ASSIGNMENT

Review Data Analysis and Processing

CE/CZ4045 Natural Language Processing

2019/2020 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.
   - All-in-one library: NLTK (Python), spaCy (Python), LingPipe (Java), Stanford NLP(Java), OpenNLP (Java)
   - Indexing and Search: Lucene (Java)

3 Assignment (100 marks)

The assignment consists of the following components: Dataset Analysis (60 marks), Development of Summarizer (30 marks), and Application (10 marks).

3.1 Dataset

***
*Please note that the collection of reviews used in this assignment is a sub-set of data released by `https://www.yelp.com/dataset/challenge`. By working on this dataset, you agree with the Dataset License (`https://www.yelp.com/dataset/download`). You are **NOT** allowed to redistribute this dataset in any format.*
***

We will use a collection of user reviews posted on Yelp as the dataset. The dataset was collected from `https://www.yelp.com/dataset/download` with further filtering (to obtain a relatively small dataset). The dataset used in this assignment contains 15,300 reviews for 153 businesses, and the data file is about 11.9 MB uncompressed. A sample data file containing 50 reviews is provided for you to understand the data format. Each review is one line in the JSON file, and each review has the following components: review_id, user_id, business_id, stars, date, text, useful, funny, cool. Detailed explanation of these components for **review.json** is available at `https://www.yelp.com/dataset/documentation/main`.

3.2 Dataset Analysis (60 marks)

Writing Style. Randomly select a few reviews and observe the writing style (*e.g.,* is the first word in a sentence capitalized; Do sentences follow good grammars; are the proper nouns capitalized; etc) in comparison to news articles published by The Straits Times. Discuss your findings.

Sentence Segmentation. Perform sentence segmentation on the reviews and show the distribution of the data in 5 plots, one for each rating star (*i.e.,* 1 to 5). In each plot, the x-axis is the length of a review in number of sentences, and the y-axis is the number of reviews of such length. Discuss your findings based on the plots.

Randomly sample 5 reviews (including both short reviews and long reviews) and verify whether the sentence segmentation function/tool detects the sentence boundaries correctly. Discuss your results.

Tokenization and Stemming. Tokenize the reviews and show two distributions of the data, one without stemming, and the other with stemming (you may choose the stemming algorithm implemented in any toolkit). Again, the x-axis is the length of a review in number of words (or tokens) and the y-axis is the number of reviews of each length. Discuss your findings based on the two plots. In these two plots, there is no need to consider different rating stars.

List the top-20 most frequent words (excluding the stop words) before and after performing stemming. Discuss the words that you expected to be popular given the nature of the dataset (*i.e.,* reviews of business, mostly about food and restaurant), and the words that you do not expect to be popular in this dataset. Stop words are the words that are commonly used but do not carry much semantic meaning such as *a*, *the*, *of*, *and*. You need to list the stop words used in your analysis in the appendix of your report.

POS Tagging. Randomly select 5 sentences from the dataset, and apply POS tagging. Show and discuss the tagging results.

Most Frequent Adjectives for each Rating. There are **two** listings in this part. First is to list the top-10 most frequently used adjectives for each rating star (*i.e.,* 1 to 5). Discuss the results. Second is to list the top-10 most indicative adjectives for each rating star. There are many ways to define or measure "indicativeness". One way is to use pointwise relative entropy. Let $P(w|R1)$ be the probability of observing word w in all reviews with rating star 1, and let $P(w)$ be the probability of observing word w in all reviews, then relative entropy for word w can be computed as: $P(w|R1) \times log\left(\frac{P(w|R1)}{P(w)}\right)$.

3.3 Development of a $\langle$ Noun - Adjective $\rangle$ Pair Summarizer (30 marks)

Given all reviews of one particular business, we would like to summarize the reviews by extracting the noun-adjective pairs, for example, service - great, food-delicious. The nouns and adjectives here can be extended to phrases (*e.g.,* very good). After extracting these pairs, you may rank these pairs by their frequencies, for example, service-great (5 times) and service-poor (2 times).

Choose any 5 businesses as examples, and summarize reviews of each business by using 5 most frequent noun-adjective pairs. Manually go through some of the reviews and discuss whether your summary does reflect the main points expressed in the reviews.

3.4 Application (10 marks)

Define and develop a simple NLP application based on the dataset. An example application is to detect the sentences containing *Negation Expression* using regular expressions. Negation is often expressed through negative words such as no, not, never, none, nobody. You may define your own application with similar (estimated) difficulty level. Note that, application here means a small tool to analysis or to mine the data. Application here does not mean a web-based application or mobile app.

4 Submission of Report and Source Code

4.1 *Final Report in Hardcopy*

- The hardcopy report must be submitted on or before **4 Nov 2019** (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 10 pages***, excluding appendix. DO NOT include in your report all the source code and complete results sets. However, you must include *code snippets* which are important for the main functions for your task. 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 CECZ4045.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.
    - An installation 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 dataset and the libraries shall **NOT** be included in the softcopy submission to minimize the file size.

- Softcopy submission deadline: ***4 Nov 2019 11:59PM***. Late submissions are allowed but will be penalized by 5% every calendar day (until zero). 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`