CS2043 Homework 2

Due: Sunday, February 9, 2014 at 11:59 PM EST on `http://cms.csuglab.cornell.edu`

Important Note: Different Unix systems have different configurations and behaviors. To be able to grade in an uniform way, we are going to use the configuration of the CSUG Lab machines as our standard environment. That means that you are free to use whatever platform you are comfortable with to complete your assignment, but you are responsible for making sure that your scripts execute as intended in one of the CSUG lab machines. You can `ssh` into one of these machines or you can install a clone of a CSUG Lab machine on your computer (it is easy and quick). Instructions are posted on the course website.

General Instrictions:

- You can form groups of two students to complete this assignment. Please form a group on CMS and submit only one solution per group.
- You must complete this assignment using only the Unix tools that were discussed in class between (and including) Lectures 2 and 6. That means that you are not allowed to use loops either in bash or in any other programming language.
- Make sure that you strictly follow the specifications for each exercise. As this is a scripting class, we are going to use scripts to grade your assignment automatically. Therefore, if you fail to comply with the specifications, you will get an automatic zero from our scripts.
- For each problem, you will write a script (not the output thereof), and save it to a file named with the problem label, e.g., **problem1.sh**. Assume that the input is in the same directory as the script. Remember that a bash script is a text file whose first line contains:

```
#!  /bin/bash
```

- Once you complete the assignment, make a compressed tarball using gzip named `submission.tgz`, containing all the scripts you have produced. Submit your compressed tarball to CMS. **Important**: your tarball should contain no directories.
- **Start this assignment early!** It is longer and more challenging than the previous one! Also, make sure that you test your access to a CSUG Lab machine and to CMS early. Requests to the IT department to fix access issues take at least 48 hours to start being considered.
- **Warning!** There will be absolutely no acceptance of late submissions.

Ithaca Foodies

We observed 100 foodies over a period of 2 years. We recorded 500 parties they formed to go eat in a restautant in Ithaca, NY. Download and decompress the file `http://www.cs.cornell.edu/courses/cs2043/2014sp/restaurants.txt.gz`, which lists the events, one per line, where each line lists the names in the party and the associated restaurant. Each line has the following format:

```
name,name,...,name;restaurant
```

Each party size varies from 1 to 9.

- **problem1.sh** : Write a script that computes a list of the top 10 foodies in town (the ones who went to restaurants the most during the observation period). Each foodie should be printed in its own line. The output should have the following format for each line:

  ```
  total person
  ```

  where `total` is the number of times the foodie went to a restaurant.

- **problem2.sh** Write a script to compute the list of the most visited restaurants, one restaurant per line, where each line should follow the following format:

  ```
  parties restaurant
  ```

  where `parties` is the total number of parties that went to the corresponding restaurant.

- **problem3.sh** Write a script to split each event into a different file, where each file should contain a list of people that joined that party, with their names separated by space.

- **problem4.sh** Using the files you produced in the preceding problem, write a script to count how many people there were in each party. Each line of the output should consist of the number of people that were in each party. The format of the output is:

  ```
  count
  count
  ...
  count
  ```

  where `count` is the number of people who were in each party.

Book play

The plain text version of the book Frankenstein is available at
`http://www.cs.cornell.edu/courses/cs2043/2014sp/frankenstein.txt`

- **problem5.sh**: Write a script to extract the 10 most used words in Letter 3 of the text Frankenstein. Words should contain no punctuation marks and the different forms of words with capital and lower case letters should be considered the same word, e.g., "The" and "the" are the same words, and "tree!" and "man," are not words, but "tree" and "man" are. (Hint: Letter 3 is between lines 255 and 298 of the file frankenstein.txt)
