NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

ASSIGNMENT (Part 1)

Domain Specific Text Data Analysis and Processing

CE/CZ4045 Natural Language Processing

2020/2021 Semester 1

NANYANG TECHNOLOGICAL UNIVERSITY

1 Objective

The objective of this assignment is to let you getting familiar with the main components in an end-to-end NLP application, the challenges faced by each component and the solutions. Through this assignment, you shall also get deeper understanding on various NLP tasks and hands on experiences on packages available for NLP tasks.

2 Assignment Format

1. This is a group assignment. Each group has 4 to 5 students.
2. One report is to be submitted by *each group* and all members in the same group receive the same grade. However, **contributions of individual members** to the assignment shall be *cleared indicated* in the report.
3. You may use ANY programming language of your choice, *e.g.,* Java, Python, C#.
4. You may use any NLP and Machine Learning library/software as long as its license allows free use for education and/or research purpose. Some example packages are listed below. However, you are not allowed to use relational data management systems like MySQL.
   - Example NLP libraries to consider: NLTK (Python), spaCy (Python), LingPipe (Java), Stanford NLP(Java), OpenNLP (Java)
   - Indexing and Search: Lucene (Java)

3 Assignment for First Half (50 marks)

The assignment consists of the following components: Dataset Analysis (20 marks), Development of Summarizer (20 marks), and Application (10 marks).

3.1 Domain Specific Dataset Analysis (20 marks)

You are tasked to form three (3) datasets for analysis. Each dataset should contain about 20 documents in one selected topical domain. Each domain shall have its own linguistic characteristics with some specific terms specific to the domain. Examples are codes in programming forums, mechanical parts in manufacturing, chemical compounds, and maths equations in test papers. Some example domains are given as follows:

- Questions on StackOverflow, e.g., `https://stackoverflow.com/questions/63883827/`
- Patents for Jet Engine, e.g., `https://patents.google.com/patent/US2474359`
- Research papers in medical or chemical areas, e.g., `https://pubs.acs.org/doi/10.1021/jacs.0c07212`
- Life insurance policy document

Tokenization and Stemming. Tokenize all documents in each domain using a selected library (e.g., NLTK) and observe the tokens obtained. Discuss your observations from the following perspectives. Has the tokenizer correctly recognized the domain specific tokens? Use examples to illustrate what the expected tokens are, and what are not. Discuss how to identify the tokens that are incorrect through programs? If you were to improve the tokenizer, what are the possible solutions.

Perform stemming and compare the token distributions before and after the stemming (you may choose any stemming algorithm implemented in any toolkit). You may compare the number of distinct tokens, and the length distribution of the tokens. The length distribution can be compared in a plot: the x-axis is the length of a token in number of characters, and the y-axis is the number of tokens of each length. Discuss your findings.

Sentence Segmentation. Perform sentence segmentation on all documents in each domain. Compare the distribution of the sentence length in the three domains. Here, the x-axis is the length of a sentence in number of words/tokens, and the y-axis is the number of sentences of such length. Discuss your findings.

POS Tagging. Randomly select 3 sentences from each dataset, and apply POS tagging. Discuss the POS tagging results. Are the results as expected? Can the POS tagger well handle the domain specific terms?

3.2 Development of a ⟨ Noun - Adjective ⟩ Pair Ranker (20 marks)

Get about 20 to 30 reviews for one particular product (e.g., reviews of a restaurant in Yelp or reviews of a smartphone in Amazon). Note that, all reviews should be about the same product. Then *manually* go through the reviews and identify the most meaningful 5 pairs of ⟨ noun (or noun phrase) - adjective (or adjective phrase) ⟩ pairs. Example pairs are: ⟨ food-delicious ⟩, ⟨ voice quality - very good⟩, ⟨ battery life - great⟩.

Design and develop a program to identify such ⟨ noun (phrase) - adjective (phrase)⟩ pairs and rank the pairs from all the review documents. Are the top 5 ranked pairs the same as your manually identified pairs? What are the challenges your group has encountered in this task? Note that, the ranking of the pairs is based on all the reviews for this product, not for each individual review.

3.3 Application (10 marks)

Define and develop a simple NLP application based on the reviews collected in Section 3.2. An example application is to classify the sentences into positive, neutral, and negative classes based on their sentiments (you may use any sentiment analysis tools or develop you own sentence level sentiment classifier). You may define your own application with similar (estimated) difficulty level. Note that, application here means a small program to analyse, or to mine the text data. Application here does not mean a web-based application or mobile app. There is no need to develop GUI for this small application.

4 Submission of Report and Source Code

4.1 *Final Report in Hardcopy*

- The hardcopy report must be submitted on or before **2 Nov 2020** (Monday, Week 12), through SCSE General Office. The report shall be formatted following the ACM "sigconf" proceedings templates[1] (either MS Word or Latex), ***maximum 7 pages***, excluding appendix. DO NOT include in your report all the source code and complete results sets. However, you may include *code snippets* or some selected results which are important to discuss. You should cite all third-part libraries used in your assignment.

- The report shall be printed in double-sided format whenever possible. A plastic cover or ring-binding leads to 2% penalty.

- Make sure any words or pictures in the report are **readable**.

4.2 *Final Report in softcopy, Source Code, and Documentation*

- A CECZ4045Part1.zip file containing the following files and folder shall be submitted: Report.PDF, Readme.txt, SourceCode.
  - Report.PDF shall be the same as the hardcopy report submitted.
  - Readme.txt shall include
    - A *link* to download the third-party library if you used any in your assignment.
    - A guide on how to setup your system, and how to use your system (*e.g.,* command lines, input format, parameters).
    - Explanations of sample output obtained from your system.
  - SourceCode folder shall contain all your source code. The libraries shall **NOT** be included in the softcopy submission to minimize the file size.

- Softcopy submission deadline: ***2 Nov 2020 11:59PM***. Late submissions are allowed but will be penalized by 5% every calendar day. The softcopy can be submitted for at most three times, only the last submission will be graded and time-stamped.

[1] https://www.acm.org/publications/proceedings-template