COMP 598 Homework 2 – Unix server and command-line exercises

30 pts
Assigned Sept 17, 2020
Due Sept 25, 2020 @ 11:59 PM

This is an INDIVIDUAL Assignment – each student's work must be their own, each student completes this assignment, there are no teams for homework 2.

The goal of this assignment is to for you to get more familiar with your Unix EC2 – both as a data science machine and as a server (as a data scientist, you'll need it as both).

Task 1: Setting up a webserver (15 pts)

The objective of this task is to setup your EC2 instance to run an apache webserver on port 8008. Your goal is to have it serving up the file comp598_hw2.txt at the www root. In other words, my web browser can access it at http://X.Y.Z.W:8008/comp598_hw2.txt, where X.Y.Z.W is the public IP address of your EC2. The file itself should be empty.

Task 2: Data filtering (15 pts)

Write a bash shell script stats.sh that accepts as a command line argument a tweet file (similar to the one in homework 1) and prints out, on subsequent lines:

- The number of lines in the file
- The first line of the file (i.e., the header row)
- The number of lines in the last 10,000 rows of the file that contain the string "potus" (case-insensitive).
- Of rows 100 – 200 (inclusive), how many of them that contain the word "fake"

All this should be done only using standard Unix commands and pipes.

To be clear, your script should work on any file which has at least 10,000 lines in it. It will be tested by TAs by calling it using:

```
stats.sh <test_file>
```

The TAs will get to choose the test file.

Submission Instructions

Your MyCourses submission should contain – at minimum - the following:

- ip_address.txt
  - contains exactly one line containing the public IP address of your EC2 instance.
- stats.sh – the bash file that performs the tasks described above

In addition to the submission file:

- Leave your EC2 server running until Sept 28 @ 11:59 PM so that the server can be checked for the file.