CS2042 Homework 2

Due: Thursday September 23rd 2010 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 will be GNU/Linux.

Assignment Notes:

- You must complete the assignment using GNU/Linux tools that were discussed in class.
- The plain text version of Moby Dick is available at `http://www.cs.cornell.edu/courses/cs2042/2010fa/moby.txt`
- A list of English prepositions was taken from Wikipedia and made available at `http://www.cs.cornell.edu/courses/cs2042/2010fa/prepositions.txt`
- An HTML page of the transcripts of a recent Glenn Beck Show (as of the time of writing this assignment) is available at `http://www.cs.cornell.edu/courses/cs2042/2010fa/beck.html`
- Treat upper and lower case characters the same, and ignore punctuation marks.
- Complete the assignment by writing the commands you used in CMS.
  - plus a simple feedback on the assignment

- **Problem#1:-** Get the 100 most used words in the text of Moby Dick.
- **Problem#2:-** Get the 100 most words in the text of Moby Dick that are not prepositions (use the given list).
- **Problem#3:-** Parse out the text (from HTML) of the provided Glenn Beck Show transcript and get the 100 most used words in the text.
- **Problem#4:-** Get the 25 words that appear before and after any mention of the following words in text: “democrats”, “obama”, “pelosi”.

Remember your two best friends are the `man` tool and your favorite search engine.