Project 6: MapReduce and network programming (100 points)

Submission guidelines:

You will submit **2 files** in Canvas.

- a PDF report file (filename: **Project_6_Report.pdf**)
- a ZIP file (filename: **LastNameFirstName_Project6.zip**) containing:
  - `Problem_1/`
    - `find_longest_words_MRJob.py`
  - `Problem_2/`
    - `TCPClient_multiply.py`
    - `TCPServer_multiply.py`
    - `UDPClient_multiply.py`
    - `UDPServer_multiply.py`

Notes:

A rubric is available in Canvas under the assignment

We will deduct **5% of the points** if you don't follow the code specifications AND submission guidelines and no matter whether the code contains the correct solution or not.

Problem 1. [CS4473 and CS5473] (40 points)

Please download “Pride and Prejudice” from http://www.gutenberg.org/files/1342/1342-0.txt. The text file, 1342-0.txt, is also available on Canvas.

Your assignment is to write a MapReduce program using MRJob to find the longest word starting with each letter from a to z. If multiple words have the same length, they should be all listed. A part of expected result is shown below:

```
"r"    ["recommendations", "representations"]
"s"    ["superciliousness"]
"t"    ["thoughtlessness"]
"g"    ["gentlemanlike", "gratification"]
"h"    ["hertfordshire", "healthfulness"]
```

Please read the instructions how to install and run MRJob on your computer.
https://mrjob.readthedocs.io/en/latest/

Learning outcome:

This is designed for you to write a MapReduce program and test it on your local machine.

You don’t need to complete this, because the setup is complicated. But if you are interested, you can try creating a Hadoop cluster on GCP and run MRJob using GCP Dataproc.
https://mrjob.readthedocs.io/en/latest/guides/dataproc.html
https://cloud.google.com/composer/docs/tutorials/hadoop-wordcount-job
In real-world applications, developers can test their MapReduce codes with a small test dataset on a local machine and then do production analytics of a large dataset on a Hadoop cluster.

What to submit

Please submit the following Python script.

- find_longest_words_MRJob.py

Make sure that the program will run with the following command:

- python find_longest_words_MRJob.py 1342-0.txt

In the report, please briefly describe the major steps that the MapReduce framework performs behind the scene to run this MapReduce program.

Problem 2. [CS4473 and CS5473] (40 points)

Suppose that multiplication is a computationally expensive calculation and you have developed a killer network application for multiplication which runs on a server equipped with a powerful CPU. Users can send a list of numbers to the server and receive their product back. The multiplication is done “in the cloud”, instead of locally on their own computer.

Please implement both the client program and the server program in both TCP and UDP. The client program will prompt the user to enter a list of comma-separated numbers, send the server this list of numbers, receive the product back, and print it out. The server program will receive a list of number, perform the multiplication, and send the product or an error message back.

In the client programs, please specify the IP address of the server program to be “127.0.0.1” so that they can be tested in your local computer.

You may use the TCP and UDP programs on Canvas as examples for this programming problem.

Application-layer Message Protocol

The message from the client to the server is a list of comma-separated numbers in the following format:

number1, number2, number3, …

The message from the server to the client is either a string “*Invalid input*” or a number for the product of all the provided numbers.

Learning outcome:

This is designed for you to write a simple network application using both UDP and TCP.

What to submit

Please submit the following programs.

- TCPClient_multiply.py
- TCPServer_multiply.py
- UDPClient_multiply.py
- UDPServer_multiply.py

  In the client programs, please specify the IP address of the server programs to be “127.0.0.1” so that they can be tested in our local computer.

In the report, please discuss the pros and cons of using UDP versus TCP for this network application.

Problem 3. [CS4473 and CS5473] (10 points)

You have worked on many parallel programming problems in the six projects in this course. Please provide some feedback on these programming problems.

Which problems are the most interesting for you to work? Which problems do you find the most difficult to solve? Which problems do you think are not well-designed? What suggestions do you have that would improve your learning experience with the projects?

Problem 4. [CS4473 and CS5473] (10 points)

This is the third year that this course is offered over the pandemic. In this iteration, you may (a) attend the lectures synchronously in person, (b) attend the lectures synchronously in Zoom, (c) watch the lecture Zoom recordings asynchronously, and (d) watch the slide narration recordings asynchronously. Which learning format(s) did you take and why? What suggestions do you have that would improve your learning experience in the lectures?