L435/L555
In-class Exercise #9
Fall 2016

Using regular expressions in Python:

1) Download the book *Sister Carrie* from Project Gutenberg (http://www.gutenberg.org/ebooks/5267) – or some other book, as that's not really crucial. After reading in the text, use .findall() to figure out what these regular expressions do:

   a) 'th[aeiou]s'
   b) 'th([aeiou])s'
   c) '(th([aeiou])s) '
   d) '(\w+(th([aeiou])s)\w+)'
      * What does \w stand for?
   e) r'\b(\w+)\b \b(\1)\b' … Note:
      * the raw string r'' notation, to ensure \b and \1 are interpreted properly
      * \1 as a *backreference* to the previous parenthetical match
   f) r'(\bthe\b\s+\w+ing\s+(\w+))'
      * What does \s stand for?
   g) Expand (f) so that you have a wider (and more precise?) range of adjective types after 'the' (i.e., what other/different suffixes would help?). Include some other determiners, too. Essentially, think of your task as trying to deduce nouns in the (\w+) part of the RE.

2) Using .search(), print all the lines that have diphthongs, i.e. two consecutive vowels (including y).

3) Are there any lines that are all caps? How about individual words?

4) Replace all occurrences of 'Chicago' by 'The Windy City'.

5) Print all the lines that have exactly two occurrences of *the*. Do this first for adjacent *the*s and then for not necessarily adjacent occurrences.