Assignment 1: Data Analysis

Individual Assignment

Table of Contents

1. Tasks

Task 1: Get data from Stack Exchange (Data Acquisition/Collection)
Task 2: Load data into chosen cloud technology (MapReduce/Pig/Hive)
Task 3: Query data using MapReduce/Pig/Hive
Task 4: Calculate TF-IDF with MapReduce/Pig/Hive

Note: plenty of versions of code available online, make sure to acknowledge the source and ***describe the changes*** *you did*

2. Task Description

2.1. **Task 1** - Acquire the top 200,000 posts by ViewCount (see Section 3 - Data Acquisition for more details)

2.2. **Task 2 & 3** - Use Pig/Hive/MapReduce - Extract, Transform and Load the data as applicable to get:
  2.2.1. The top 10 posts by score
  2.2.2. The top 10 users by post score
  2.2.3. The number of distinct users, who used the word “cloud” in one of their posts

2.3. **Task 4** - Use Mapreduce/Pig/Hive to calculate the per-user TF-IDF of the top 10 terms for each of the top 10 users

3. Data Acquisition

Use the Data Explorer feature of the StackExchange system using following link to run your queries:
https://data.stackexchange.com/stackoverflow/query/new

Clarification is needed regarding the data acquisition: you can only download maximum 50,000 records at a time. This means you would need to run at least 4 queries from the Stack Exchange site to get 200,000 posts!

The queries provided in the notes are just examples: you need to figure out how to change the parameters of these queries to get **exactly** the top 200,000 posts!

If you run the query example in the notes:

Example Query 1:
```
select top 50000 * from posts where posts.ViewCount > 1000000
ORDER BY posts.ViewCount
```

you get around 1500 records. This tells you that there are more posts with ViewCount lower than 1,000,000. This means you need to run different instance of the following query example

Example query 2:
```
select count(*) from posts where posts.ViewCount>15000 and
posts.ViewCount < 20000
```

This query helps us figure out what range of ViewCount contains the top 50,000 posts, then the second batch of 50,000 and so on till you reach 200,000.

Example query 3: if you run the query
```
select count(*) from posts where posts.ViewCount > 100000
```

you will get a number of records close to 50,000, which means you can retrieve them all in one query. The subsequent batches of 50,000 have probably a view count lower than 58,000 and greater than another number you need to find (using the second example query) and so on till you reach a download of 4 or 5 CSV files that sum up to 200,000 records you would then need to extract, transform and load as per **Task 1**.

4. Submission

- Submission open: Week 4, 14-October-2021 (7pm)
- **Due date: Week 6, 28-October-2021 (11:59pm)**
- Submit 1 docx or pdf file as per instructions below
- Worth 20% of the final marks

Assessment criteria (see Section 5 for more details):
- Task completion quantity
- Task completion quality

Use Gitlab (http://gitlab.computing.dcu.ie/) or GitHub (http://www.github.com/) to create a new repository, use this repository to store your source code for each task. The URL link for your repository should be added in your submission document.

Main submission document should be **report - a word (.docx) or PDF (.pdf)** file submitted on **Loop (http://loop.dcu.ie)**.

The report document should follow the format (5 pages maximum, excluding the screenshots):

1. Student details
    a. Name
    b. Student ID
    c. Email
2. Link for the Git repository
3. Short description of the dataset acquired along with the steps taken to acquire the dataset
4. Technologies used for each task – describe why you chose the technology for the task
5. Description of the steps taken to achieve each task along with query or code (link to the file on the repository)
6. Snippet of the important part of the source code (wherever possible)
7. Relevant screenshots (describe what the screenshot represents) to show the work and completion of the tasks (use Appendix at the end)

5. Assessment Criteria

Criteria	0-4	4-8	8-12	12-17
Task Completion Quantity	One task or less fully completed	Two tasks fully completed	Three tasks fully completed	Four tasks fully completed
Task Completion Quality	Major errors	Several minor errors	Only a few minor errors	No errors

3 additional marks if all tasks implemented on GCP/AWS or other cloud system