Week 9 Exercises

Physics 91SI Spring 2011 Handout 22

Alex Ji and Zahan Malkani

Today, you will be using a combination of regular expressions, bash scripts, and awk to process some simple data files. **DO NOT USE PYTHON TO DO ANY OF THIS.** The point is to get practice with the other things. Clone over the `exercise9` directory from our class `src` directory into your work folder. Put all of your work in a script called `processdata.sh`.

If you're really efficient, this might actually take you only 20-30 minutes. If you'd like to get more practice with shell scripts, you can go here: `http://www.freeos.com/guides/lsst/ch08.html` Otherwise, just take off and we'll see you next Tuesday.

Given data

In the `data` folder, you'll find two types of files that describe galaxy positions and velocities: `pos_xx` files and `url_xx` files.

The `pos_xx` files have 7 space-separated columns in them. The columns correspond to x, vx, y, vy, z, vz, and id number.

The `url_xx` files are lists of (fake) urls. They are all of the form: `http://www.galaxyzoo.org/<surveyname>/<telescope>/<id>.html`. (These telescopes, surveynames, and ids have no correlation with real life surveys, we generated them arbitrarily for this exercise.)

Combine and Sort Data Files

Your first task is to combine the several different data files into one aggregated data file named `galaxies.dat`. The final aggregated data file should have the following columns in this order, sorted by order of increasing id:

`id, telescope, surveyname, x, y, z, vx, vy, vz, v2`

You'll have to calculate v^2, which we suggest you do with awk.

Regarding pulling out `telescope` and `surveyname`: `grep` can match regular expressions, but it actually can't extract groups the way python `re` can. Google up another shell tool that will do what you want (there are at least two).

Separate Aggregated File into Separate Data Files

Now take this giant file and separate them into different files for each telescope, `<telescopename>.dat`. Instead of sorting by id, sort by v^2. Keep the columns in the same order as `galaxies.dat`.

Note that in order to do that, you'll need to identify the unique survey names.