CS2043 Assignment 2

Due: Saturday February 7th 2015 at 11:59 PM on `http://cms.csuglab.cornell.edu`.

You can form groups of two students to complete this assignment. Please form a group on CMS and submit only one solution per group.

Note: This assignment (and future ones) require you to have access to a Unix-like (Linux, Mac OS X, etc) machine. If you do not have such an operating system installed on your local machine, make sure to get a CSUGLab account (by contacting the course staff). Different systems have slightly different configurations. The main environment in this class will be GNU/Linux.

You must complete this assignment using only the Unix tools that were discussed in class between the first lecture and Lecture 7. That means that you are not allowed to use loops either in bash or in any other programming language.

For each problem, you will write a script (not the output thereof), and save it to a file named with the specified problem label, e.g., **problem1name.sh**. Assume that the input is in the same directory as the script. Remember that a bash script is a text file with: `#! /bin/bash` as the first line.

Start this assignment early. It is somewhat longer and more challenging than the previous one!

Tweets

1. Copy the file:
   `http://www.cs.cornell.edu/courses/cs2043/2015sp/assignments/tweets.tar.gz`
   and extract it into your home folder. This is a compressed tarball of a folder with 2000 text tweets.

2. Write a script **wordfreq.sh** to compute the number of words per tweet and output:

   (a) in a file called `min_words.txt` the minimum # of words in any tweet

   (b) in a file called `max_words.txt` the maximum # of words in any tweet

   (c) in a file called `avg_words.txt` the average # of words per tweet
   Round the average down to the nearest integer.

3. Write a script **charfreq.sh** to compute the number of characters per tweet and output:
   (a) in a file called `min_chars.txt` the minimum # of chars. in any tweet
   (b) in a file called `max_chars.txt` the maximum # of chars. in any tweet
   (c) in a file called `avg_chars.txt` the average # of chars. in any tweet
   Round the average down to the nearest integer.
4. Write a script **commword.sh** to compute the most common words among all tweets and output in a file called `most_common.txt` the top 10 most common words sorted by popularity (with the most common first i.e. descending order)

Note that the contents of your output files from the above scripts should only be numbers.

Reading Frankenstein

1. Download the plain text version of the book Frankenstein from:
   `http://www.cs.cornell.edu/courses/cs2043/2015sp/assignments/frankenstein.txt`
2. Write a script called **frankenread.sh** to extract the 10 most used words in Letter 3 of the Frankenstein text. 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 the trimmed version "tree" and "man" are.
   Hint: Letter 3 is between lines 255 and 298 of the file `frankenstein.txt`.

Write a short feedback document, **README.txt**, regarding this assignment and include it in your submission folder:

1. How did you find the assignment? Was it easy/hard? boring/interesting?
2. Is there something general or specific you would like to see covered later in the course?
3. Any other questions or concerns you may have

Once you are done, create a compressed tarball (compressed using bzip2) of the folder containing your answers and submit your tar archive using CMS. Please include all the scripts and output files that you used to complete all of the above tasks. Remember that your best friends, if you get stuck, are the `man` tool, piazza and your favorite search engine.