Unix Tools
Courant Institute of Mathematical Sciences
Homework assignment 2 – Solution
March 11, 2007

1. Download the data using the command `curl`:

   ```
   $ curl "http://www.usatoday.com/news/politicselections/vote2004/\
   PresidentialByCounty.aspx?oi=P&rti=G&tf=l&sp=FL" > \
   fl.county.2004.html
   ```

   How would the shell interpret this if you did not use quotes around the URL?

   Would produce three back-grounded jobs:

   ```
   [2] 50670
   [3] 50671
   [4] 50672
   [3]   Done                    rti=G
   [4]+  Done                    tf=l
   $ <html><head><title>Object moved</title></head><body>
   <h2>Object moved to <a href='/news/politicselections/\
   nation/primariescaucus/results.aspx'>here</a>.</h2>
   </body></html>

   [2]+  Done                    curl http://www.usatoday.com/news/\
   politicselections/vote2004/PresidentialByCounty.aspx?oi=P
   ```

2. Use a pipe of `awk` and `sed` with no more than one instance of each to extract county level data from the HTML file and write it to file `fl.county.2004`. The result should have six columns separated with the separator `*`. Why is space not a good separator here?

   ```
   $ gawk 'BEGIN{FS="<td"} NF == 7 && !/Bush/' < fl.county.2004.html |\
   sed -e 's/<[^>]*>/*/g' -e 's/,//g' -e 's/\*\**/\*/g' \
   -e 's/^[^*]*\*//' -e 's/\*[^*]*$//' > fl.county.2004
   ```

   Choice of the delimiter: two counties with a space in them,

```
Santa Rosa 43 43 51988 14637 295
St. Lucie 75 75 47575 51816 436
```

3. Use `awk` to determine how many precincts are there in FL and how many reporting.

```
$ cat fl.county.2004  | \
gawk 'BEGIN{FS="*"}{sum+=$2; sum_rep+=$3}END{print sum, sum_rep}'
7241 7241
```

4. Use `awk` to find how many votes were registered for Bush, Kerry, and Nader.

```
$ cat fl.county.2004  | \
gawk 'BEGIN{FS="*"}{sum_bush+=$4;sum_kerry+=$5;sum_nader+=$6}\
END{print sum_bush, sum_kerry,sum_nader}'
3955656 3574509 32890
```

5. Use `awk` to find out how many counties did Bush win.

```
cat fl.county.2004  | gawk 'BEGIN{FS="*"}$4 > $5 && $4 > $6'\
| wc -l
      56
```

6. Use `awk` and `sort` to determine who won the largest county.

```
$ gawk 'BEGIN{OFS=FS="*"} \
{print $1, $4, $5, $6, $4+$5+$6}' < fl.county.2004 | \
sort -t\* -nrk5 | gawk 'BEGIN{FS="*"} \
NR == 1 {if ($2 > $3 && $2 > $4) {print "Bush"}\
else if($3 > $4) {print "Kerry"} else print "Nader"}'
Kerry
```

7. Use `awk` and `sort` to determine how far down in the list of counties sorted in decreasing order of total number of votes we need to go to find a county Bush won.

```
$ gawk 'BEGIN{OFS=FS="*"}{ print $1, $4, $5, $6, $4+$5+$6 }'\
 < fl.county.2004 | sort -t\* -nrk5 | \
gawk 'BEGIN{FS="*"}($2 > $3 && $2 > $4){print NR; exit}'
4
```

8. Download the Florida 2000 Presidential elections data set from

   `http://www.stat.unc.edu/faculty/rs/source/Data/fldat1.txt`

   The row starting with `50` gives the results for the Palm Beach County.

   ```
   curl -O "http://www.stat.unc.edu/faculty/rs/source/Data/fldat1.txt"
   ```

9. Use `awk` to print the count and the county number of the county where Buchanan obtained the largest number of votes.

   ```
   $ gawk 'NR > 1 && ($18 > max) {max=$18; max_county=$1}\
   END{ print max, max_county}' < fldat1.txt
   3407 50
   ```

10. Use `awk` to determine the average number of votes Buchanan obtained in Florida.

    ```
    $ 260.672
    ```

11. Use `awk` to determine the standard deviation $\sigma$ of Buchanan's votes in Florida, that is the square-root of the average of $(|\text{votes}|-|\text{average\_votes}|)^2$.

    ```
    $ gawk 'NR > 1 {square+=($18 - 260.672)^2 }\
    END{ print sqrt(square/(NR -1)) }' < fldat1.txt
    446.554
    ```

12. Use `awk` to find out the number of standard deviations separating the number of votes Buchanan obtained in Palm Beach from the average number of votes he obtained in Florida, in absolute value.

    ```
    $ gawk '$1==50 { x= ($18 - 260.672)/446.554;\
    print (x>0?x:-x) }' < fldat1.txt
    7.0458
    ```

13. Use `awk` and `sort` to print that number in decreasing order for all counties.

```
$ gawk 'NR > 1{ x= ($18 - 260.672)/446.554; print $1, \
(x>0?x:-x) }' < fldat1.txt | sort -nrk2 | head
50 7.0458
52 1.68474
28 1.31301
6 1.18088
15 0.876329
51 0.6927
5 0.6927
41 0.677025
43 0.670306
53 0.607604
```

14. Use `awk` determine by how many votes $V$ Bush won in Florida. Give the ratio $V/\sigma$.

```
$ gawk '{bush+=$12; gore+=$13 }END{V=bush - gore; \
print V, V/446.554 }' < fldat1.txt
961 2.15204
```