Assignment 2 (10%)
Date Given: Feb 12, 2019
Submission Due: Feb 26, 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 BigData and NoSQL, and research phase of a data management project. Consider this assignment as the second phase of an industry project. The designed bigdata framework and data gathered in this assignment will be used in the next assignments.

Grading Scheme:

- Cluster Setup: 20%
- Data Extraction & Transformation: 40%
- Data Processing: 30%
- Answers of question given in section D: 10%

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:

HalifaxInfo is a startup in Halifax, which is planning to build a data management portal for the Halifax region. The system can be conceptualized as a content management system. The project has three components,
(1) Data management,
(2) Visualization-Analytics, and
(3) Front-end design.

HalifaxInfo is trying to identify key performance indicators (KPIs) in the Halifax region to improve the *business*, *education*, *lifestyle*, and *safety*. In the first phase of the project, the company has gathered relevant structured data and information that are collected by various sources. The company found key entity sets, their attributes, and relationships that are critical to the future system.

The second phase of the project focuses on implementing a BigData infrastructure and processing data extracted from Twitter. The company believes 280 characters in a tweet on "Halifax" may contain essential information related to the city, which might help improve the *business*, *education*, *lifestyle*, and *safety*. In this phase, a pilot research is performed on bigdata processing.

*** Your Tasks for this Assignment ***

CSCI-5408 assignment series covers the (1) Data management, and (2) Visualization-Analytics component. The first two assignments will focus on the data management component, and the next two assignments will focus on the visualization-analytics component.

As an *information specialist,* you are expected to perform a series of tasks that are mentioned in section A to section D. To start building the project, *HalifaxInfo* has identified a keyword, "Halifax",

Specific Tasks (section A to D) – complete them as specified and submit on Brightspace based on the submission instructions on page #3

A. Cluster Setup:

1. Create a cloud account (if you do not have one) with any cloud service provider.
2. Initialize Apache Spark on your cloud account. Follow the tutorials provided in Lab 4
3. Install MongoDB to store the data

B. Data Extraction & Transformation:

4. Create a Twitter developer account
5. Explore the Twitter search and streaming APIs and data format
6. Write a well-formed script/program using (Java or Python or php or Perl) to extract data from Twitter. (Do not use any online program codes or scripts. You can only use API specification codes given by Twitter)
    a. The search keyword is "Halifax"
    b. You need to extract the tweets related to the given keyword
    c. The method/program querying search API should run for 1 hour (minimum 1000 data points)
    d. The method/program querying streaming API should run 6 times (2 times/day for 3 days)
    e. You should extract tweets, and retweets along with provided meta data, such as location, time etc.
    f. The captured data should be kept in MongoDB.
7. The data you captured from tweets using search/streaming APIs could be cleaned and transformed before uploading to the cloud infrastructure.
    a. Remove special characters, URLs, emoticons etc.
    b. You can upload the JSON/XML/TXT etc. files containing the tweets to cloud

C. Data Processing (MapReduce):

8. Write a MapReduce program (Following the structures given in Apache tutorials) to count (frequency count) the following substrings or words. You can perform an exact matching
    a. "not safe"
    b. "safe" (must exclude "not safe", "no safe" etc.)
    c. "accident"
    d. "long waiting"
    e. "expensive"
    f. "friendly"
    g. "snow storm
    h. "good school" or "good schools"
    i. "bad school" or "bad schools" or "poor school" or "poor schools"
    j. "immigrants" or "immigrant"

k. “pollution”
l. “bus” or “buses”
m. “parks” or “park”
n. “parking”

D. Answer the Questions:

9. Which substring or word are most frequently used in your extracted tweets.
10. Can you justify the reason by performing a new data extraction using twitter streaming API? Keep the search keyword as “Halifax”.
11. If the latest captured tweets do not contain the mostly used word or substring, mention which of the word(s) or substring(s) are available in the new tweets.

Submission Instruction:

- Create a Folder with your name and B00 number, and store all your files –
  - PDF file with at most 2-page report that includes the following
    - Your cloud setup steps,
    - Data extraction process,
    - Cleaning process,
    - Sample JSON/XML/or any other formats of data file
  - Screenshots of your cloud dashboard, and processing of data
  - Program or script file (Source Code)
  - Any dictionary or supporting file(s) required for the program to run
  - An output file (.txt format). 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: **Feb 26, 2019 at 11:59 pm (midnight)**