Loading data and the command line

Charlotte Wickham

April 10, 2014

General points

https://github.com/gitelman/BigData/blob/master/Project1/01-get-data.R

- **start small**
    - there's no point reading in a million rows if you can't read in 10 without error.
    - there's no point reading in (50 states)/(10 years) if you haven't figured out what to do with one.
- **read function documentation**: functions are often written to do sensible things if you don't specify certain arguments, if you are specific about arguments the function doesn't have to waste time figuring out what is sensible
- **look outside R** - a tool specific to the job you need to do is often faster
- **split the problem up** into "small" pieces

Command line

A "typing" interface to your computer

You can automate what might ordinarily be a point and click operation.

There are heaps of cool utilites you get access too.

If you are working on a remote computer, a command line might be the only way you can interact with it.

For Windows: http://www.cygwin.com/
Mac & Linux: terminal

cut

An example of a shell command

```
cut -d, -f13,35,37,75 data/ss12por.csv > data/ss12por-cut.csv
```

Breaking it down

`cut`	command name
`-d,`	option d with argument **`,`**
`-f13,35,37,75`	option f with arguments 13, 35, 37 and 75
`data/ss12por.csv`	the file to cut
`> data/ss12por-cut.csv`	take the output from the command and feed it out to a new file (> is called a redirect)

File paths

A *file path* is the location of a file or directory:

```
/Users/wickhamc/Documents/BigData/Project1/data/ss12por.csv
```

They can be specified absolute to the root directory (as above) or relative to where you are currently. I.e. if I'm in `/Users/wickhamc/Documents/BigData/Project1/` then:

```
/data/ss12por.csv
```

would refer to the same file as above.

`.` means the directory I'm in.
`..` means the directory above the one I'm in.
`~/` means my home directory (`/Users/wickhamc/` for me)

Hit `tab` and the terminal will try to complete what you have written so far.

General commands

Task	command
Where am I?	`pwd`
Change directory	`cd`
Make a directory	`mkdir`
Move a file	`mv`
Copy a file	`cp`
Delete a file	`rm`
Help on a command	`man`

Useful for data

Task	command
Look at a file	`less`, `more`, `head`, `tail`
remove sections from each line of files	`cut`
print lines matching a pattern	`grep`
pattern-directed scanning and processing language	`awk`
filtering and transforming text	`sed`

Find a cheat sheet you like and do a tutorial:

A Command Line Primer for Beginners

Basic Unix Shell Commands for the Data Scientist