

An intro to shell commands and unix / linux

Practical Data Science

@jattenberg

May 2014

Unix: what is it?

- ✦ Operating system developed by AT&T in 1971
- ✦ Linux: A family of operating systems based on an open source “kernel” developed by Linus Torvalds in 1991
- ✦ Long history and open source have lead to wide adoption and very high performance.
- ✦ Basis for OS-X, with emulators in Windows (cygwin)
- ✦ Almost ubiquitous in modern web and data systems.
- ✦ Includes a powerful set of commands and utilities

Let's Dive in!

- ✦ OS-X: Terminal app or iTerm2
- ✦ Cygwin or connect to a unix server with PuTTY
- ✦ Unix shell in ipython: (http://nbviewer.ipython.org/github/jattenberg/PDS-Spring-2014/blob/master/ipython_notebooks/Basic%20Unix%20Shell%20Commands%20for%20the%20Data%20Scientist.ipynb)
- ✦ Example data here: <https://raw.githubusercontent.com/jattenberg/PDS-Spring-2014/master/data/sample.txt>
- ✦ Try downloading with wget or curl

Shell Script Files

- ✦ In addition to just typing commands in the terminal, one may save complex commands in a file.
- ✦ Typically, one logical command is given per line.
(commands in unix may also be separated by ;)
- ✦ Comment delimited by the # character
- ✦ By convention, the suffix of shell command files is .sh
- ✦ To execute the commands in the file, `sh your_file.sh`

Unix Example

- ✦ See here.

Cron

- ✦ Often one wants to automate some process, cron provides a mechanism for periodically executing a command according to a predefined schedule.
- ✦ To schedule a new periodic task, edit a systems “crontab”: type `crontab -e`

Cron

- ✦ Crontab has a special format to control the timing of automatic commands, followed by the command itself.
- ✦ [minute] [hour] [day of month] [month of year] [day of week].
- ✦ Wildcards (*) can be used to mean all in a specific granularity. Multiple values can be separated by commas

Cron

- ✦ Crontab has a special format to control the timing of automatic commands, followed by the command itself.
- ✦ [minute] [hour] [day of month] [month of year] [day of week].
- ✦ Wildcards (*) can be used to mean all in a specific granularity. Multiple values can be separated by commas
- ✦ / can be used to divide wildcards, eg, /2 for every 2 hours

Cron

- ✦ Example (write hi to the end of a file every minute)
- ✦ `0-59 * * * * * echo "hi" >> /tmp/hi.txt`
- ✦ Example (same thing, every 2 hours)
- ✦ `0 */2 * * * * echo "hi" >> /tmp/hi.txt`
- ✦ See: <http://www.nncron.ru/help/EN/working/cron-format.htm> for more info

Unix Assignment

- ✦ [See here](#) for details