CMPSC/DS 410: PROGRAMMING FINAL

Instructions:

(1) **What you can access**
   (a) the cluster
   (b) the canvas pages for this class (but no external sites, even if linked from canvas)
   (c) your github classroom repository (e.g., on github.com or cluster)
   (d) tophat
   (e) gradescope
   (f) zoom for recording your work.
   (g) nothing else: no email, no google, no how-to sites, no facebook, no netflix, twitter, no other zoom sessions, etc. No chatgpt, no catgpt, no other large language models, no small language models, no medium language models. Improper internet access will result in an academic integrity violation.

(2) **Zoom**
   (a) Share full screen (not a window).
   (b) Record (only end recording after you have submitted to gradescope).

(3) Get the exam using updatestarter.

(4) **Submiting your exam:** gradescope

(5) **if terminal is not working, use Putty to connect to the cluster**

Question 1 (MapReduce)**.** This question uses the dataset in **/datasets/facebook**. In this dataset, each line is a pair of nodes (left node and right node) separated by a space. For each node, we want to know how many times it appears on the left and how many times it appears on the right, but only if the **total number of appearances is 3 or more**. So, if **234 [4, 5]** appears in the output, it means that node 234 appeared 4 times on the left and 5 times on the right. The output should not have a line like **432J [1, 1]** because node 432J only appears 2 times.

- Write mapreduce code to solve this problem.
- Fill in mr/mr_final.py
- Fill in the run script
- Fill in the test data files and expected output.
- **Your code must use a combiner to get full points.**

Question 2 (Spark RDD)**.** Now do the same problem with spark RDDs (no dataframes at all for this question).

- Fill in the file rdd/rdd_final.scala
  - The doFinal() function should do the main processing
  - Also fill in getSC(), getRDD(), getTestRDD(), expectedOutput()
  - Do not change main() or saveit()
- Fill in the other files needed for compilation.

Question 3 (Spark DataFrames)**.** Now do the same problem with spark DataFrames (no RDDs at all for this question).

- Fill in the file df/df_final.scala
  - The doFinal() function should do the main processing
  - Also fill in getDF(), getTestDF(), expectedOutput()
  - Do not change main() or saveit()
- Fill in the other files needed for compilation