Assignment #5

CSCI 5408 (Data Management, Warehousing, Analytics)
Faculty of Computer Science, Dalhousie University

Date Given: Mar 15, 2022
Due Date: Mar 29*, 2022 at 11:59 pm
(with or without SDA, this assignment cannot be submitted after Mar 30, 2022.)

Submissions after Mar 30th are not accepted due to end of term evaluation

Disclaimer: This assignment requires students to work on BI Framework, and sentiment/semantic analysis. Submissions related to this assignment will not be used for commercial purposes.

Objective:

- The objective of this assignment is to understand BI framework, creating star/ snowflake schema, and concept of sentiment and semantic analysis.

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

	Excellent (25%)	Proficient (15%)	Marginal (5%)	Unacceptable (0%)	Problem # where applied
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	Problem #1 Problem #2 Problem #3
Correctness	All parts of the given tasks are correct	Most of the given tasks are correct However, some portions need	Most of the given tasks are incorrect. The submission	Incorrect and unacceptable	Problem #1 Problem #2 Problem #3

		minor modifications	requires major modifications.		
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	Problem #1 Problem #2 Problem #3
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	Problem #1 Problem #2 Problem #3

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

Tasks

- This assignment requires you to submit programming codes on gitLab, and a single PDF file on Brightspace.

Problem #1

Business Intelligence Reporting using Cognos

1. Download the weather dataset available on https://www.kaggle.com/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region?select=sudeste.csv

2. Explore the dataset and identify data field(s) that could be measured by certain factors or dimensions. (Follow recorded lecture #18, and synchronous session #18)

   Example: In a Sales dataset, you may find a measurable field "total sales", which could be analyzed by other factors such as, "products", "time", "location" etc. These factors are known as dimensions. Depending on the data, you may also find possibilities of slice and dice, i.e. analysis could be possible in more granular level; From **total sales by city** to **total sales by store**

3. Write ½ page explanation on how did you select the measurable filed, i.e. fact and what are the possible dimensions. Include this part in your PDF file.

4. Clean the dataset, if required perform formatting. You can perform the cleaning and formatting using spreadsheet operation or programming script. If you use program add that in GitLab, if you use other methods, write the steps in the PDF file.

5. Create Cognos account and import your dataset. Identify the dimensions, and create/import the dimension tables.

6. Based on your understanding of the domain (please read the information/metadata available on the dataset source, i.e. Kaggle), create star schema or snowflake schema. Provide justification of your model creation in the PDF file. At this stage build your schema (dimension modelling) using a drawing tool like draw.io.

7. In addition to justification, attach screenshot of the model created using Cognos (star schema or snowflake schema) in the PDF file.

8. Display visual analysis of the data in a suitable format, e.g. bar graph showing temperature change in terms of a suitable dimension. Add the screenshot of the analysis on the pdf or add a screen recording of the analysis on your .zip folder.

Problem #2

Sentiment Analysis – Java Program only with no additional libraries. Use the parser you developed in previous assignment.

1. To perform this task, you need to consider the processed tweets (**messages** only, ignore other fields) that you obtained and stored in MongoDB in your previous assignment. If you could not perform/complete the task, then obtain the processed MongoDb tweets collection by contacting your TA Manraj

2. Write a script to create bag-of-words for each tweet message. (**code from online or other sources are not accepted**)
   e.g. tweet1 = “Canada is cold cold. I don’t feel good; but is not bad”
   bow1 = {“Canada”:1, “is”:2, “cold”:2, “I”:1, “don’t”: 1, “feel”:1, “good”:1, “but”:1, not”:1, “bad”:1}
   You do not need any libraries. Just implement a simple counter using loop.

   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 tweets 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 “cold”=2, “not” =1, “bad”=1 (**negative -4**), “good”=1 (**positive +1**). Overall score = -3

Tweets #	message	match	polarity
1	Canada is cold cold. I don’t feel good; but is not bad	cold, good, not, bad	negative

Problem #3

Semantic Analysis – Java Program only with no additional libraries. Use the parser you developed in previous assignment.

1. For this task, consider the processed tweets collection that you created in Assignment 4.

2. Use the following steps to compute TF-IDF (term frequency-inverse document frequency)
    a. Suppose, you have 50 tweets (messages only) that are stored in 50 JSON arrays. You need to consider these data points as the total number of **documents (*N*)**. In this case *N*=50
    Now, use the search query "people", "condition", "weather" and search in how many documents these words have appeared.

Table 3.1: tf-idf table

Total Documents	50		
Search Query	Document containing term(df)	Total Documents(N)/ number of documents term appeared (df)	Log_{10}(N/df)
weather	30	50/30	0.221848749
people	5	50/5	1
condition	10	50/10	0.698970004

   b. Once you build the above table, you need to find which document has the highest occurrence of the word "weather". You can find this by performing frequency count of the word per document.

Table 3.2: term frequency table

Term	weather	
Canada appeared in 30 documents	Total Words (*m*)	Frequency (*f*)
tweet #1	6	2
tweet #2	10	1
:	:	:
tweet #30	8	1

   c. You should print the tweets (programmatically), which has the highest relative frequency. You can find this by computing (*f*/*m*).

*** Just to create tabular structure (table 3.1, 3.2), you can use a 3rd party library

---

Assignment 4 Submission Format:

1) Compress all your reports/files into a single .zip file and give it a meaningful name.

You are free to choose any meaningful file name, preferably - **BannerId_Lastname_firstname_5408_A5** but avoid generic names like assignment-5.

2) Submit your reports only in PDF format.

Please avoid submitting .doc/.docx and submit only the PDF version. You can merge all the reports into a single PDF or keep them separate. **You should also include output (if any) and test cases (if any) in the PDF file in the proper format (e.g. tables, charts etc. as required).**

3) Your code needs to be submitted on https://git.cs.dal.ca/