Homework 2: Shell Scripting

Due: Friday, Feb 5, 2016 by 11:59pm.

You can use whatever shell/OS you like, but I will assume you have access to a modern GNU userland/flags, and your work will be graded accordingly. As such, you are strongly encouraged to use your Ubuntu VM or Newton.

Each of these exercises uses the file `iris.csv`, available on blackboard. You can reproduce this file by entering `write.csv(file="iris.csv", iris)` in R. You are only allowed to use the tools introduced in class (e.g., no perl, R, python, ...).

1. (10 pts) Show valid shell syntax using `cut` to remove the first column (the index column) and store the resulting file as `iris.csv`. (Hint: don't try to do it in place; remember the example from class).

2. (10 pts) Show valid `sed` syntax to convert the csv file into a tab-delimeted file and store the result as `iris.tsv`. Briefly explain the pros and cons to this approach (i.e, using `sed` as opposed to using a CSV parser).

3. (10 pts) Show how to downsample the file from problem 2 to a new file containing only the observations from the "setosa" species. Call the new file `setosa.tsv`. (Hint: `grep`). Don't try to keep the "header" line (your downsampling will not preserve it).

4. (10 pts) Sort the setosa data from problem 3 by sepal width (the second variable) and remove duplicate entries. Store this as `setosa.tsv`. How many records were removed? Show all of the commands you used.

5. (10 pts) Suppose you had done the steps in problems 2, 3, and 4 for the virginica species in a separate file called `virginica.tsv` (you don't need to actually do this), and that you wished to recombine the results into a single file `iris_new.tsv` containing the records for both species. Between the two, which command is most appropriate for this re-combining, `cat` or `join`? Show how you would do it. (Don't add a "header" line to the final file, just the data).