Quiz 02: Working with large files

Tumble down the rabbit hole into Wonderland, where nothing is as it seems [50 points]

Introduction

In the land of Big Data, the usual techniques don't quite work and we must be prepared for surprises. This quiz is designed to have you experience some of the differences and workarounds.

Background

Over the summer of 2019, The Washington Post published a series of reports on the distribution of opioids in the US. Their data team published a portion of a database that tracks the path of every opioid pain pill, from manufacturer to pharmacy, in the United States between 2006 and 2012.

Details of how the data may be accessed are available here. This file represents the national data, it is about 7GB *compressed.* Uncompressed, it is about 75 GB in size.

Linux Commands

Most coding assignments for big data require students to run programs on the EECS server or Google Cloud Platform web instance, which both are linux-based servers. To prepare you for the incoming coding assignments/quizzes, we encourage you to learn about these frequently used linux commands (`man`, `ssh`, `ls`, `cd`, `pwd`, `mkdir`, `rm`, `cat`, `mv`, `cp`, `wc`). To find usage of each command, you're encouraged to use Google or the `man` command in a linux shell (Use Terminal app - MacOS OR login to Halligan/EECS Server).

1. Answer the following questions in one or two sentences each:
    a. [½ point] What does the `man pwd` command tell you about the pwd command? You should learn what the man command is if you tried with different options, such as man ls, man cd.
    b. [1½ point] Explain piping in Linux.
    c. [3 points] Explain the concepts of `stdin`, `stdout` and `stderr`.

Handling Compressed Data

For this part of the quiz, you may use your own laptop. On a Mac or a Chromebook, use Terminal. On Windows laptops, we recommend using the ubuntu[1]. If you don't have enough space on your laptop, you may use the eecs Linux servers[2]. The eecs Linux servers already have the compressed file available as `/comp/119/arcos_all_washpost.tsv.gz`.

2. [7½ points] Find the column names in the Opioid dataset. The naive way is to `gunzip` the .gz file and run `head -1` on the result, but you likely don't have enough disk space. Conveniently, `zcat` can read the file and write the unzipped contents into stdout, which can be piped into `head -1`.
3. [7½ points] Find the number of rows in the Opioid dataset by processing the `zcat` output, stripping the header row, and counting the remaining lines using `wc`.
4. [10 points] Find the names of all the drugs named in the dataset[3].

[1] If you don't have ubuntu on your Windows laptop, you may have to enable it. A quick google search will tell you how. If you run into any problems, please post a request on Piazza under qz1.

[2] Visit http://systems.eecs.tufts.edu/managing-your-password/ to obtain a login and password, then ssh linux.eecs.tufts.edu to access your linux account.

[3] There are multiple ways of solving this problem, but many of them result in "write failed: No space left on device." The task is to find one that works!

Analysis of the Opioid Dataset Using Sampling

5. [10 points] Estimate the number of rows for each year in the dataset[4]. There may be enough space in the shell, but we'd prefer not to rely on that. So here's a potential strategy: Use the `shuf` command to extract, say, random 7,500 rows from the output of `zcat`. Find the proportion of rows for each year in this extract. Assuming that the distribution of the random 7,500 rows is similar to the distribution in the whole file, estimate the number of rows for each year.[5]

6. [10 points] Extract any 4,000 rows for the year 2012 from `arcos_all_washpost.tsv.gz`. Write the resulting lines to `arcos_2012.tsv`.

Submit the answers to all problems as a single PDF file in Gradescope. All questions (except q1) have partial credit, and work/explanations/justifications ARE required. *Be sure to not paste screenshots into what you submit. Don't post pictures of text. We would like to cut'n'paste the input text to run your code to develop a solution, but we can't do that with screenshots. Please include all commands you issued and also any programs you wrote.*

[4] There is no year field in the dataset. The year is embedded in the TRANSACTION DATE.

[5] Make sure to avoid including the header in coming up with your answer.