Homework 1: Tweet analysis with MapReduce

Data Intensive Computing Spring 2020

**Due Date: March 6, 2020 at 11:59 pm**

In this homework, you’ll write a MapReduce algorithm to analyze sample twitter dataset containing approximately 3.8 million tweets.

• Install Hadoop to your own server or use cs433.cse.unr.edu.

• You need to use jump host to access cs433.cse.unr.edu from outside of UNR campus. So, you can first login to nxlogin.engr.unr.edu and from there to cs433.cse.unr.edu

• Download ZIP file in here. Its size is around 405 MB. The files are already uploaded to HDFS in cs433.cse.unr.edu under “/” directory. Check by running “Hadoop dfs -ls /homework1/”

• Unzip the file and upload “training\_set\_tweets.txt” (tweets) and “training\_set\_users.txt” (users) files to HDFS

Once your Hadoop cluster is up and running do the following tasks:

• Show HDFS daemons (hint: search for processes called namenode, datanode) **(5 pts)**

• Show how many blocks created in HDFS for “tweets” file, either through command line or namenode web ui **(5 pts)**

• Show how many map tasks are created when you try to process “tweets” file in HDFS **(10pts)**

• Set the number of reduce tasks to 3 and show that Hadoop created 3 reduce tasks **(10 pts)**

• Write a MapReduce code to count the number of hash tags occurrences and find the most repeated 10 hashtags. **(20 pts)**

• Write a MapReduce code find the most tweeted 10 days. (Tweets are associated with time stamps so you need to count all the tweets posted in same days) **(20 pts)**

• Write a MapReduce code to find the most tweeted 10 cities along with the number of tweets (“training\_set\_users.txt” file has user\_id city relation to extract city information) **(30 pts)**

Important Notes

• It is NOT allowed to use global variables in Q5 and Q6 as they are easy to implement with single MR job.

• Although it is not an ideal solution, you can use a global variable in Q7 to keep the solution simple. However, I offer 10pt bonus points if you implement without using a global variable. You'll need to write multiple jobs in one application and use reduce-side join to implement this way.

What to deliver

Create following files/folders and compress them in a single zip file with name <**LASTNAME>\_<NAME>\_HW1.zip** and submit on WebCampus

• Take screenshots for Question 1-4 to a file answers1-4.pdf

• Copy the most repeated 30 hashtags along with number of occurrences to a file called “popular\_tweets.txt” file

• Copy the most tweeted 20 days along with number of tweets to a file called “most\_tweeted\_days.txt” file

• Copy the most tweeted 10 cities along with number of tweets to a file called “most\_tweeted\_citites.txt” file

• Create three directories Q5, Q6, and Q7 and copy your source code for question 5, 6, and 7 into those directories.

**• [Important]** Create **README** file that shows how to run compile and run your code

**• [Important]** Do not include input files in your final submission

Statement on Academic Dishonesty (from syllabus):

"Cheating, plagiarism or otherwise obtaining grades under false pretenses constitute academic dishonesty according to the code of this university. Academic dishonesty will not be tolerated and penalties can include filing a final grade of "F"; reducing the student's final course grade one or two full grade points; awarding a failing mark on the coursework in question; or requiring the student to retake or resubmit the coursework. For more details, see the University of Nevada, Reno General Catalog."