CS2044 Homework 1

Due: Sunday March 6th 2011 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.
- Complete the assignment by submitting your script on CMS.
  - plus a simple feedback on the assignment

In this assignment you will write a script to build a very rudimentary index of a corpus of Twitter updates. Download the list of the following package with a list of tweets and an expanding shell script:

```
wget http://www.cs.cornell.edu/courses/cs2044/2011sp/hw1.tar.gz
tar -xzf hw1.tar.gz
```

There are two files in this tarball:

- `tweets.txt` contains a list of 2000 tweets each on a separate line.
- `expand_tweets.sh` is a simple bash script to put each tweet in a separate file. It would be very useful for you to take a look at that script as it demonstrates the use of `while` loops in bash, and the `read` command.

Your objective in this assignment is to write a bash script, `find_by_keywords.sh`, that takes in a list of words and creates a directory for each word containing all the tweets that contain that word (ignore capitalization when looking for matches). The contents of the output directory should be multiple files with a single tweet per file. Additionally, if the first argument to your script is `-all` then the output should be a single directory containing the tweets that contain **all** the passed in arguments to your script.

In summary:

- The result of calling `./find_by_keywords.sh restaurant` should be a directory called `restaurant` containing all the files of tweets containing the word 'restaurant'.
- The result of calling `./find_by_keywords.sh restaurant ithaca` should be two directories, `restaurant` and `ithaca`, each containing all the files of tweets containing their respective keywords.
- The result of calling `./find_by_keywords.sh -all restaurant ithaca` should be a single directory, `restaurant_ithaca`, that contains all the files of tweets containing both of the words 'restaurant' and 'ithaca".

Finally, when you're looking for matches, make sure to only consider files that contain full word matches and not only substrings, there is a useful `grep` flag for that if you're interested.

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

Good Luck!