ASSIGNMENT

Product Review Data Analysis and Processing

CE/CZ4045 Natural Language Processing

2018/2019 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 hands on experiences on various 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#, 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), LingPipe (Java), GATE (Java), Stanford NLP(Java), OpenNLP (Java)
    - Indexing and Search: Lucene (Java)

3 Assignment (100 marks)

The assignment consists of the following components: Dataset Analysis (40 marks), Development of a Noun Phrase Summarizer (30 marks), Sentiment Word Detection (20 marks), and Application (10 marks).

3.1 Data Format

**** Please note that the dataset is solely for the purpose of this assignment, and you are **NOT** allowed to redistribute this dataset in any format. ****

We will use a collection of user reviews posted on Amazon (product category: Cell Phones and Accessories) as the dataset. The dataset was collected from `http://jmcauley.ucsd.edu/data/amazon/` with further cleaning (*e.g.,* removing reviewer names, and removing lines with formatting or encoding errors). The dataset used in this assignment contains 190,919 reviews, and the data file is about 127 MB. A sample data file containing 49 reviews is provided for you to understand the data format. Each review has the following components:

- reviewerID, *e.g.,* "ASY55RVNIL0UD", the ID of a reviewer and his/her name has been removed from the dataset.
- asin, *e.g.,* "120401325X", the Amazon standard identification number which uniquely identifier a product in Amazon.
- reviewText, *e.g.,* "These stickers are super stylish ....", the full review text.

- overall, *e.g.,* 5.0, the rating of the product by this reviewer.
- summary, *e.g.,* "Really great product.", a short summary of this review.
- unixReviewTime, *e.g.,* 1389657600, the Unix timestamp
- reviewTime, eg "01 14, 2014", the timestamp in MM DD, YYYY format.

3.2 Dataset Analysis (40 marks)

Popular Products and Frequent Reviewers Identify the top-10 products that attract the most number of reviews, and the top-10 reviewers who have contributed most number of reviews. List the product id/user id with the number of reviews in a table.

Sentence Segmentation. Perform sentence segmentation on the reviews and show the distribution of the data in a plot. The x-axis is the length of a review in number of sentences, and the y-axis is the number of reviews of each length. Discuss your findings based on the plot.

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.

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 cell phones and accessories), 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.

3.3 Development of a Noun Phrase Summarizer (30 marks)

Design and implement a *noun phrase detector* to identify and extract noun phrases from ALL reviews and list the top-20 most frequent noun phrases used by reviewers in this dataset. Discuss your findings from the results. You need to give clear definition of noun phrase in your report.

Choose any 3 popular products which has the largest number of reviews, and summarize the reviews of each product by using 10 representative noun phrases. You need to define the meaning of "representative noun phrase" which could consider the frequency (or number of times) a noun phrase appears in this product's reviews, as well as how frequent this noun phrase is among all product reviews.

Randomly sample 5 reviews and evaluate the effectiveness of the noun phrase detector in terms of Precision and Recall. You will need to read through and manually annotate the noun phrases in these 5 reviews and then compare the annotated noun phrases (as groundtruth) and the noun phrases detected by the noun phrase detector.

3.4 Sentiment Word Detection (20 marks)

Sentiment words are the words expressing feelings or emotions towards a product or an aspect of the product (*e.g.,* price) in this context. Each review contains a rating in the range of 1 to 5. Using the rating as an indication, find the top-20 representative words (*e.g.,* great, good) for expressing positive sentiment and top-20 representative words (*e.g.,* disappointing) for expressing negative sentiment, respectively, for the entire dataset. Detail the procedure on how to identify these two lists of representative words.

3.5 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 **5 Nov 2018** (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 of your system. 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.

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

  - 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: ***5 Nov 2018 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.