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, 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.

    - A tarball is a utility function used to collect a bunch of files into one archive file. The function's primary use is to create backup files.
    - The compression in a .zip file is already built-in and happens independently for every file in the archive whereas for a tar file compression is an extra step that compresses the entire archive.
    - Code used is in code file.

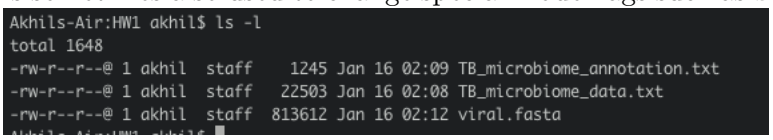
2. Learn more about tools for downloading files from external servers (e.g., `scp`, `ftp`, `sftp`, `rsync`), and for to downloading data from webpages (e.g., `curl`, `wget`, `mget`). Use an appropriate function to download the scores for all college basketball games for the 2021-2022 season (http://kenpom.com/cbbga22.txt).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

    - `scp`: Secure Copy Protocol, this protocol is used to move files onto servers. The benefit of using this is that it supports encryption and authentication.

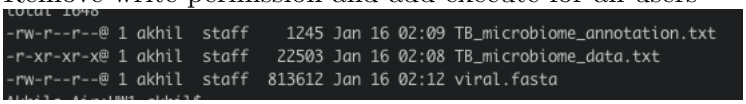
- `ftp`: File Transfer Protocol is used to download, upload, and transfer files from one location to another on the internet and between computer systems.
- `sftp`: Secure File Transfer Protocol is the same as `ftp` but has the use of `SSH` which is a secure way of doing things with files. Example: Allows businesses to securely transfer billing data, funds and data recovery files.
- `rsync`: Remote Sync, it is a remote and local file synchronization tool. The algorithm that is uses minimizes the amount of data copied by only moving portions of files that have been changed.
- `curl`: Client URL is a command used to transfer data to and from a server.
- `wget`: This is a free software for retrieving files using HTTP, HTTPS, FTP, FTPS. This will be a good command line tool to download files like the basketball data.
- Code used to download and save file is in the code file.

3. 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`).

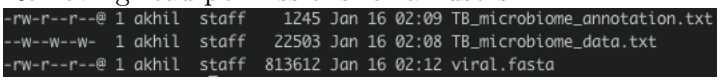
   - `chmod` is a command line call used to change access permissions to file objects. It is sometimes also used to change special mode flags such as `setuid` and `setgid`.



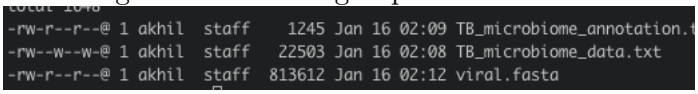
   - Remove write permission and add execute for all users



   - Removing read permissions for all users



   - Removing read access for group and other users



4. 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:

(a) How many games has SUU played so far this season? (hint: search for 'Southern Utah' in the file)

- Number of SUU games = 16

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

- Number of games other than SUU = 3414

(c) 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)

- Southern Utah 76 vs. Dixie State 83
- SUU had played one game before the game against Dixie.

(d) How many coronavirus genomes are present in the 'virus.fa' file? How many of these are SARS-COV-2?

- Total number of genomes = 10082
- Coronavirus genomes = 3
- SARS-COV-2 genomes = 1

(e) 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).

- There are 14,122 occurrences of the letter 'A' ignoring the case.

(f) 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?

- *Staphylococcus aureus*, *Staphylococcus epidermis*, *Staphylococcus haemolyticus*, *Staphylococcus saprophyticus*. There are 4 species of *Staphylococcus* present.
- There is one species of *Mycobacterium tuberculosis* present.
- There are 16 species of *Streptococcus* present.

5. 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 ☺.

   - The 'less' command can be used to read contents of a text file one page at a time.

- It would be useful to use less when we need to do some fast debugging or pattern finding in a file. It will also be useful to jump ahead n number of lines since it provides that functionality.
- Searching for a text can be done using the -p option. 'j' is used to go forward one line, 'k' is used to go backward one line. 'G' goes to end of file, 'g' goes to start of file. '10j' and '10k' are used to jump 10 lines forward and backward respectively.

6. Open a text file in `vim` and change the file. How do you move 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?

   - Opening a text file: vim cbbga22.txt
   - gg: moving to the start of the line
   - G: moving to the end of the line
   - i: insert into text file
   - yy: copy a line
   - yw: copy a word
   - p: paste whatever has been copied
   - d: delete highlighted
   - dd: delete line
   - dw: delete word
   - :w : save file
   - ;wq : quit file
   - Advantages: Vim can provide the ability to do super fast editing. Vim can be adapted to change color and backgrounds according to what the user wants. It can be run on any operating system. Since there is no need for a mouse it is better than an IDE.
   - You need to spend a lot of time to learn both vim and less. It takes a lot more effort to write reproducible code using less and also testing independent modules is quite hard.
   - If one would like to edit a file very quickly, vim would be the go to editor whereas if one wanted to check the contents of a file without having to open it up or search for a specific start in the file, one would go to use less.

7. 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:

- Redirection is used to redirect the stdout/stdin/stderr. Eg. ls > log
- Pipes are used to give the output of a command as input to another command. E.g. ls | grep file

(a) `ls -l | less`

- This will allow us to read the access permissions one page at a time. This will make it easy to look at the permission of a large directory.

(b) `ls -l > directory_contents.txt`

- This will create a text file with the contents and access permissions of the directory.

(c) `ls -l >> directory_contents.txt`

- This will create another copy of the directory contents and access permissions in the same file.

(d) `cat directory_contents.txtt | head -3 | tail -2`

- This command will print out the contents of the file, the −3 command will remove the last 3. The head command prints out the first 10 entries. The tail command is the complement to the head command and does the same.

(e) `ls | grep -c html`

- This will show the files in the directory and the grep function will then try to count how many *html* counts are in the directory.

(f) `ls | wc -l`

- This will count the number of files in the directory.

(g) `cat file1.txt file2.txt > file3.txt`

- This will combine the two files `file1.txt, file2.txt` into one file names `file3.txt`

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

- `mtcars <- mtcars %>% transform(cyl = cyl * 2)`
- This will multiply the cylinder column in mtcars by 2 and save it.

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

- SSH command

- The ssh command provides a secure encrypted connection between two hosts over an insecure network. This connection can also be used for terminal access, file transfers, and for tunneling other applications
- This has been such an important command for some of my projects especially the last one, home assistant software on a raspberry pi to automate the lights and appliances in the house. I had to use the ssh commands to log in to the raspberry pi and do lots of other things related to ssh. Here are some examples
- `ssh -v`: Displays ssh client version. To find out whether or not we need to update the ssh version or not.
- `ssh -l "name" remotehost.example.com`: To login to remote host
- `ssh@"IP_address"`: This will also log in to ssh client or device.

9. You need to complete the following two tasks for the Git and GitHub lecture:

   (a) 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.

   - Done!

   (b) 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!).

   - Done!

.