CS-839 NoSQL Systems
Dr. Vinu Venugopal, Divyam Agrawal, Aniket Mitra & Nandakishore S Menon

Assignment #1

This is a group assignment. The deadline for submitting the assignment is 30th Jan 2023 (midnight). The solutions for the assignment problems along with a detailed report should be submitted via the link provided on LMS. Only one submission per group is required. Each group should demonstrate the solutions before the assigned TA. The demo schedule will be communicated on LMS.

Set Up Your Local Working Environment

For this assignment, we will use the PostgreSQL open-source database system for loading and querying the prepared dumps of the Wikipedia articles provided on LMS.

- Download and install the **PostgreSQL** (version 14.6 or above) database server by choosing the respective installer binaries for your operating system from: `https://www.postgresql.org/download/`
- PostgreSQL provides a graphical user interface (GUI) for its installation under all operating systems. Simply follow the installation steps provided by this interface and make sure that you remember your admin password for the provided example database instance. This example database instance is called `postgres` by default and can be accessed by using your password from either the admin GUI (`pgAdmin 4`) or the command-line shell (`psql`).
- See also Chapter 2 of the "Seven Databases in Seven Weeks" book for useful hints about installing and administrating PostgreSQL (if needed).

The goal of this first exercise sheet is to develop an understanding for what the technical challenges in processing queries over a large collection of text documents are. We will practice this by designing a simple search engine for Wikipedia. For this purpose, the two files `Wikipedia-EN-20120601_KEYWORDS.TSV.gz` and `Wikipedia-EN-20120601_REVISION_URIS.TSV.gz` are provided for download on the course homepage on `https://learn.iiitb.net/`.

The file `Wikipedia-EN-20120601_KEYWORDS.TSV` is a tab-separated file that contains a dump of 10,000 Wikipedia articles, which have been preprocessed and parsed into the following format:

```
ID<tab>TERM<tab>SCORE
```

- `ID` represents a unique numerical identifier for each Wikipedia article in the dump file.
- `TERM` denotes the distinct keywords that are contained in the corresponding Wikipedia articles in UTF-8 format. All keywords have been stemmed according to the Porter stemming algorithm[1] and stopwords, such as `the`, `and`, `have` and so on, have been removed.
- `SCORE` represents a decimal numeral value that captures the relevance of the keyword in the given article. The higher the score, the more relevant the article should be for this keyword.

Additionally, the file `Wikipedia-EN-20120601_REVISION_URIS.TSV` contains mappings of the article identifiers to the archived revisions of these Wikipedia articles. These URLs still are mostly intact, such that you can verify your results by opening the Wikipedia articles in a browser.

Bulkloading Data into a PostgreSQL Database **6 Points**

Problem 1.

1. Create a new database instance in PostgreSQL and create two relation schemas (i.e., tables) in order to store the content of the two files `Wikipedia-EN-20120601_KEYWORDS.TSV` and `Wikipedia-EN-20120601_REVISION_URIS.TSV` in your database. Think of appropriate primary- and foreign-key constraints for your two relation schema. **2 Points**

[1] `http://tartarus.org/martin/PorterStemmer/`

2. Bulkload the two TSV files into your relation schema using the PostgreSQL command `COPY`[2]. Make sure that the original UTF-8 encoding of the TSV files is preserved in your database. **2 Points**

3. Repeat the above bulkloading step once by including all your primary- and foreign-key constraints into the relation schema, and once by omitting all the primary- and foreign-key constraints from your relation schema. Compare the runtimes of the two options. **2 Points**

Running Keyword Queries over Wikipedia **16 Points**

Problem 2. In the next step, we will compare different retrieval modes for evaluating keyword queries against the Wikipedia database you created in the previous problem. Formulate appropriate SQL queries to solve the following tasks.

Note: All of the following query tasks can be solved by using `GROUP BY` queries in SQL over the table containing the contents of the `Wikipedia-EN-20120601_KEYWORDS.TSV` file. A final join operation is only required between the table with the contents of `Wikipedia-EN-20120601_KEYWORDS.TSV` and the table with the contents of `Wikipedia-EN20120601_REVISION_URIS.TSV`.

1. *Boolean Retrieval 1*: Find URLs of Wikipedia articles that contain *all* of the stemmed keywords `infantri`, `reinforc`, `brigad`, and `fire`. **2 Points**

2. *Boolean Retrieval 2*: Find URLs of Wikipedia articles that contain *any* of the stemmed keywords `infantri`, `reinforc`, `brigad`, or `fire`. **2 Points**

3. *Boolean Retrieval 3*: Find URLs of Wikipedia articles that contain the stemmed keyword `reinforc` but *not* any of the stemmed keywords `infantri`, `brigad`, or `fire`. **2 Points**

4. *Ranked Retrieval 1*: Find URLs of Wikipedia articles that contain *all* of the stemmed keywords `infantri`, `reinforc`, `brigad`, and `fire`, ordered by the sum of the scores of these keywords in the articles. **3 Points**

5. *Ranked Retrieval 2*: Find URLs of Wikipedia articles that contain *any* of the stemmed keywords `infantri`, `reinforc`, `brigad`, and `fire`, ordered by the sum of the scores of these keywords in the articles. **3 Points**

   Note: An article may be returned here if it contains any subset of these keywords, but the higher the sum of the scores of the keywords in an article, the higher this article should be ranked.

6. *Ranked Retrieval 3*: Find URLs of Wikipedia articles that contain the stemmed keyword `reinforc` but *not all* of the stemmed keywords `infantri`, `brigad`, or `fire`. Consider the following note while assigning the ranks. **4 Points**

   Note: An article shall be returned even if it contains any proper subset of the keywords `infantri`, `brigad`, and `fire` – it should of course contain the keyword `reinforc`. However, if the score for `reinforc` is higher than the sum of the scores for the terms `infantri`, `brigad`, and `fire` for an article, then the article should be ranked higher.

[2] `http://www.postgresql.org/docs/current/interactive/populate.html`