Lecture 28 - common data manipulation

Videos

https://youtu.be/v8Z668Y-EQM - Lect-28-2150-pt1-data-manipulation.mp4
https://youtu.be/6GlAQE0B9XQ - Lect-28-2150-pt2-more-data-manip.mp4
https://youtu.be/_XHCsatomqM - Lect-28-2150-pt3-log-search.mp4
https://youtu.be/mndlF4WSGjw - Lect-28-2150-pt4.mp4

From Amazon S3 - for download (same as youtube videos)

http://uw-s20-2015.s3.amazonaws.com/Lect-28-2150-pt1-data-manipulation.mp4
http://uw-s20-2015.s3.amazonaws.com/Lect-28-2150-pt2-more-data-manip.mp4
http://uw-s20-2015.s3.amazonaws.com/Lect-28-2150-pt3-log-search.mp4
http://uw-s20-2015.s3.amazonaws.com/Lect-28-2150-pt4.mp4

Quick Look into data

```
$ wc log-file.txt
```

Get number of lines, words, characters in a file.

```
$ wc -l Users.xml
```

Too big - to play with so let's just take a peek in it and get the first 1000 lines.

```
$ head Users.xml
$ tail Users.xml
```

Now for the first 1000 lines

```
$ head -1000 Users.xml > first1000.xml
```

Now give it a spin in `vi` .

Tools Used

1. awk : https://www.grymoire.com/Unix/Awk.html
2. sed : https://www.gnu.org/software/sed/manual/sed.html
3. R : https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
4. gnuplot : http://physics.ucsc.edu/~medling/programming/gnuplot_tutorial_1/index.html
5. sort
6. uniq
7. head
8. tail

Scritps

gnu plot bar graph

```
#!/bin/bash

gnuplot -p -e 'set boxwidth 0.25 ; plot "-" using 1:xtic(2) with boxes'
```

R summary

```
#!/bin/bash

R --slave -e 'x <- scan(file="stdin", quiet=TRUE); summary(x)'
```

Sort/Unique on users

```
#!/bin/bash

grep '<row' $1 | sed -E 's/^.*DisplayName="//' | sed -E 's/".*$//' | get-length | tee ,
```

Line Length

```
#!/usr/bin/awk -f
{print length}
```