

Tabular Data File Processing

Read
Select
Merge (join)
Compute



Homework

- A. Report two sites where GAWK manual-like information is available
- B. Some species have gaps in their classification (i.e. some higher taxa (taxonomic group) they belong to information is). In `/home/pafilis/gawk/classification.csv`: Can you spot these with a. a **grep** search? b. a **gawk** script?

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)

- C. Based on the `/home/pafilis/gawk/homework/classification_10_entries.tsv`, create a: “species X family” 2D matrix (see next two slides) (`classification_10_entries.tsv` contains only the top 10 entries of `classification.csv` -- for development and educational purposes). Please report the final output and the script used

Tip: You may have to read the classification file twice and Store family names in an array initiated in BEGIN so could print the in the header in the output

Tip: consider “composite keys” in associative arrays

Tip: Gawk built-in parameters can help you handle the header lines (first line in each file)

NB: unpublished material, please do not distribute further

Homework

Input file: `/home/pafilis/gawk/homework/classification_10_entries.tsv` (NB: not `/home/pafilis/gawk/classification*`)

row_names, species, genus, family, order, class, phylum

Acanthocephaloides distinctus,Acanthocephaloides distinctus,Acanthocephaloides,Arhythmacanthidae,Echinorhynchida,Palaeacanthocephala,Acanthocephal

Acanthocephaloides geneticus,Acanthocephaloides geneticus,Acanthocephaloides,Arhythmacanthidae,Echinorhynchida,Palaeacanthocephala,Acanthocephal

Acanthocephaloides incrassatus,Acanthocephaloides incrassatus,Acanthocephaloides,Arhythmacanthidae,Echinorhynchida,Palaeacanthocephala,Acanthoceph

Acanthocephaloides propinquus,Acanthocephaloides propinquus,Acanthocephaloides,Arhythmacanthidae,Echinorhynchida,Palaeacanthocephala,Acanthoceph

Acanthocephalus anguillae,Acanthocephalus anguillae,Acanthocephalus,Echinorhynchidae,Echinorhynchida,Palaeacanthocephala,Acanthocephala

Acanthocephalus clavula,Acanthocephalus clavula,Acanthocephalus,Echinorhynchidae,Echinorhynchida,Palaeacanthocephala,Acanthocephala

Acanthocephalus lucii,Acanthocephalus lucii,Acanthocephalus,Echinorhynchidae,Echinorhynchida,Palaeacanthocephala,Acanthocephala

Andracantha phalacrocoracis,Andracantha phalacrocoracis,Andracantha,Polymorphidae,Polymorphida,Palaeacanthocephala,Acanthocephala

Andracantha tunitae,Andracantha tunitae,Andracantha,Polymorphidae,Polymorphida,Palaeacanthocephala,Acanthocephala

Homework

Requested file: species X family 2D matrix example

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