NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

ASSIGNMENT

Review Data Analysis and Processing

AI6122 Text Data Management & Processing

September 2024

NANYANG TECHNOLOGICAL UNIVERSITY

1 Objective

The objective of this assignment is to familiarize you with the main components of end-to-end text management, processing, and search applications, as well as the challenges faced by each component and their corresponding solutions. Through this assignment, you will also gain hands-on experience with various packages available for information retrieval and natural language processing tasks.

2 Assignment Format

1. This is a group assignment. Each group has at most 6 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), and deep learning based libraries.
   - Indexing and Search: Lucene (Java), with Python port available.

3 Assignment (100 marks)

The assignment consists of the following components: Dataset Analysis (20 marks), Development of a Search Engine (30 marks), Review summary (40 marks), and Application (10 marks).

3.1 Dataset

We will use the Yelp Open Dataset (`https://www.yelp.com/dataset`). You should read the dataset documentation before working on the assignment: `https://www.yelp.com/dataset/documentation/main` In this assignment, we will use two files from the dataset: business.json and review.json.[1]

3.2 Dataset Analysis (20 marks)

Data Sampling. The Yelp dataset includes reviews from 11 metropolitan areas. Your first task is to sample businesses and reviews from one of these areas. Clearly indicate which metropolitan area your group has selected. Then, construct a dataset exclusively for the chosen area, containing only the businesses located within it and their associated reviews. You should also provide basic statistics for the newly sampled dataset. From this point onward, all tasks will be conducted using the dataset from the selected metropolitan area.

[1] Due to its big size, one dataset download per group is strongly encouraged.

Tokenization and Stemming. Randomly select a business, b_1, from the sampled dataset and extract all reviews associated with b_1 to form a small dataset, B_1. Display the word frequency distributions in B_1 before and after applying stemming. You may use any stemming algorithm available in existing toolkits. Consider plotting the word frequency distributions on a log scale (experiment with different visualizations to best present your data and insights).

Repeat the process for another randomly selected business, b_2, and compare the findings between the two businesses using the generated plots.

Additionally, list the top 10 most frequent words (excluding stopwords) before and after stemming for both businesses. Provide a discussion of your findings based on these results.

Writing Style. Randomly sample the following: (i) two posts from StackOverflow, (ii) two posts from Reddit, (iii) two news articles from Straits Times, and (iv) two patents in any domain. Discuss the differences on their writing styles (*e.g.,* is the first word in a sentence capitalized; do sentences follow good grammars; are the proper nouns capitalized; etc). You need to provide the URLs of the sampled posts/articles in your report.

3.3 Development of a Search Engine (30 marks)

Write a search engine to index and search reviews, by using Lucene or other libraries specific to IR.[2] 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. Your search engine shall support the following types of searches:

- Keyword search of business *e.g.,* searching for restaurants by words in their names.
- Keyword search of reviews *e.g.,* searching for reviews that contain certain keywords or phrases.
- Geospatial search *e.g.,* searching for businesses located within a geospatial area defined by a bounding box.

Detail your definition of "document" in your indexing, and your choice of parsing/linguistic processing on the words/terms in the chosen fields, *e.g.,* whether to perform stemming, case folding, stopword removal. 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 return top N (the number of N is configurable) results for each search via the console[3] along with rank, scores, docID, and snippets whenever possible. Discuss whether the results returned by the search engine are as expected with sample queries. You may also record the time taken to process a query.

3.4 Review Summary (40 marks)

On most of websites, reviews are associated with the items (*i.e.,* businesses in Yelp dataset). In this assignment, we would like to present a summary of reviews for a specific user.

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

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

Before presenting a user's review summary, your task is to show a distribution of reviews contributed by the users. You may plot a figure with x-axis showing the number of reviews contributed by a user, and y-axis showing the number of users having a particular number of reviews. Each point in this plot would show the how many users have 10, 20, or 30 reviews where 10, 20, and 30 are along the x-axis.

Given a user id, his/her review summary contains the following information:

- The number of reviews he/she has contributed.
- The bounding box of the businesses that he/she has reviewed. This bounding box may suggests the activity area of this user.
- The top-10 most frequent words used in his/her review, excluding stopwords. It would be a plus to show the top-10 most frequent phrases used in his/her reviews.
- Three most representative sentences selected from his/her reviews. In your report, you should describe what does it mean by "representative" and how the sentences are selected.

The ideal implementation is that this review summary is built on top of the search engine that you have built earlier, where we can input/select a user id and then get his/her review summary based on the search results on the fly.

3.5 Application (10 marks)

Define and develop a simple NLP/IR application based on the dataset. An example application is to detect the sentences containing *comparison* in reviews. In a sentence containing "comparison" the reviewer compares the current business being reviewed with other businesses that he/she has visited. Note that, application here means a small tool (or piece of code) to analyse or to mine the data. Application here does not mean a web-based application or mobile app. Command line interface is sufficient, and GUI or Web-based interface does not contribute to grading.

4 Submission of Report and Source Code

4.1 *Final Report in Hardcopy*

- The hardcopy report must be submitted on or before **1 Nov 2024** (Friday), through CCDS General Office.
- The report must use the provided cover page, and the main content shall be formatted following the ACM "sigconf" proceedings templates[4] (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.

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

- 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*

- Report_Gxx.PDF shall be submitted through NTULearn using the link for *report submission*. This PDF file shall has the same content as the hardcopy report submitted. Note: xx is your group number. Only one member from each group needs to submit the softcopy on behalf of the group.

- An Src_Gxx.zip file containing the following files and folder shall be submitted through NTULearn using the link for *sourcecode submission*: Readme.txt, SourceCode. Only one member from each group needs to submit the softcopy on behalf of the group.

  - 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 own source code**. The dataset and the libraries shall *****NOT******* be included in the softcopy submission to minimize the file size.

- Softcopy submission deadline: ***1 Nov 2024 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.