Assignment 2: Boolean Search

phuonglh@gmail.com

March 6, 2023

1 Problem

In this assignment, you will develop a program capable of searching for documents in an indexed corpus. The indexed corpus was created in the first assignment. (See *A1: Corpus Indexing* of the Reuters corpus.)

Suppose that the index contains the following posting lists, each list is sorted by document id:

tom	$\rightarrow$	*List(1, 4, 8, 10)*
jerry	$\rightarrow$	*List(1, 4, 7)*
dog	$\rightarrow$	*List(1, 7, 11, 13)*
like	$\rightarrow$	*List(2, 4, 7)*

Your program takes as input an index and a query, and it should return the search result, which is a list of document ids. There are five simple, boolean query types as follows:

1. A single-term query, which contains a single token (or term), for example *q=tom*; the result should be *List(1, 4, 8, 10)*.
2. A two-term query with the boolean operator AND, for example *q=tom AND jerry*; the result should be *List(1, 4)*.
3. A two-term query with the boolean operator OR, for example *q=tom OR jerry*; the result should be *List(1, 4, 7, 8, 10)*.
4. A two-term query with the boolean operator AND, and NOT, for example *q=tom AND (NOT jerry)*; the result should be *List(8, 10)*.
5. A two-term query with the boolean operator OR, and NOT, for example *q=tom OR (NOT jerry)*; the result should be *List(1, 2, 4, 8, 10, 11, 13)*.

Note that the returned result should contains sorted document ids.

2 Guide

Think about functional programming style, given two collections A and B:

- How to express $A \cap B$ as a function?
- How to express $A \cap \overline{B}$ as a function?
- Is computing $A \cap B$ more efficient than $B \cap A$, why?

Your should design and organize your search functionalities into multiple functions, for example *search(index, tom)*, *searchAnd(index, tom, jerry)*, etc.

3 Submission

If you develop your program in Scala, Python, or Java, submit your source file *YourFullName.scala*, *YourFullName.py*, or *YourFullName.java*, respectively; for example *NguyenPhuTrong.scala*. If you have multiple source files then put them in a directory and zip it to *YourFullName.zip* and submit.

- Do not submit a Notebook (say **.ipynb*). Your teacher will not run any notebook server for evaluation of this assignment. Do not submit **.rar*, your teacher does not want to be made install an external program to decompress your file.
- You need to submit only your source code, no data is included to save space. You can assume that a user needs to provide a data folder (i.e., *.../dat/reuteurs/test*) as an argument for your program to run.
- Submit your file to Classroom before the due date. You have 7 days to turn in your program.

Further Developments*

This section is not required for this assignment: Write a front-end application which allows users to interact with your back-end program more friendly. This front-end app can be desktop-based or web-based and has some functionalities. For example, it has an GUI which allows a user:

1. Click a button to select the data folder that contains input documents;
2. Click a button to start indexing the corpus;
3. Show the vocabulary once the indexing finishes;
4. Type a query into a textbox, hit Enter or a Submit button, search result is displayed in a list.
5. Permit specifying boolean operators (AND, OR, NOT).