DATA 200 Spring 2020 Homework 6

Due March 27, 2020 at 11:59pm

Note: Please only use python standard libraries covered in class lectures (i.e. json, csv, urllib etc). Other libraries such as pandas, numpy are *not* allowed in your work.

1. Redo HW3-3 (word count problem) by utilizing **regular expressions** and the **Counter** class from the collections library. In this case, you will make it an interactive program which prompts the user for length of word and the n^{th} topmost common words and prints out the results as follows:

```
Enter length of word and # terms to return:  2 10

Top 10, length of word: 2
to 414
my 393
of 370
in 323
is 170
me 164
so 145
be 142
as 121
it 111
```

You must use regular expressions to extract the words and use Counter to tally in this exercise. The split() function is not allowed to parse the strings. You may get a few extra to's etc depending on whether you count the *to* in "*to*-day" or "*to*-morrow" in the text. Also, you shouldn't need to replace the special characters ?!,;: as you did in HW#3.

(10 points)

2. Write down regular expressions (one regex for each case) for the following:
   a) Extract all domain names in a string that contains email addresses. For example, your regex should return ["sjsu", "gmail"] for the string "An email was sent from student@**sjsu**.edu to inquiries@**gmail**.com".
   b) Extract all words that start with a vowel.
   c) Extract all words between 5 to 7 alphabets long.
   d) Extract all phone numbers that are in the following format:
      123-456-7890
      1234567890
      (123) 456-7890

(10 points)