Assignment 3
Posted Date: Jul 17, 2023
Submission Due: Jul 31, 2023 (11:59 pm)
Late assignments will not be accepted and will result in a 0 on the assignment

Objective: This assignment covers two learning objectives (lo).

- ***lo#1:*** Perform research on NoSQL and data processing – To achieve this task, you need to read and understand the usage of spark framework, MongoDB and then implement a programming framework for big data processing, and store.
- ***lo#2:*** Build a light-weight analytics engine, which will perform custom ETL operation, and one specific analysis (sentiment and semantic)

Plagiarism Policy:

- This assignment is an individual task. Collaboration of any type amounts to a violation of the academic integrity policy and will be reported to the AIO.
- Content cannot be copied verbatim from any source(s). Please understand the concept and write in your own words. In addition, cite the actual source. Failing to do so will be considered as plagiarism and/or cheating.
- The Dalhousie Academic Integrity policy applies to all material submitted as part of this course. Please understand the policy, which is available at: https://www.dal.ca/dept/university_secretariat/academic-integrity.html

Assignment Rubric - based on the discussion board rubric (McKinney, 2018)

	Excellent (25%)	Proficient (15%)	Marginal (5%)	Unacceptable (0%)
Completeness including Citation	All required tasks are completed	Submission highlights tasks completion. However, missed some tasks in between, which created a disconnection	Some tasks are completed, which are disjoint in nature.	Incorrect and irrelevant
Correctness	All parts of the given tasks are correct	Most of the given tasks are correct However, some portions need minor modifications	Most of the given tasks are incorrect. The submission requires major modifications.	Incorrect and unacceptable
Novelty	The submission contains novel contribution in key segments, which is a clear indication of application knowledge	The submission lacks novel contributions. There are some evidences of novelty, however, it is not significant	The submission does not contain novel contributions. However, there is an evidence of some effort	There is no novelty
Clarity	The written or graphical materials, and developed applications provide a clear picture of the concept, and highlights the clarity	The written or graphical materials, and developed applications do not show clear picture of the concept. There is room for improvement	The written or graphical materials, and developed applications fail to prove the clarity. Background knowledge is needed	Failed to prove the clarity. Need proper background knowledge to perform the tasks

Citation: McKinney, B. (2018). The impact of program-wide discussion board grading rubrics on students' and faculty satisfaction. Online Learning, 22(2), 289-299.

Explanation of the rubric: Suppose you received different grades in Clarity for the 2 problems
Problem #1: 25% in clarity
Problem #2: 15% in clarity
Then your overall grade for the clarity will be avg of (25+15) % = 20%

Problem 1A: Reuter News Data Reading & Transformation and storing in **MogoDb**. Objective is lo#1

1. From the two given news files (reut2-009.sgm, and reut2-014.sgm), create MongoDb Database – ReuterDb, where each Document contains a news article. The task must be done using a Java Program "ReutRead.java".
    a. To perform this operation, you need to write a Java code to scan the required texts between two <REUTERS></ REUTERS > tags, <TEXT></ TEXT> tags, and <TITLE></ TITLE > tags.
    b. In the ReuterDb, you may consider each news as a document. You can also include nested or sub-document. {
    title: "",
    text: ""
    }
2. You need to include a flowchart and algorithm of your Reuters Data cleaning/transformation program on the PDF file.

Problem 1B: Reuter News Data Processing using **Spark**. Objective is lo#1

1. Using your GCP cloud account, configure and initialize Apache Spark cluster. (Follow the tutorials provided in Lab session).
2. Create a flowchart or write ½ page explanation on how you completed the task, include this part in your PDF file.

Note: If for some reason, you fail to work on GCP cloud account (valid reasons required), you need to create local standalone Hadoop/Spark cluster to perform the next set of operations.

3. Write a MapReduce program using Java (WordCounter.java Engine) to count (frequency count) the unique words found in "reut2-009.sgm".
4. You need to include a flowchart/algorithm of your MapReduce program on the PDF file.
5. In your PDF file, report the words that have highest and lowest frequencies.

Problem 2: Sentiment Analysis using BOW model on title of Reuters News Articles

Use Core Java Program only with no additional libraries. Use the parser (regex based parser)

1. Write a Java program to create bag-of-words for each News title. (code from online or other sources are not accepted)

e.g. news1 = "best best deals on laptop in Canada"
bow1 = {"best":2, "deals":1, "on":1, "laptop":1, "in": 1, "Canada":1}

You do not need any libraries. Just implement a simple counter using loop.

2. Compare each bag-of-words with a list of positive and negative words.

You can download list of positive and negative words from online source(s). You do not need any libraries. Just perform word by word comparison with a list of positive and negative words that you can get from any online platform. E.g. negative words can be found here https://gist.github.com/mkulakowski2/4289441

3. Tag each news title as "positive", "negative", or "neutral" based on overall score. You can add an additional column to present your finding.

E.g. frequencies of the matches "best"=+2, Overall score = +2 (positive)

News#	Title Content	match	Polarity
1	best best deals on laptop in Canada	best	Positive

Submission Guidelines:

1. All written reports, images, code etc. must be added in a folder, and compress it with **.ZIP** format only.
2. If not mentioned by TAs, then please rename the .zip file with your B00xxxxx_FnameLname_A2
3. Submit your Java code in gitlab. Your TA must have provided guidelines for that. If not, please ask the TA.
4. You must include Test Cases (at least 3 – manual testing of functionality or validation testing) for the developed application and provide necessary screenshots as evidence of testing. Note: This is not Junit test, this is functional test of each problem.
5. Check the next point "Suggestions" for quality improvement and time management.

Suggestions:

Better Quality: To obtain good grades, you should follow the points given below:

- Try to understand the assignment requirement and follow all the steps required.
- Do not miss adding citations. If you write a single sentence taking the idea from somewhere else, then give credit to the author. Therefore, provide citation for any report you write, or any code you implement
- When you add citation, make sure to add it in a standard format and uniform format. E.g. if I refer 3 sources for writing a report, then I must cite the 3 sources in same format. One source in MLA, two sources in APA citation format will be a mismatch. Therefore, follow any one standard citation format
- Make sure to provide inline citations within report, and programming code
- Any image/picture/flowchart/diagram you add, make sure to provide a caption and a number for that image. It should be placed at the bottom of the image. E.g. "**Fig 1: Weekly time management chart for CSCI 5408**"
- Any table you add, must have a number and caption. This should be added on top of the table. E.g. "**Tab1: Table highlights the requirements in a ordered format**"

Time Management: Follow proper time management to reduce stress, and last-minute preparations. I am suggesting you follow the pie chart, which will require you to spend 5 hours in a week outside the classroom time for this course.

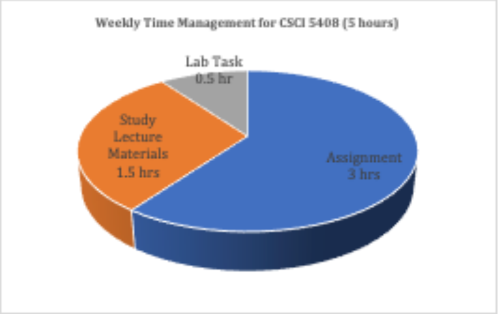

Fig 1: Weekly time management chart for CSCI 5408