**Lab Problems 7**

**Problem 1**

Get a collection of tweets (new collection that you collect on your own).

Put the tweets in a Mongo DB or you can use them from a file.

Either write a short Python program or run a Python interpreter session as we did in class and produce a list with frequencies of the top 30 frequency tokens. Use lower-case on words and eliminate stopwords and words with all non-alphabetic characters.

**Problem 2**

Consider this example text, consisting of a tweet, an excerpt from an html page, and an extra sentence.

Tweet

text = “Celebrate #NationalPetDay with our puppy playlist: https://t.co/eBHHFPW0z7 https://t.co/uix5AY2FFQ”

Excerpt of HTML page

<a href="http://msande.stanford.edu"> Management Science and Engineering </a>

<p class="MsoNormal">

Address: Terman 311, Stanford CA 94305<br>

Email: ashishg@cs.stanford.edu<br>

Phone: (650)814-9999 [Cell], Fax: (650)723-9999<br>

Admin asst: Roz Morf, Terman 405, 650-723-9999, rozm@stanford.edu</p>

Extra Sentence

“The U.S.A. olympic teams have east-west training centers with up-to-date equipment.”

Write regular expressions using python to find the following items, where these can all be separate regular expressions:

1. match the URLs (there should be 3 results)

2. match the phone numbers (3 results)

3. match the email addresses (2 results)

4. match the (all) words with one or more internal hyphen (choice is either to have 2 results or to have many results for all the words, but including the 2 with internal hyphens)

5. match acronyms (there should be two results, with or without dots: (CA and U.S.A.)

You can try your regular expressions first in regexpal and then put them into python. But note that occasionally, the behavior of regular expressions in regexpal and python are a little different, possibly with the use of + and ?. (My guess is that regexpal gets the results a little differently than python.)