ASSIGNMENT

Review Data Analysis and Processing

AI6122 Text Data Management and Processing

2022/2023 Semester 1

NANYANG TECHNOLOGICAL UNIVERSITY

1 Objective

The objective of this assignment is to let you getting familiar with the main components in end-to-end text management and processing applications, the challenges faced by each component and the solutions. Through this assignment, you shall also get hands on experiences on various packages available for information retrieval and natural language processing 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. Group size is not a factor in grading.
3. You may use ANY programming language of your choice, *e.g.,* Java, Python, C#.
4. You may use any NLP, IR, 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, relational database like MySQL is not allowed.
    - 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 (40 marks), Development of a Simple Search Engine (20 marks), Development of a Review Summarizer (20 marks), and Development of an Application (20 marks).

3.1 Data Format

We will use a collection of user reviews posted on Amazon in this assignment. Typically a review shall contain 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)

Dataset Download and Sampling. The datasets are available at `https://jmcauley.ucsd.edu/data/amazon/`. Please read the assignment carefully before downloading any datasets as we do not need to use all the datasets on the page.

We will use **two datasets** under the section titled "Small" subsets for experimentation, *i.e.,* the 5-core datasets. From the 24 datasets that have been preprocessed with 5-core filtering, randomly choose and download 2 datasets. For example, you may use the first two datasets "Books" and "Electronics" or any other two categories.

From each of the two downloaded datasets, randomly sample 200 products and keep all the reviews of these 200 products in each dataset. All our subsequent analysis will be conducted on these two sampled datasets, each with reviews of 200 products.

Writing Sytle. 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; are there many spelling errors; etc.) in comparison to news articles published by The Straits Times. Discuss your findings.

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

Sentence Segmentation. Perform sentence segmentation on the reviews and compare the distribution of the two datasets in a single 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.

Tokenization and Stemming. Tokenize the reviews and compare the review length distribution in terms of number of tokens. The x-axis is the length of a review, defined by the number of tokens, and the y-axis is the number of reviews having that length. Are there any difference between the two product categories?

At the dataset level, show two distributions to observe the impact of stemming. You may choose the stemming algorithm implemented in any toolkit to perform token stemming. Show two distributions of the data, one without stemming, and the other with stemming. The x-axis is the number of unique tokens in a dataset, and the y-axis is the number of times each token appears in the dataset. The tokens shall be sorted by their frequency in your plots. Discuss your findings based on the two plots.

Indicative Words List the top-10 most indicative words in each of the two datasets. There are many ways to define or measure a word "indicativeness". One way is to use pointwise relative entropy. Let $P(w|D_1)$ be the probability of observing word w in all reviews in dataset D_1, and let $P(w|D_2)$ be the probability of observing word w in all reviews in dataset D_2, then relative entropy for word w can be computed as: $P(w|D_1) \times log\left(\frac{P(w|D_1)}{P(w|D_2)}\right)$. You may also consider other measures.

3.3 Development of a Simple Search Engine (20 marks)

Write a search engine to index and search reviews, by using Lucene or other libraries specific to IR.[1] In this part of the assignment, you may use (i) One main IR specific library for most of the operations; (ii) Any other third-party libraries if and only if the main library does not provide the required functionality; and (iii) Any stopword list of your choice. However, you are not allowed to use very high-level libraries like Elasticsearch.

In this search engine, each review is a "document" and you may discuss what filed(s) (*e.g.,* reviewID, asin, reviewText, overall rating) shall be indexed and searchable. Detail your choice of parsing/linguistic processing on the words/terms in the chosen fields, *e.g.,* whether to perform stemming, case folding, stopword removal, in these fields. Based on the number of "documents" to be indexed in the dataset, collect the time needed to index every 10% of the documents. Discuss your findings on the indexing time.

Your search engine should at least support free text keyword queries (including single keyword query and phrase query) on the "text" field in a review. Top N (the number of N is configurable) results should be returned via the console[2] along with rank, scores, docID, and snippets whenever possible. Your search engine shall also support search for reviews that are specific to a product defined by asin or product id.

Randomly choose a few queries (including both single keyword query and phrase queries), discuss whether the results returned by the search engine are as expected. You may also record the time taken to process a query.

3.4 Development of a Review Summarizer (20 marks)

Given a product, the summarizer should summarize all *reviewTexts* received for this particular product. Your are not allowed to use the "summary" component in the review dataset.

You will define and justify what an ideal summary should be, based on your understanding. Example summaries could be *e.g.,* a list of indicative keywords, a list of key phrases, a list of noun-adjective or adjective-noun pairs or phrase pairs, a list of representative sentences. Discuss the technical challenges to achieve the ideal summarization, and your solution. Present your solution and justify why the proposed approach (*e.g.,* each component in your solution) is the best option and the limitations of your solution. Discuss how to evaluate your solution and possible alternative solutions (or baselines). Randomly choose 2 products, one from each dataset, and produce the review summary. The chosen products should have a reasonable number of reviews.

3.5 Application (20 marks)

Define and develop a simple application based on the dataset. An example application is to derive a collection of sentiment words from this dataset. 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.

[1] See `http://en.wikipedia.org/wiki/List_of_information_retrieval_libraries` for a list.

[2] Note, a text-based command line system is sufficient; a GUI or web-based interface to the search engine is NOT encouraged.

4 Submission of Report and Source Code

4.1 *Final Report in Hardcopy*

- The hardcopy report must be submitted on or before **24 Oct 2022** (Monday, Week 11), through SCSE General Office.
- The report must use the provided cover page, and the main content shall be formatted following the ACM "sigconf" proceedings templates[3] (either MS Word or Latex). The main content of the report ***must not exceed 10 pages***, *i.e.,* excluding cover page and 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.
- Before submission, please read the hardcopy of your own report. **Make sure any words or pictures in your report are readable**.

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

- An AI6122-Gxx.zip file containing the following files and folder shall be submitted: Report.PDF, Readme.txt, SourceCode.
  - The Gxx is your Group ID.
  - 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 only your source code. The dataset and the libraries shall **NOT** be included in the softcopy submission to minimize the file size.
- Softcopy submission deadline: ***24 Oct 2022 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 considered and time-stamped.

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