Project 1 - due 9/23 at 12pm

Write one Jupyter notebook with your solutions to all of the following problems. Document each step of your process in a reproducible manner; when you are finished, your instructor and anyone else should be able to run your notebook from start to finish without error.

Add text notes on your process as appropriate, documenting any assumptions and explaining key decisions you make along the way. Use markdown cells for this text.

Be sure to answer each question directly and precisely, using the data to justify your answers.

As always, you are welcome to seek and give assistance to others who might become stuck along the way. Please acknowledge any assistance you receive. At the same time, you must each perform and submit your own work, in your own voice, in accordance with the GWU Code of Academic Integrity.

This assignment is due on Friday, September 23, at 12pm. Package your notebook and any additional script files into a single zip file and submit that zip file to Blackboard. In addition, commit and push your individual project files to your repository in Github.

Problem 1 - Word counts (40 points)

Part A. Characters in Little Women

How many times are each of the following characters mentioned by name in the text of Little Women?

- Jo, Beth, Meg, Amy

Use the text available at https://raw.githubusercontent.com/gwsb-istm-6212-fall-2016/syllabus-and-schedule/master/projects/project-01/women.txt for this part.

Part B. Juliet and Romeo in Romeo and Juliet

How many times do each of the characters Juliet and Romeo have speaking lines in Romeo and Juliet? Keep in mind that this is the text of a play.

Use the text available at https://raw.githubusercontent.com/gwsb-istm-6212-fall-2016/syllabus-and-schedule/master/projects/project-01/romeo.txt for this part.

Problem 2 - Capital Bikeshare (40 points)

Use the data available at https://raw.githubusercontent.com/gwsb-istm-6212-fall-2016/syllabus-and-schedule/master/projects/project-01/2016q1.csv.zip for this problem.

Part A (20 points)

Which 10 Capital Bikeshare stations were the most popular departing stations in Q1 2016? Which 10 were the most popular destination stations in Q1 2016?

Part B (20 points)

For the most popular departure station, which 10 bikes were used most in trips departing from there? Which 10 bikes were used most in trips ending at the most popular destination station?

Problem 3 - Filters (20 points)

In class lectures, previous exercises, and Problems 1 and 2 above, you use Unix commands like `grep` and `tr` as filters, changing the text lines streaming through the pipeline. In this problem, write small Python programs that act as filters in the same way, where each program serves one filtering purpose. Name each new filter program clearly.

For this problem, use the basic Python filter template shown in class and available at https://raw.githubusercontent.com/gwsb-istm-6212-fall-2016/syllabus-and-schedule/master/projects/project-01/simplefilter.py as the basis of your own filters.

Part A (10 points)

Demonstrate a pipeline that performs a count of the top ten unique words in Little Women. This may be exactly the same pipeline we have used before.

Write a Python filter than replaces `grep -oE '\w{2,}'` to split lines of text into one word per line, and write an additional Python filter to replace `tr '[:upper:]' '[:lower:]'` to transform text into lower case.

With your two new filters, repeat the original pipeline, and substitute your new filters as appropriate. You should obtain the same results.

Part B (10 points)

Write a Python filter that removes at least ten common words of English text, commonly known as "stop words". Sources of English stop word lists are readily available online, or you may generate your own list from the text.

Add your stop word filter to a word count pipeline and show the top 25 words in Little Women with stop words removed. You may re-use your filters from Part A if you wish, although this is not required for full credit.

Extra credit (10 points)

Use GNU parallel to count the 25 most common words across all the 109 texts in the zip file provided, with stop words removed. You may re-use your filters from Problem 3.

Use the texts available at https://raw.githubusercontent.com/gwsb-istm-6212-fall-2016/syllabus-and-schedule/master/projects/project-01/texts.zip for this part.