Assignment 2 (10%)
Date Given: Feb 14, 2020
Submission Due: Feb 26, 2020 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 Graph Database, and visualization of a management project. Consider this assignment as a visualization & marketing phase of an industry project. The designed graph and data gathered in this assignment may be used in the next assignments.

Grading Scheme:

- Section A: Exploring and Understanding the Dataset: 5%
- Section B: Program for Data Extraction/Transformation: 25%
- Section C: Installation of the required tool(s): 10%
- Section D: Graph Building & Query: 40%
- Section E: Technical Report: 15%
- 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:

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 (CMS). 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 **second phase** of the project, the company is trying to gather structured data and information on Parks of Canada, which will be used as a marketing strategy for the company's CMS. Once the relevant data are extracted and cleaned using scripts/program the company will use a visualization tool and a graph database to show the relationships in the gathered data. This is a showcasing phase of the industry project.

*** Your Tasks for this Assignment ***

Section A: Visit the website (https://data.novascotia.ca/Lands-Forests-and-Wildlife/DNR-Camping-Parks-Reservation-Data-2016/4zt7-x443) and explore the dataset on Parks. Based on your observation, write minimum ½ page summary (Font: Type= Times New Roman, size=11) about the given dataset.

Section B: Write a program using Java/Python/C++ to perform the following:

- Extract all data on parks of Canada from the given CSV file, and create a new CSV file. (e.g. file1.csv)
- From the newly created CSV file (file1.csv), extract data on ParkName, State, PartySize, BookingType, RateType, Equipment, and create another CSV file (e.g. file2.csv)
- In the newly created CSV file (file2.csv), scan the Equipment column, and replace all "less than" with "LT" [e.g. less than 30 ft. after transforming LT30ft], and replace "Single tent" with "ST", and create another CSV (e.g. file3.csv).

Note:

- You cannot use spreadsheet/Excel charting tools to perform the extraction/ filtration/ transformation.
- Once you download the CSV file from the given URL, all operations on the CSV file must be performed using your program.
- Libraries such as "Beautiful Soup" is not allowed. You can use regex.

Section C: Explore Graph Database "Neo4j", and install it with the required dependencies.

Section D: Considering a subset of final CSV file (file3.csv) [Consider parks of "NS" only] as the input (file4.csv), build a graph using Neo4j. You must use Cypher query language for graph related operations.

- Consider each park as a node
- For each park node, add ParksName, RateType, PartySize, BookingType, Equipment as attributes
- All parks are part of Nova Scotia, so the parks must be connected using edges.
- Parks with identical "RateType" should be connected using a "NeghbourByRate" relationship or edge.
- Parks with identical "Equipment" should be connected using a "NeghbourByEquipment" relationship or edge.
- Once the graph is constructed using Neo4j, using analytics and visualization display the parks that supports highest number of PartySize.

Section E: Write a two-page technical report on Neo4j, where you will include your views on Neo4j, such as its significance, and limitations etc. In addition, mention if there is any suitable alternative approach to present the given dataset. [Font: Times, 11 pt., Single spacing]. You must include proper reference

Submission Instruction:

- Create a Folder with your name and B00 number, and store all your files –
  - PDF file with answers,
  - 3 CSV files,
  - Graph images (if any) in the folder.
- Program source code.
- Compress the folder and create a .ZIP file (do not use other compression formats)
- Upload the .ZIP file on Brightspace.
- Submission Due: Feb 26, 2020 at 11:59 pm (midnight)