CS6502: Week 5 Lab Exercises

Dr. Andrew Ju

February 22, 2021

Exercise 1

The lab in this week is **supervised**. The lecturer/lab assistant will be available between 14:30 – 16:30 for any questions you may have.

The main task for today's lab is to continue getting familiar with Google Cloud Platform, and practice the WordCount example in Python.

In Sulis class page, there is a forum "Labs (week 5)" created. In the forum, there are two topics:

- *Discussion*, use this topic to post your questions and discuss with your fellow classmates.
- *Submissions*, use this topic to submit your video recording of the code practice.

The deadline is 12am (noon time) of Sunday, 28th Feb.

Below are steps that you should follow for this week's lab exercise.

1. Quickstart using the console [Please only read through, and do not enable the API in this step]
2. Dataproc pricing [Task: Understand how to estimate the cost of running a Dataproc job[1]]
3. Practice the WordCount example in your preferred Python environment. [Task: pick a book from Project Gutenberg, then run the Python scripts with selected book. Note down the size of the book (in txt format), and how long it takes to complete the run.]
4. [SE students only] Use Hadoop Streaming to run the WC example on your single node cluster. [Task: make a video recording of this step, in the recording make sure to include how you copy data to HDFS, run the job and view the result.]
5. Submission requirement:
    - As usual, use your student Id as the conversation title

[1] From next week, you will be practicing on Dataproc, and you should know how to estimate the cost in advance.

- In the message: write down the book you selected, size of the book (in txt format), and how long it takes for the Python scripts to complete
- For SE students, in addition to above, please include the running time on your Hadoop cluster, and attach the video recording.
- Click submit.

6. That's it, good job!