Open Source Tools Assignment 1

Due: Thursday, October 2, 2014 at 11:59pm

Overview

In this assignment, we'll use some of the UNIX tools we have seen so far to analyze some real world data. Specifically, we will look at New York City's records pertaining to restaurants. Data from the NYC OpenData website was extracted and put into the directory `/home/unixtool/data/restaurants`. Inside this directory, there are a few scripts and files that were used to fetch and extract the data (you are welcome to explore but the contents are not necessary for the assignment) as well as the data itself. We have extracted all the grades given to restaurants and put them under the `grades` subdirectory, and within this directory we have created a subdirectory for each borough, and finally within each borough subdirectory have created a subdirectory for each zip code. Within each zip code directory there is one file for each restaurant, named using NYC's id for the restaurant (for example, `41583757`). The contents of each file contains 9 columns: the business id (same as the file name), the business name, the borough number (1=manhattan, 2=bronx, 3=brooklyn, 4=queens, 5=staten island), the building number, the street name, the zipcode, an inspection date (in year-month-day format), the inspection score, and the grade. There is one line for each inspection in date order. Scores represent a number violations (0-13 results in an A, 14-27 is a B, and 28+ is a C). Some restaurants will have a score of Z which means "Grade Pending" where a future inspection will be done rather than accept a C grade.

For example, the file `~unixtool/data/restaurants/grades/manhattan/10012/41169084` contains the contents:

```
41169084,THINK COFFEE,1,248,MERCER STREET,10012,2012-03-22,2,A
41169084,THINK COFFEE,1,248,MERCER STREET,10012,2012-07-30,13,A
41169084,THINK COFFEE,1,248,MERCER STREET,10012,2013-07-15,11,A
41169084,THINK COFFEE,1,248,MERCER STREET,10012,2014-07-14,9,A
```

Which means that the business *Think Coffee* at 248 Mercer St was inspected four times and got an A each time with varying scores.

Before you begin

Before answering the questions, explore the directories and experiment with UNIX commands. In particular, familiarize yourself with the commands: **cd**, **cat**, **head**, **tail**, **cut**, **sort**, **uniq**, **tr**, **wc**, **find**. **Do not use any other commands in answering the questions in this assignment**.

Questions

Each of these questions should be expressible as a single command or pipeline of commands. You should not need multiple lines or semicolons for any question. Do not make assumptions that rely on the specific set of files that exist; keep your answers generally applicable (if new data were entered or existing data were changed, your answers should still be correct; do not make any assumptions about valid baby names!)

Part I

Using **find** (and possibly other commands), write commands to do the following from `/home/unixtool/data/restaurants`. Do not use the **ls** command or any .csv files.

1. Print the total number of businesses.
2. List all zip codes in the data set that end with 11.
3. Print how many business are not in zip code starting with 10.
4. Print the business ids that are writable by the group.
5. List all business *names* in Staten Island (if there is a business name in two locations, only list it once).
6. List the alphabetically last 10 business names across all boroughs except Manhattan.
7. Print the name of the borough with the most restaurants.

Part II

This part asks questions that should be solved using the file *all_grades.csv*, which is simply a concatenation of all the data files into a single file (and a faster alternative to using find). Again, do not use any commands other than the ones listed above.

8. Show name of restaurant with most violations.
9. Show distribution of grades given (each line should contain a count, white space, and a grade)
10. Show distribution of the most recent 10,000 grades given (and for fun compare to the earliest 10,000 grades given).
11. Show the names of restaurants that got two different grades over time.
12. Show number of inspections by year (count followed by whitespace followed by year).
13. Print the three most common restaurant names. For this question, unlike the others, be case *insensitive* for names.
14. Print how many restaurants there are that have a restaurant of the same name in another borough.
15. Print the zip code with the most restaurants.
16. Print the then most common restaurant words (words are a series of any characters including `&`, split by a space).

Turning in the assignment

Copy the file `/home/unixtool/data/assignment1` into your directory and fill in the answers in the appropriate places in this file. It is important that you start with this file, which will help the graders partially automate testing. Only submit the commands themselves, not the output of the commands. Here is an example:

```
## Assignment 1
## Danny Meyer N53943243

#############################################################
## Question 1 ###############################################
#############################################################

find . -name myfile -print
#############################################################
## Question 2 ###############################################
#############################################################

find . -name otherfile -print | wc
```

Make sure each command has been tested! You can use the following tool to verify that your homework is in a valid format:

```
/home/unixtool/bin/check_assignment1 assignment1
```

When you have finished, submit using the homework submission system.