

CS2043 Homework 3

Due: Sunday February 12th 2012 at 11:59 PM on <http://cms.csuglab.cornell.edu>.

General 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 CSUG Lab account. Different systems have slightly different configurations. The main environment in this class is GNU/Linux.

Assignment Notes:

- You must complete the assignment using GNU/Linux tools that were discussed in class.
 - This assignment is composed of 3 parts.
 - For the first 2 parts, treat upper and lower case characters the same and ignore punctuation marks.
 - **Note:** Words are not just lists of alphabet letters. You should also match words that contain an apostrophe like: *don't*, *Frank's*, *can't* ...etc. But not arbitrary instances of single quotes as in *'this'*.
 - Complete the assignment by writing the commands you used in CMS.
 - * plus a simple feedback on the assignment
 - **Start this assignment early! It is longer/harder than the previous ones!**
-

Text play

- The plain text version of Frankenstein is available at <http://www.cs.cornell.edu/courses/cs2043/2012sp/frankenstein.txt>
 - A list of English prepositions was taken from Wikipedia and made available at <http://www.cs.cornell.edu/courses/cs2043/2012sp/prepositions.txt>
 - **Problem#1:-** Get the 100 most used words in the text of Frankenstein.
 - **Problem#2:-** Get the 100 most words in the text of Frankenstein that are not prepositions (use the given list).
-

HTML scraping

- An HTML page of a recent New York Times article about the Super Bowl is available at <http://www.cs.cornell.edu/courses/cs2043/2012sp/superbowl.html> :
 - For the purposes of this assignment, HTML text is anything that is not a tag. Meaning, `<someTag>text to scrape</someTag>`. You do not need to worry about JavaScript or other things that you might catch. We will consider all of that to be text for the purposes of this assignment. One easy way to identify tags here is that they begin with a `<` and contain characters that are **not** `>`.
 - **Problem#3:-** Parse out the text (from HTML) of the provided NY Times article and get the 100 most used words in the text.
 - **Problem#4:-** Get the 5 words that appear before and after any mention of the following words in text: “giants”, “patriots”.
-

Data processing

- This part is mainly to help give you a taste of `awk/gawk`.
 - A plain text, comma separated value, file containing an activities log for some person (*not me .. I am way too lazy for that*) is available at http://www.cs.cornell.edu/courses/cs2043/2012sp/activity_log.csv
 - The log has 3 columns: date, activity title, and time. Values from different columns are separated by commas.
 - There are three types of activities: *work*, *run*, and *farmers market*.
 - Each activity entry marks either the starting time of an activity or the end time. The notation used is: `start <activity>` and `end <activity>`.
 - **Problem#5:-** Calculate the total number of hours spent on each activity during the period in the log.
-

Remember your two best friends are the `man` tool and your favorite search engine. If you need to ask questions, use the class discussion board on Piazza.