MATH 3190 Homework 1

Essential Tools for Data Science (due 1/26/22) 1/15/22

**Advanced Unix Tools**

Most Unix implementations include a large number of powerful tools and utilities. (Unix has been in development for more than 50 years!). We were only able to scratch the surface in our class time. It will take time to become comfortable with Unix, but as you struggle, you will find yourself learning just by looking at man files and finding solutions on the internet. For this HW, you will explore several more advanced Unix functions. You can use any resource available to you–classmates, the internet, and Dr. Johnson. Ask all the questions you want, just make sure you do the work and you learn!

1. Learn about the tar function. What is a tarball? How is it different from a .zip file? Download the HW1.tar.gz file from Canvas and unzip the contents, and report that code you used. How effective is the compression for this tarball? After you complete Question [2,](#_bookmark0) add the basketball data to the directory and generate a gzipped tarball for all the HW1 data (plus the basketball data) and provide the code you used.

Zip is a collection of compressed files, .tar.gz is a compressed collection of uncompressed files. Code to decompress

$ tar -xf HW1.tar.gz

$ mv ~/Documents/Fundamentals\_of\_Data\_Science/cbbga22.txt ~/Documents/Fundamentals\_of\_Data\_Science/HW1

$ tar cvf output\_tarball.tar ~/Documents/Fundamentals\_of\_Data\_Science/HW1

1. Learn more about tools for downloading files from external servers (e.g., scp, ftp, sftp, rsync), and for to downloading data from webpages (e.g., curl, wget, mget). Use an appropriate function to download the scores for all college basketball games for the 2021-2022 season ([http://kenpom.com/cbbga22.txt).](http://kenpom.com/cbbga22.txt) Give the code you used to download these data

$ curl -O https://kenpom.com/cbbga22.txt

1. Research the chmod function. Give short explanation of what this function does, its syntax, and examples when you would use it. Practice chmod by changing the permissions on the ‘TB microbiome data.txt’ file in the HW1 directory from the previous questions. Give examples of the code you used and show that the code works (e.g., use ls -l).

This changes permissions. It has the format chmod [references][operator][modes] filename

References are u- user o-other g-group

Operator + to add permission – to remove it, = to set

Modes are r-read w-write x-exectue

chmod u+w .\_TB\_microbiome\_data.txt

1. The grep function is an extremely powerful tool for search (potentially large) files for patterns and strings. One advantage is that you don’t have to open the file to conduct a search! Using the internet, find a short tutorial on the basics of grep, and give the code and results for the following tasks:
   1. How many games has SUU played so far this season? (hint: search for ‘Southern Utah’ in the file)

15

$ grep -c "Southern Utah" cbbga22.txt

* 1. How many games have been played by teams other than SUU? (i.e., inverse search)

3262

$ grep -c -v "Southern Utah" cbbga22.txt

* 1. What was the score when SUU played Dixie St.? How many games had SUU played before that game? (i.e., add line numbers to the result)

$ grep -n "Southern Utah" cbbga22.txt

87-59

11 games before that

* 1. How many coronavirus genomes are present in the ‘virus.fa’ file? How many of these are SARS-COV-2?

$ grep -i coronavirus viral.fasta

3 coronavirus genomes, 1 is SARS-COV-2

* 1. How many times does the letter ‘A’ (capital or lowercase) appear in all the files from the HW1 tar file plus the college basketball data? (i.e., ignore case).

$ grep -c -i a \*

output

TB\_microbiome\_annotation.txt:31

TB\_microbiome\_data.txt:226

cbbga22.txt:2963

viral.fasta:10072

* 1. What *Staphylococcus* species are present in the ‘TB microbiome data.txt’ file? (hint: each separate microbe has its own row in the file). Print out the counts for *Mycobacterium tuberculosis*. How many *Streptococcus* species are present?

$ grep Staphylococcus TB\_microbiome\_data.txt

Staphylococcus aureus

Staphylococcus epidermidis

Staphylococcus haemolyticus

Staphylococcus saprophyticus

$ grep -c "Mycobacterium tuberculosis" TB\_microbiome\_data.txt

1

$ grep -c Streptococcus TB\_microbiome\_data.txt

16

1. Learn how to use less to display large text files in the terminal using the man help page. Using the “OPTIONS” section of the man page, open the ‘virus.fa’ file to display so that it does not wrap long lines (default), displays line numbers, and opens at the first occurrence of ‘coronavirus’. Provide the command you used to open the file in this way. Within less, learn and practice how to scroll forward/backward, scroll forward/backward *n* lines, jump to the middle or end of the file, and search for text in the document. When would it be advantageous to use less over a tool like Microsoft Word? Ask Dr. Johnson why in Unix more is less and less is more ,.

$ less -N -pcoronavirus viral.fasta

It is more useful than Word for large files, since it allows much faster navigation through them.

1. Open a text file in vim and change the file. How do you How do you move the the beginning/end of a line, insert text, copy and paste, delete text and lines? How do you save your file or exit vim with/without saving your result? What are the advantages and disadvantages of vim versus less? In which scenarios would you use each of these?

Use vim filename to open a file

i is used to enter insert mode where you can insert characters

0 to beginning of line, $ to end

d starts delete operation

d starts the delete operation.

dw will delete a word.

d0 will delete to the beginning of a line.

d$ will delete to the end of a line.

dgg will delete to the beginning of the file.

dG will delete to the end of the file.

u will undo the last operation.

Ctrl-r will redo the last undo.

v highlights character

V highlights line

CTRL V highlights columns

p paste text after the current line.

P paste text on the current line.

1. Learn about pipes and redirects in Unix. In which scenarios would you use them, and why are they helpful? describe what the following commands do:
   1. ls -l | less

opens the long version of the list of files in the working directory. Would be useful to search for see the file list in the less format.

* 1. ls -l > directory contents.txt

puts the list of the directory in a file called directory\_contents.txt

* 1. ls -l >> directory contents.txt

adds the list of directory contents to the directory\_contents.txt file

* 1. cat directory contents.txtt | head -3 | tail -2

selects the last 2 of the 1st 3 line of directory\_contents.txt

* 1. ls | grep -c html

searches for “html” in the list of files in the working directory

* 1. ls | wc -l

since wc-l counts the number of lines, this counts the number of files in the working directory

* 1. cat file1.txt file2.txt > file3.txt

creates a file3.txt that is file2.txt concatenated onto file1.txt

You can also use pipes in R! Investigate how to do this and give the code for a great example.

1. Learn about another Unix command that we have not discussed. Give a short de- scription of this function, when you would use it, its syntax, and give some examples of its use.

Locate is a good one. It is used similarly to windows search. It has the syntax

$ locate filename

-i makes it case insensitive, and asterisks are used when more than one word is contained. You’d use it to locate a file.

1. You need to complete the following two tasks for the Git and GitHub lecture:
   1. Fork the <https://github.com/wevanjohnson/my.package> directory and clone it to your local machine. Then add your name as an author in the DESCRIPTION file local repository and add a multiplication function to the R package (R folder). Then push the changes to your GitHub fork, and send me a pull request with your changes.
   2. Set up a git repository for this HW assignment on your computer (repo named “MATH 3190 HW”), add files/changes to it, and upload it to GitHub. This is how you will turn in your HW for this semester! (including this one!).

.