

Lab2 <Good front — Courier New>

```
wget -r -e robots=off --wait 1 -A updates.20180129.*.bz2 http://archive.routeviews.org/bgpdata/2018.01/UPDATES/
for f in ../UPDATES/updates.20180129.*; do ./bgpdump -M $f -O output.txt; done.
awk '/12.0.1.63/ && /W/' output.txt | sort -rnk5 | head -10
```

```
tar -xvzf CodeRedTraces.tar.gz
```

Terminal :

<https://www.datacamp.com/community/tutorials/shell-commands-data-scientist>

wc ; cat ; sed/grep ; head/tail ; uniq ; cut ; loop; variables;

**wc** -- word, line, character, and byte count

**-c** The number of bytes in each input file is written to the standard output. This will cancel out any prior usage of the **-m** option.

**-l** The number of lines in each input file is written to the standard output.

**-m** The number of characters in each input file is written to the standard output. If the current locale does not support multibyte characters, this is equivalent to the **-c** option. This will cancel out any prior usage of the **-c** option.

**-w** The number of words in each input file is written to the standard output.

cat file\_1.csv file\_2.csv > target\_file.csv will merge the content of both file\_1.csv and file\_2.csv into target\_file.csv, adding file\_2.csv at the end of file\_1.csv.

```
head -n 120 adult.csv | tail -n 20 > adult_sample.csv
```

```
sed "s/<string to replace>/<string to replace it with>/g" <source_file> >
```

```
<target_file>.  
sed "s/, ?,/,/g" adult.csv > adult.csv
```

Save with > :

```
ls -l | head
```

To save this information into a file in the current directory named mylist, you could enter:

```
ls -l | head > mylist
```

The -i option causes a case-insensitive search.

The -w option matches only whole words.

The -l option lists only the files in which matches were found, but not the matching lines.

The -r (recursive) option searches files in the current working directory and all subdirectories below it.

The -n option lists the matching lines, together with line numbers.

- `uniq -c`: which adds the repetition count to each line;
- `uniq -d`: which only outputs duplicate lines; And
- `uniq -u`: which only outputs unique lines.

```
$ sort adult.csv | uniq -c | sort -r | head -n 3  
3 25, Private, 195994, 1st-4th, 2, Never-married, ...  
2 90, Private, 52386, Some-college, 10, Never-married, ...  
2 49, Self-emp-not-inc, 43479, Some-college, 10, Married-civ-spouse, ...
```

The great thing about CSV files and shell commands is that you can also work at the column level by using `cut` to select a particular column. `cut` takes two main flags: `-d` to specify the column delimiter and `-f` to specify the columns you want to work on.

```
$ cut -d "," -f 2 adult.csv | sort | uniq -c  
1837  
960 Federal-gov  
2093 Local-gov  
7 Never-worked  
22696 Private  
1116 Self-emp-inc  
2541 Self-emp-not-inc  
1298 State-gov
```

14 Without-pay  
1 workclass

Filtering rows and columns :

**Filtering rows is easy, for example with AWK:**

```
cat largefile | awk 'NR >= 10000 && NR <= 100000 { print }'
```

**Filtering columns is easier with CUT:**

```
cat largefile | cut -d '\t' -f 10000-100000
```

**As Rahul Dravid mentioned, cat is not a must here, and as Zsolt Botykai added you can improve performance using:**

```
awk 'NR > 100000 { exit } NR >= 10000 && NR <= 100000' largefile
```

```
cut -d '\t' -f 10000-100000 largefile
```

```
root@b317c63c8ac5:/home/lab2/CRED# tcpdump -n -r CRed.07-19-01.dump |  
awk '/192.168.1.105/' > test.txt  
root@b317c63c8ac5:/home/lab2/CRED# cut -d ">" -f 1 test.txt > test1.txt  
root@b317c63c8ac5:/home/lab2/CRED# cat test1.txt | awk '/192.168.1.105/' | sort  
| uniq -c | wc -l  
2711
```

```
root@b317c63c8ac5:/home/lab2/CRED# cat test1.txt | awk '/192.168.1.105/' | sort  
| uniq -c > test  
root@b317c63c8ac5:/home/lab2/CRED# cut -d "." -f 1 test > test2.txt  
root@b317c63c8ac5:/home/lab2/CRED# cat test2.txt | sort | uniq -c | wc -l  
38  
root@b317c63c8ac5:/home/lab2/CRED# cat test2.txt | sort | uniq -c
```

```
root@b317c63c8ac5:/home/lab2/CRED# cat data1 | awk '/198.137.240.91/'  
>data1_2  
root@b317c63c8ac5:/home/lab2/CRED# cat data1_2 | sort | uniq -u | wc -l  
300
```

```
docker cp b317c63c8ac5:/home/lab2 /Users/zyx/PycharmProjects/yuxuanz6/lab2
```

