**Assignment 3 (10%)**
**Date Given: Nov 20, 2019**
**Submission Due: Dec 5, 2019 at 11:59 pm (midnight)**
**** Late submissions are not accepted and will result in a 0 on the assignment**

## Objective:

This assignment covers concepts related to data analysis. The primary objective of this assignment to use concepts and tools related to Semantic Analysis, Sentiment Analysis, and Business Intelligence. Consider this assignment as the third and last phase of an industry project.

## Grading Scheme:

- Data processing and Sentiment Analysis: 30%
- Data processing and Semantic Analysis: 30%
- Creation of correct Fact and Dimension tables, and BI Framework: 25%
- Working on BI Tool(s) and Query processing: 10%
- Adding citation in IEEE/ACM Format only. Use reliable information source: 5%

## Academic Integrity:

- This assignment does not require group work. Therefore, each student is expected to complete their work by themselves. Collaboration of any type amounts to a violation of the academic integrity policy and will be reported to the AIO.
- Do not copy texts verbatim from online or printed materials
- Do not copy texts from other's work
- Do not submit other's work
- If you obtain help from Tutor(s), please acknowledge
- Provide citation for texts, images, tables, data etc.
- 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

## Hypothetical Scenario:

A Company, "Analytics-5408" is trying to establish its position in the business domain in Canada. Recently it hired you as an information specialist to work for its client (Dalhousie University). You designed the Data Models, and gathered some data for the client. Now, the project has extended, and you started working on a content management system (CMS), which will have information related to Education in Canada. The project has two components,

(1) Data management, and
(2) Business Intelligence

In this phase, the project focuses on data analysis. Data extracted from Twitter, and NEWS API will be used to perform semantic, sentiment analysis. As an information specialist you will design a BI framework for Dalhousie University, which will be used to improve the University performance, and reduce operation cost.

## *** Your Tasks for this Assignment ***

### A. Sentiment Analysis:

1. To perform this task, you need to consider the tweets ("messages" or "texts" only, ignore metadata). Processing 500+ tweets should be enough.
2. Write a script to remove URL and/or any special characters. (Online code not accepted)
3. Write a script to create bag-of-words for each tweet. (Online code not accepted)
   e.g. tweet1 = "hey i m happy in Canada"
   bow1 = {"hey":1, "i":1, "m":1, "happy":1, "in":1, "Canada":1}
4. 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).
5. Tag each tweet as "positive", "negative", or "neutral". You can add an additional column to present your finding.

| Tweet | message | match | polarity |
|---|---|---|---|
| 1 | hey i m happy in Canada | happy | positive |

6. Visualize the most frequently occurring words in the positive and negative tweets you collected in a word cloud using Tableau.

### B. Semantic Analysis:

7. Write a well-formed script/program to clean and transform the news articles (previously obtained NEWS API data).
   (Do not use any online program codes or scripts. All cleaning and transformation logic must be written by you. You cannot copy any method from another online available program).
8. Consider a clean chunk of text (such as one news article) as a document. Use the news file(s) that you obtained in Assignment 2.
9. Each news file should contain "title", "description", and the news "content".
10. Use the following steps to compute TF-IDF (term frequency-inverse document frequency)
    a. Suppose, you have 500 news articles that are stored in 500 text files. You need to consider these files as the total number of documents ($N$). In this case $N$=500
    Now, use the search query "Canada", "University", "Dalhousie University", "Halifax", "Canada Education", and search in how many documents these words have appeared.

| Total Documents | 500 | | |
|---|---|---|---|
| Search Query | Document containing term(df) | Total Documents(N)/ number of documents term appeared (df) | $Log_{10}$(N/df) |
| Canada | 20 | 500/20 | 1.39 |
| Halifax | 40 | 500/40 | 1.09 |
| Dalhousie University | 10 | 500/10 | 1.69 |

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

| Term | Canada | |
|---|---|---|
| Canada appeared in 20 documents | Total Words ($m$) | Frequency ($f$) |
| Article #1 | 600 | 3 |
| Article #2 | 200 | 5 |
| : | : | : |
| Article #20 | 400 | 2 |

    c. You should print the news article, which has the highest relative frequency. You can find this by computing ($f/m$).

## C. Business Intelligence:

Build a BI framework for "*Analytics-5408*", where you need to focus on Dalhousie University data model. Perform the following steps:

11. Define the main facts to be analyzed (Hint: These facts become the source for the design of the fact table). E.g. Number of departments, and/or number of professors, courses, programs etc.
12. Define and describe appropriate dimensions. (Hint: These dimensions become the source for the design of the dimension tables.). E.g. location, time, programs (e.g. MACS, MCS..) etc.
13. Define the attributes for each of the dimensions. E.g. program dimension can have program code, program description, degree level etc.
14. Recommend the appropriate attribute hierarchies
15. Implement your data warehouse design, using the star schema or snowflake schema. Include your star or snowflake schema for each of the domains in the report.
16. Connect your database to Cognos BI or Power BI and create the report
17. Perform multidimensional data analysis using the concept of OLAP that you have learned in the BI lesson. Use Cognos BI or PowerBI and necessary tools.
18. Using your BI framework, can you answer the following?
    a. Does Computer Science offer highest number of programs?
    b. How many courses are there in each department or faculty?

## Submission Instruction:

- Create a Folder with your name and B00 number, and store all your files –
  - PDF files containing answers, tables, charts etc.
  - Screenshots of your cloud/ local server dashboard, processing of data, and output.
  - Program or script file (Source Code)
  - Any dictionary or supporting file(s) required for the program to run
  - An output file (if applicable). You may also include output file as part of the PDF file

- Compress the folder and create a .ZIP file (do not use other compression formats)
- Upload the .ZIP file on Brightspace.
- Submission Due: **Dec 5, 2019 at 11:59 pm (midnight)**