Assignment 3 (10%)
Date Given: Mar 7, 2019
Submission Due: Mar 21, 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 Information Retrieval (IR), Semantic Analysis, and Sentiment Analysis. Consider this assignment as the research phase of an industry project. The designed IR framework and analysis in this assignment will be used in the last assignment.

Grading Scheme:

- Data Upload: 10%
- R1: Report on Sentiment Analysis: 20%
- Sentiment Analysis (tweets): 30%
- Semantic Analysis (Reuters): 30%
- R2: Report on findings and screenshots: 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 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. In the second phase, the company has setup a Big data infrastructure and implemented a parallel computing framework using MapReduce. In the first two phases, the company gathered relational data and unstructured data (tweets) on Halifax and performed some basic study. In addition to sentiment analysis of tweets, *HalifaxInfo* has decided to launch a pilot project, which will analyze some historical news data. The primary objective of this pilot project is to identify the important documents, and terms related to "Canada".

The third phase of the project focuses on implementing an Information Retrieval infrastructure to perform Sentiment and Semantic Analysis of captured/ collected data. The company believes data gathered from tweets, and news articles on "Halifax", and "Canada" may provide essential information related to the city of Halifax, and Canada. Furthermore, this research phase will help to identify polarity of topics that are frequently discussed on social media.

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

As an *information specialist,* you are expected to perform a series of tasks that are mentioned in section A to section D.

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

A. Data Upload:

1. Use your cloud instance(s) or local server to load "tweets", and the given dataset ("reuters").
    a. If you have less number of tweets (<1000), then run your Assignment 2 tweet extraction engine and extract new data points.
    b. The reuters dataset contains twenty two (.sgm) files, which must be uploaded on your local or cloud server.
    c. You can perform your analysis on tweets on the cloud server, and run the pilot project (reuters data analysis) on your local server. This will reduce your load on the cloud server.

B. Exploratory Study (R1):

2. Perform a preliminary research on "sentiment analysis", and produce a two-page summary. You should not copy-paste any content from the source(s). Understand the concept and write the summary. (Do not forget to cite the source, and avoid using Wikipedia as a source)

C. Data Extraction, Transformation & Analysis:

Tweets

3. Write a well-formed script/program to clean and transform the tweets.
   (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).
4. You do not need to consider the meta data, such as time, place, ID etc.
5. For the tweets, you need to perform a basic sentiment analysis.
   Hint: How are you going to handle stop words, stemming, and parts of speech. May be a dictionary is helpful.
6. Your final sentiment analysis script should tag a tweet as positive or negative.

Reuters data

7. Write a well-formed script/program to clean and transform the news articles.
   (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).
   Hint: For this semantic analysis, you do not need to work on stop words, special characters, and symbols.
8. Consider a clean chunk of text (such as one news article or one paragraph etc.) as a document.
9. Utilizing the concept of vector space IR model, determine the TF-IDF (term frequency – inverse document frequency).
10. Use "Canada" as your search query, and retrieve all the documents related to Canada. You need to rank the documents based on distance calculation as explained in Lecture 12.
    (e.g. consider "Canada" appeared in 3 documents, based on distance calculation we ranked the documents "reut_test11.txt", "reut_test5.txt", reut_test0.txt" etc.
11. Finally, retrieve sentence(s) or paragraph(s) from the top ranked document, where Canada appeared multiple times.

D. Report (R2) and Screenshots:

12. Write a report on your analysis. You should include the following in the report (R2)
    a. How many tweets are positive.
    b. How many tweets are negative.
    c. A sample tweet with polarity.
    d. For reuters, how many documents you created after cleaning
    e. The extracted sentences or paragraphs from the top ranked document, where "Canada" appeared multiple times.
13. Take screenshots of your cloud/ local server dashboard, and your output.

Submission Instruction:

- Create a Folder with your name and B00 number, and store all your files –
    - PDF files of R1, and R2.
    - 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 (.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 21, 2019 at 11:59 pm (midnight)**