DSCI 5350 – Big Data Analytics **Assignment# 1: Due Date: 02/28/2020 11:59pm**

Twitter Data Analysis – 50 points

Assignment Details:

In Assignment 1, you will perform a rudimentary sentiment analysis using Hive. The purpose of this assignment is to 1) make you comfortable with using Hadoop, in general, and Hive, in particular, and 2) allow you to understand the capabilities of the tools so that you can perform more complex analysis using these tools in future. To perform this analysis, you are provided with two files:

- Twitter.json – a JSON file with Twitter data
- Dictionary.txt – a dictionary file with rating for individual words

You are allowed to enhance the dictionary.txt file, or find a better one online. You are also allowed to process and/or simplify the JSON file using Hadoop tools. However, simplification of the JSON file outside of Hadoop is not allowed.

Below are the details about the task:

1. Correctly process and store the files in Hive. All tables created for the solution must have your student_id as a prefix to table name. For example, if I were to store the dictionary table, I would name it *dictionary_ks0776* (5 points)
2. Using the files provided, answer the following:
   a. What were the hashtags used in the file, and how many times each hashtag was used? (10 points)
   b. Identify the most trending hashtag by the day. How many times the most trending hashtag was tweeted? (10 points)
   [Note: day should be in the format 'yyyy-mm-dd']
   c. Determine the score for each tweet that was posted? Identify whether the tweet had a positive or negative sentiment? Use the dictionary.txt file for determining the score. Note: Include the date ('yyyy-mm-dd'), tweet_id, user_name, and the score in the resulting query. (20 points)
3. Propose a better solution for the sentiment analysis as compared to 1(c). Cite the source. (5 points)
   Note: You just need to provide the solution, you are not required to solve the problem using the solution.

Deliverables:

Submit the screenshots in a word document as mentioned below:

a. You must provide the screenshots of each query run, along with the answers on the terminal. The screenshots must include your name of the taskbar.

b. Under the screenshots, provide answers to each of the problem solved.

Plagiarism:

Given the nature of the assignment, it is highly improbable that two students approach the problem in the exact same manner. The grader will be tasked to look for similarities in the assignment. Any cases of plagiarism will result in a zero in the assignment.