COM4513 Natural Language Processing

Assessment I: Open Domain Topic Detection

Note: This assessment accounts for 50% of your total mark of the course.
Information for Plagiarism: The source code and your report may be submitted for plagiarism check (e.g., Turnitin). Please also read the following information provided by the university: https://www.sheffield.ac.uk/ssid/unfair-means/index

Quick Summary

1. In this assessment, your task is to perform open domain topic detection and analysis on a given dataset.
2. You will write a report to present your findings and results on the task. You will complete this project **individually (no team work)**.
3. The structure of your report should follow the order of the questions with appropriate subsection titles.
4. Submit your code (if any) and any associated files along with your report. If you use Python, please submit the Jupyter Notebook containing both the Python scripts and the running results.

For any clarifications on this assessment, please email **Dr. Chenghua Lin** (c.lin@sheffield.ac.uk).

Tools and Programming Languages

You are free to use any programming languages or tools. You are also free to use any machine learning packages available online, but you must acknowledge them.
If you use open-source code, you must point out where it was obtained (even if the sources are online tutorials or blogs) and detail any modifications you have made to it in your tasks. You should mention this in both your code and report. *Failure to do so may result in zero marks being awarded on related (sub)tasks*.

Marking Criteria

1. Quality of the report, including structure, clarity, and brevity.
2. Quality of your experiment, including model design and implementation.
3. Quality of result analysis and evaluation.
4. Quality of the source code, including the documentation of the code.

Dataset Description

Back in 1985, the BBC initiated an ambitious project called the Domesday Project, which aims to documenting the daily life of people in the UK. Over a million of people took part in this project. You will be provided with a cleaned version of the Domesday dataset consisting of 120K+ descriptive articles written by people lived across the UK (i.e., England, Scotland, Wales and Northern Ireland). These articles cover a wide range of topics such as education, sports, unemployment, etc. Geographical information is also provided for each article, i.e., the city/town where the article was written. The dataset can be download from the course area of Blackboard.

Task Specification

1. *Dataset analysis*:
    a. The dataset is in the XML format. You will need to extract the 'title' and 'content' for each article for further processing. You may also consider other pre-processing steps such as stop word removal and lemmatisation. [5 pts]
    b. Analyse the characteristics of the dataset, e.g., the distribution of articles across different countries, average article length, average number of sentences per article, vocabulary size, etc. [5 pts]
2. *Topic Detection*: design and implement methods to extract the topics covered in the Domesday dataset, e.g., using the topic model.
    a. Design the topic detection algorithm/framework. [15 pts]
    b. Implementation of the topic detection algorithm/framework. [10 pts]
3. *Result analysis and evaluation:*
    a. What are the top 15 most prominent topics of the whole dataset? [5 pts]
    b. What are the key topics that are shared across different countries and what are the topics that are unique to each of the counties? [10 pts]
    c. What are the top 10 most prominent topics of the articles from the shires/counties of your choice based on the topic modelling results? [5 pts]
    d. To conduct more fine-grained analysis, you will be given a list of 100 articles for the shires/counties of your choice (50 each), and then for each article, (i) manually examine the topics discussed and tag each article with one or more topic tags (e.g., job, sports, strike, etc). Please use this topic tag set and can add more tags as needed; (ii) label each article with one or more of the following sentiment tags (i.e., positive, negative, and neutral). [15 pts]
    e. For the articles given in Task 3-d, (i) compare the topics labelled by yourself and the topics detected by the topic model and analyse how well they align; (ii) provide some general discussion regarding the sentiment of the provided articles. [10 pts]
4. *Report writing:* [20 pts]
    a. Provide a brief description of the dataset including the background of the Domesday project and a summary of the dataset statistics.
    b. Describe and discuss your method, experimental setup and results.
    c. The word limit for the report is no more than 2500 words.

Submission Guideline

You should submit a PDF version of your report along with your code via the BlackBoard

by 23:59 5th April (Monday) 2021. The name of the PDF file should have the form “COM4513_Assessment1_< your Surname>_<your first name>_<Your Student ID>”. For instance, “COM4513_Assessment1_Smith_John_XXXXX.pdf”, where XXXXX is your student ID. You should submit your code and any associated files along with your report. If you use Python, please also submit the Jupyter Notebook containing both the Python script and the running results. If you have more than two files to submit, please compress all your files into one “zip” file. Other format of compression files is not recommended.