Assignment 3 (10%)
Date Given: Mar 3, 2020
Submission Due: Mar 17, 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 BigData and NoSQL, and research phase of a data management project. The designed bigdata framework and data gathered in this assignment will be used in the next assignment.

Grading Scheme:

- Spark Setup: 10%
- Twitter Data Extraction & Transformation: 25%
- News Article Data Extraction & Transformation: 15%
- OMDb Data Extraction & Transformation: 15%
- Data Processing: 30%
- 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). Th 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 this phase, the project focuses on implementing a BigData infrastructure and processing data extracted from Twitter, NEWS API, and Open Movie Database. The company believes tweets on “Canada”, “University”, “Education” etc. may contain essential information related to education in Canada and incoming. Similarly, news content and movies on or related to various aspects of Canada may provide meaningful information.

*** Your Tasks for this Assignment ***

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 Labs.
3. If you do not wish to work on cloud account, you must create local standalone Hadoop cluster to perform the operation.
4. Install MongoDB to store the data

B. (i)Twitter Data Extraction & Transformation:

5. Create a Twitter developer account (Approval might take 3 - 4 days, therefore, create account ASAP)
6. Explore the Twitter search and streaming APIs and data format
7. Write a well-formed script/program using (Java or Python or php or Perl etc.) to extract data from Twitter. (Do not use any online program codes or scripts. You can only use API specification codes given by Twitter - “tweepy”)
    a. The search keyword is “Canada”, “University”, “Dalhousie University”, “Halifax”, “Canada Education”.
    b. You need to extract the tweets related to the given keyword
    c. Running the method/program querying search API and streaming API for 3000+ records will be enough.
       Note: Working on small dataset will not use huge cloud resource or your local cluster memory.
    d. You should extract tweets, and retweets along with provided meta data, such as location, time etc.
    e. The captured data should be kept in MongoDB.
8. The data you captured from tweets using search/streaming APIs could be cleaned and transformed before uploading to the cloud infrastructure or local cluster.
    a. Remove special characters, URLs, emoticons etc. Retain “RT”
    b. You can upload the JSON/XML/TXT etc. files containing the tweets to cloud/ local cluster.

(ii) News Article Data Extraction & Transformation:

9. Visit the news API https://newsapi.org/
10. Create a developer account
11. Search the same keywords as mentioned before - “Canada”, “University”, “Dalhousie University”, “Halifax”, “Canada Education”, “Moncton”, “Toronto”.
12. Clean and format the data. You need to remove special tags (if any).
13. Upload your newly created files on the cloud server or local cluster in MongoDB.

(iii) Movie Data Extraction & Transformation:

14. Visit the news API http://www.omdbapi.com/
15. Request Free API key
16. Search keywords - “Canada”, “University”, “Moncton”, “Halifax”, “Toronto”, “Vancouver”, “Alberta”, “Niagara”.
17. Clean and format the data. You need to remove special tags (if any).
18. Upload your newly created files on the cloud server or local cluster in MongoDB

C. Data Processing (Spark):

19. Using Spark framework perform a frequency count (using MapReduce) of the following substrings or words. You need to consider the stored tweets and the stored news articles for frequency count.
    a. “education”
    b. “Canada”
    c. “university”
    d. “dalhousie”
    e. “expensive”
    f. “good school” or “good schools”
    g. “bad school” or “bad schools” or “poor school” or “poor schools”
    h. “faculty”
    i. “computer science”
    j. “graduate”

20. Write a separate program/ script to extract movie rating, genre, and plot from the stored movie data

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,
        - API setup process
        - Data extraction process,
        - Cleaning process,
        - Sample JSON/XML/or any other formats of data file
    - Screenshots of your cloud spark framework or local Hadoop cluster, and processing of data
    - Program or script files (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: **Mar 17, 2020 at 11:59 pm (midnight)**