Intermediate Linux Exercises

Get the file `linux.intermediate.exercises.zip` from one of:
1) `https://its.unc.edu/rc-services/research-computing-presentations` or
2) on longleaf.unc.edu in `/pine/scr/m/a//markreed/linux` or
3) using wget

```
wget https://its.unc.edu/files/2018/07/linux.intermediate.exercises.zip
```

Unzip it in your directory space as follows:

```
unzip linux.intermediate.exercises.zip
```

Do the following exercises:

1. Grep and regular expressions exercise using the the data file `Decl.Ind.txt`, which is the text of the Declaration of Independence.
   a) Print all lines with a number in them.
   b) Print the line that contains the famous phrase concerning life, liberty and the pursuit of happiness.
   c) To put this in context, print the two lines before it along with the line.
   d) Print just the names of the states. Note that these lines start with a capital letter and have a ":" after the name. Hint: you can pipe two `grep` commands together. How many lines are there?
2. Simple `gawk` example. Use `gawk` to print only the lines in the file `Decl.Ind.txt` with two columns. How can you print just the lines that are the names of the signers?
3. Simple `gawk` example. Use the data file `nba.Mar.2015.stats`.
   a) Print all players on the list who are averaging 20 or more points per game (PTS). You should find 14 (you can use `wc` to count them).
   b) Print all players on the list who are averaging 20 or more poinits per game and also averaging double figures in rebounds (REB). You should find three.
   c) Print out the stat line for all players whose first name is Kyle.
4. Sorting exercise, again use `nba.Mar.2015.stats`.
   a) Sort alphabetically by last name.
   b) Sort numerically by most rebounds from highest to lowest.
   c) Show only the top 10 in scoring (points).
   d) Bonus: print how many times each first name appears in the list. Hint use gawk piped into sort piped into another command to count (and perhaps a final sort to order them as you like). Which name appears the most?
5. Now for a more advanced `gawk` exercise, let's create the `gawksumcol` command. This should be a simple shell script that invokes `gawk`. I use this command frequently in

all kinds of contexts. The idea is to sum an arbitray column of numbers. You should pass in two arguments. First the input to use (either a filename or stdin (-)) and two, the number of the column to sum. Variations on this would be to compute an average, find a maximum or minimum, etc. You may find it simpler to create a script which calls `gawk` and passes it a program file with the `-f` flag.

6. Use the input file `AC091491.tbl` which is a long DNA sequence. First find the length of the sequence. Next print the percentage of each base, namely Adenine, Cytosine, Thiamine or Guanine (a, c, t, g) in the sequence. To make the ouput pretty, print the letter abbreviation of the base and the number of occurences as a percentage. Hint: you will need to use the "`fold -1`" command in your pipleline. You can do this second step all in one command line and your output should look like this:

   ```
   a 32.3073 %
   c 20.0802 %
   g 18.9257 %
   t 28.6868 %
   ```

   This is adapted from http://genome.crg.es/courses/genefinding/P6/index.html by Roderic Guigò.

7. See the first `grep` exercise. Now count the number of original colonies (i.e. states) in the Declaration of Independence by piping the output of the `grep` commands through a series of filters. How many original colonies were there? US natives should know that there were 13 original colonies. :)

8. Lets do a little analysis of the signers of the Declaration of Independence. Using the command line you constructed in the `gawk` example to print names, can you construct a command line to strip out the names and print a sorted list of each first name and how many times it appears? Which names are the most common? As a check now pipe your answer into the `gawksumcol` command you built previously (use - as the input file). Are there 56 signers? Why not? We could change `gawk` to print lines with either two or three columns to get most of the signers but would still be one short (damn you Charles Carroll! :).

Happy Linux-ing from the UNC Resarch Computing Team!