GawkExercises

Sofia Kampaki

December 21, 2018

R Markdown

A.

Report two sites where GAWK manual-like information is available

1) http://web.mit.edu/gnu/doc/html/gawk_toc.html from MIT

2) https://www.gnu.org/software/gawk/manual/gawk.html

B

Some species have gaps in their classification, i.e. some higher taxonomic groups (like the order or the class) they belong to are missing. In /home/pafilis/gawk/classification.cs

Can you spot write the command that will print the names of such species using these with a. a grep search? b. a gawk script?

a)

```
grep  ',,' classification.csv | cut -d, -f 1
```

b)

```
gawk -F","  'match($0,",,") {NR>1;print $1}' classification.csv  #ONE LINER GAWK
```

GAWK SCRIPT

```
#!/usr/bin/gawk/ -f

BEGIN{
        FS=",";
        IGNORECASE=1;


}

{ARGIND==1}{
 if(FNR==1){next;}
 match($0,",,")
 print $1;

}
```

Regular expressions in gawk have not been presented. Can you describe how a regular expression applied in gawk can answer the above? (in your reply include both a gawk-one-liner and an alternative version as a gawk script. In the gawk script SET the FIELD SEPARATOR (FS) in the BEGIN clause)

With regular exrpessions we could write in the command to find two or more commas at the file,and if it finds it, it should print only the first column with the corresponding species.That can be achieved with gawk or grep.

C

Based on the /home/pafilis/gawk/homework/classification_10_entries.csv, create a:

"species X family" 2D matrix (see next two slides) (classification_10_entries.csv contains only the top 10 entries of classification.csv – for development and educational purposes). Please report the final output and the script used

SCRIPT:

```
#!/usr/bin/gawk/ -f

BEGIN{
        FS=",";
        IGNORECASE=1;
}



{ARGIND==1}{ #classification_10.tsv
 if(FNR==1){next;}
 sp_family_presence[$2]=$4;          #first associative array to match species with families(families as
 families[$4]="1";                   #second associative array for the families only(families as keys)
}


END{
header_line="species_name";         #creating the header
for (family in families){
   header_line=header_line","family;
}
print header_line;
n=asorti(sp_family_presence,sorted);  #sort the array in order to have the output in alphabetical order
 for(species = 1; species<=n; species++){
        output_line="";                     #creating the output line(one for each species)
        output_line=sorted[species];
        for(family in families){
                if(sp_family_presence[sorted[species]] == family){
                        output_line=output_line",1";
                }
                else {
                        output_line=output_line",0";
                }
        }
```

```
        print output_line;
 }
```

The output:

```
species_name,Echinorhynchidae,Arhythmacanthidae,Polymorphidae
Acanthocephaloides distinctus,0,1,0
Acanthocephaloides geneticus,0,1,0
Acanthocephaloides incrassatus,0,1,0
Acanthocephaloides propinquus,0,1,0
Acanthocephalus anguillae,1,0,0
Acanthocephalus clavula,1,0,0
Acanthocephalus lucii,1,0,0
Andracantha phalacrocoracis,0,0,1
Andracantha tunitae,0,0,1
```