Course Assignment 2 – Search Engine

CI-6226 Information Retrieval & Analysis

In this course there are a total of three individual assignment. This is assignment 2 of 3. A single PDF file is to be submitted by every student containing the report.

Feel free to use any external materials but don't forget to reference your sources.

Reporting on Completion of Assignment

Read this section very carefully. Failure to adhere to the simple rules of reporting will lead to lowered marks.

The assignment report is due before Sunday April 4th, 2021 @ 23:59. Every student is to submit electronically one PDF file (and nothing else) called '*<MATRIC>*`-2.pdf`' containing the assignment report, where *<MATRIC>* is your matriculation number. E.g., a student with the matriculation number A123456B should submit a file named '`A123456B-2.pdf`'.

The report should not have a separate title page. At the top of the first page please put the following mandatory elements

- Title: **CI6226 Information Retrieval & Analysis / Assignment 2 / AY20-21**
- Full name
- Matric number
- Your NTU email address

The report should not exceed 3 pages excluding references. The report should cover what was done in each step of the assignment, provide reasoning for the chosen course of actions, demonstrate examples (where applicable). The report should be written as a coherent text. You are not required to submit your code, but you can showcase portions of your code in the report.

Submission should be done in NTULearn.

The report must be neatly formatted. Reports that are hard to read due to formatting (or any other reason) will be marked low or not marked at all in extreme cases.

Grading

The assignment is overall graded on a 0–100 scale. This assignment comprises 25% of the course grade. In this assignment you can earn **Bonus Points**. They will be added to your final score for the course. E.g., if your marks for assignment 1-3 and the quiz were 95, 90, 85, and 80 and you scored 10 bonus points, your total marks for the course will be $\frac{95+90+85+80+\mathbf{10}}{4} = 90$.

Dataset

You are provided a dataset for this assignment, which you are free to use. You can use your own dataset as well.

Assignment

In this assignment we continue building an information retrieval system based on the results of Assignment 1, which was a system that can output a sorted list of term-document pairs.

The task in this assignment is:

1) Build an inverted index;
2) Enable simple Boolean search;
3) Implement compression techniques.

1. Inverted Index

Input: a file with sorted term-doc pairs
Output: inverted index

In this part you would need to take the file containing the sorted list of term-doc pairs and transform it into a simple inverted index. In this assignment you **don't have to** worry that the list or the inverted index can be too big for main memory; however, you can take that into account.

Bonus Points: persist the inverted index as a file so you don't need to rebuild it every time you launch the program.

2. Boolean Search

Input: a search query, an inverted index
Output: a list of documents satisfying the query

Implement a simple AND-based Boolean search, i.e., a query "horse car phone" should be treated as "horse AND car AND phone" and return only documents that contain all three words.

Bonus Points: Implement OR and NOT in addition to AND.

3. Index Compression/Optimization

Implement compression and optimization techniques that were discussed in the lectures. In particular, implement *at least* the Dictionary-as-a-String approach. Implementing other techniques (blocking, front-coding, skip pointers, variable-length gap encoding) is encouraged.

Compare your search engine performance and memory requirements before and after implementing compression and optimizations. Reflect the comparison in your report.

Bonus Points: If you implement many techniques and manage to achieve impressive savings in speed and/or memory, that may earn you bonus points.

Good luck, have fun!