who has bought the blogs with the objective of innovating new products/services based on what people have been talking about on popular blog sites.

In particular your boss wants to know the two most popular topics that the bloggers have been talking about in the following demographics :

a) Males

b) Females
c) Age brackets <=20 and over 20.
d) Everyone

Task Requirements

In order to achieve the objectives of the project, you will firstly need to read in the data, extract the meta data and segment it into the required demographics.
You will then need to design strategies to extract and cluster topics.

To be consistent within the class, let us use the same definition for a topic. Let us define a topic to be the mention of an OBJECT or a THING. So, for instance you could simply take the THING that is mentioned as the highest number of times as the popular topic and correspondingly the second most popular topic.

Once you get the two most dominant "things" mentioned, expand the topic to be 2 verb/noun before and 2 verb/noun after the topic. Output them as "what has been said about the dominant "thing" in terms of the 4 surrounding nouns/verbs.

Repeat what you done using frequency above, but this time use TFIDF. For this consider all the blogs from one person as a document.

Compare the results from the two modes of counting and comment on which one is more accurate in your opinion with justifications.

Note that you will need to use various techniques such as stemming, lemmatization, PCA, stop word removal, inter alia, in order to get as accurate results as possible. The results will need to be evaluated manually and the strategy for evaluation should be described in your writeup.

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

3. Your 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 details of your strategy to solve the problem. In this part you should describe the details of how you processed the data from start to finish including the details of how the data got processed in any external library you have used (if you have used it).
   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.

Assessment

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

Approximate marking scheme.

Part of Assignment	**Mark**
Research question and rationale description	10
Data description and analysis	15
Research Design	30
Implementation (code) submitted as appendix	15
Analysis and Evaluation	20
Conclusion, formatting and references	10
Total	**100**

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

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