CSC 4760/6760 Big Data Programming

Assignment 2

Due Date: 11:59 pm, September 25th, 2019

Point value: (100 points – Undegraduates, 140 points graduate students)

Dataset:

The StackExchange data science dump used **found on iCollege – Datasets Table of contents section**.

Or found here: https://archive.org/details/stackexchange

We need the datascience Users.xml, PostHistory.xml and Comments.xml files only.

Problem:
With Spark (and Python), use the Users, PostsHistory and Comment files to answer the following questions on a Jupyter Notebook. You can use only an RDD, and you must use lambda functions, transformations and actions and no SparkSQL is allowed:

1) From the Users.xml file, find all users which are from Georgia and output to screen their DisplayName only. (20 points)
2) Using the Users.xml file, provide the count for all users which joined (CreationDate) in 2017. (30 points). Output this to the screen.
3) Using the PostHistory file, count the number of Posts that feature the words "Spark" and "Scala". Output this to the screen. (20 points)
4) Using the PostHistory file, provide a total count of the words used by each distinct user. In other words, count all words in all posts for each user and display this to screen. You can only identify users by the UserID (30 points). You get 15 bonus points if you get the actual DisplayName of the user.
5) GRADUATE STUDENTS: Using the users.xml, comments.xml and PostHistory.xml files, produce **a single file** that includes the following information: DisplayName, Number of Comments, total Score and Number of posts. This file should have the users (DisplayName) sorted by score, descending from higher to lower. (40 points)

NOTE: You CAN NOT use Colab for this assignment, you must submit a Notebook created on your own environment. Any Colab files will not count.

Submission Materials:

a) Your Notebook file, with the question in a text block, the code in a code block, and the executed answer.
b) A PDF print out of your Notebook file.
c) Submit all on a zipped file through iCollege, the zip file must be called Assignment2_LASTNAME_FIRSTNAME.zip
d) Link (in the iCollege submission folder) to your Github repository (Named Assignment 2 – Spark). The repository must have your Notebook file and the zipped file***. Files without this naming convention will not be accepted. And you will receive a zero on the assignment.***