COMP814 Text Mining
Assignment Specification

Due Date

The assignment is due on 7 June at midnight. This is the Friday of the 13th week of AUT semester.
The assignment should be submitted via the Assignment submission Link on Canvas.

Assignment

This is assignment must be done in pairs to develop teamwork skills as well. 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. You should have the 2 authors names as well as the ID as indicated in the IEEE template.

The assignment should be written up in a maximum of 12 pages excluding reference 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. The format of paper is given as a Latex template file.

Task Resources

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

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) Students
e) Everyone

The dataset contains 19,320 blogs. If this happens to be a size that your computer takes too long to process, the task is such that you can segment the dataset and process it in batches. Also, this is a raw dataset so you might have to deal with noise in the data, such as taking care of non-ascii characters.

The table below gives the number of files in each of the demography to be analysed for this assignment.

Demography	**Number of Files**
Male	9661
Females	9661
Age over 20	10936
Age under 20	8241
Students	5121
Everyone	19321

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.

We could use a few different strategies to extract the dominant topic in a corpus. Some examples are :

- Count all types of nouns.
- Count all clauses and then extract the subjects.
- Count all Subjects, Direct Objects and Prep Objects.
- Count all types of nouns and their modifiers.
- Count all subjects and direct objects.
- Count all objects that participate in actions.
- Etc.

We could also use TFIDF instead of counting as this will also take rare terms into account.

Minimum Requirements For the Assignment

1. Use at least 2 different strategies to extract the 2 most dominant topics.
2. As an attempt to decipher what was said about the dominant topics, output the clauses containing the 2 dominant topics.
3. As a third variation, instead of counting, use TFIDF vectorization to choose the two most dominant topics. Ensure that your TFIDF strategies are consistent with step 1.

4. Do step 2 with TFIDF as vectorizer.

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 Canvas.
2. In order to run your program by just clicking it, you need to share the script file with unrestricted privilege to everyone as you have been doing for lab submissions. Include this link in the appendix.
3. In order for your script be able to read the datafile you can either share the data file and read it as a url using *pydrive* package which is a bit complex, or, you can read the datafile from a directory path which is pre-specified. When I run it I can keep the datafile in the pre-specified location so I will be able to run your program by reading the same datafile from the same specified on my google drive. To do this, use the following:
    a. Include the following code as the first cell in your program.

    ```
    from google.colab import drive
    drive.mount('/content/drive')
    ```

    b. On your MyDrive, create a directory called '*COMP814Data*' and then put your datafile '*AssignmentBlogData.zip*' in this directory.
    c. Use the following sample code to read the zip file contents into a string.

    ```
    import zipfile

    z = zipfile.ZipFile('/content/drive/MyDrive/COMP814Data/AssignmentBlogData.zip', "r")
    import zipfile

    output = ''
    with zipfile.ZipFile('/content/drive/MyDrive/COMP814Data/AssignmentBlogData.zip') as z:
        for filename in z.namelist():
            with z.open(filename) as f:
               output += f.read()
    ```

4. Your paper should describe:

    a) A brief section on the contribution from each member of the pair labelled "Contributions".
    b) The task you set out to solve.
    c) A literature review of same or similar tasks attempted by other researchers.
    d) 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). This should explicitly describe the strategy used to extract the 2 most popular topics.

e) How you ensured the accuracy of your results.

f) A comparison of the results from the two topic extraction strategies and the two modes of counting. Comment on which one is more accurate, in your opinion, with justifications.

g) 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, language and references	10
Total	**100**

Treat this as a learning experience rather than an assessment exercise. The assignment is constrained enough for you to do exactly as specified and pass it, and flexible enough for you to be creative and attempt to get more accurate results, hence a better grade.

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