Lab 4: File Processing

Goals

A. To learn how to read, write, update, copy, and delete from files (touch, vi, sed, awk, rm, cp)
B. To learn how to create, move, remove, and copy files and directories (touch, vi, mv, cp, rm, rmdir, mkdir)
C. To understand standard output, standard input, standard error
D. To understand the concept of redirection (<, >, <<, >>), piping, and streams
E. Converting streams to files, files to streams
F. using the find utility
G. combining files with cat and paste
H. cutting files to pieces
I. sorting files and removing duplicates with uniqTha
J. using join
K. writing a shell script
L. using sed and awk
M. trimming output with tail, head, and grep
N. to understand file structures for storing and retrieving data
O. To learn to filter stdout with grep, head, and tail
P. To use comm and diff to compare files

Directions

1. Enter your terminal bash shell and come to your home directory. Create a new directory called lab4. cd into this directory and type pwd to verify your new location.

2. Inside here you are going to download and unpack two datasets. One is a recipes dataset following a tree structure and another is a dating app dataset that is a variable length record file called a csv file which separates values by commas and usually encloses values in double quotes. You will be using the utilities curl, unzip, and git to retrieve and process the data.

curl https://storage.googleapis.com/recipe-box/recipes_raw.zip > recipes_raw

mkdir recipes

unzip recipes_raw.zip -d recipes

git clone https://github.com/rudeboybert/JSE_OkCupid.git

3. View the file recipes_raw_nosource_ar.json in the recipes directory and also view the file profiles.csv in the JSE_OkCupid directory. Notice that one file separates values by commas (though there is much whitespace in the dataset) and the other dataset encases data in a kind of tree using curly braces { }.

4. Using all of the gnu utilities you have learned about up to this point, try to learn the following pieces of information:

a. which word occurs more frequently in the OkCupid data set: love, money, confidence, tall, friendly, or sex
b. how many recipes vs how many dating profiles are in each dataset, which one has the most
c. according to the OkCupid dataset, do more males or more females identify as bisexual
d. what percentage of men identify their body type as “athletic”
e. how many OkCupid profiles want kids
f. how many recipes call for curry

5. Using the OkCupid profile, create a new file that only lists the gender, height, and preference for pets.