33-beautiful-soup

April 29, 2016

1 BeautifulSoup

- great 'screen scraping' package
- tons of interesting data on webpages
- makes it easy to extract information from complex web pages and XML documents
- can figure out what to do by playing interactively
- doc

In [1]:
```
# another way to do hamlet

import urllib.request
import collections
import bs4

url='https://courseworks.columbia.edu/access/content/group/'
url+='COMSW3101_002_2015_3/data/hamlet.html'

def hamlet(url):
    page = urllib.request.urlopen(url)
    sp = bs4.BeautifulSoup(page, 'html5lib', from_encoding='utf-8')
    lam = lambda t : t.name == 'a' and ('name' in t.attrs ) and t['name'].startswith('speech')
    al = sp.findAll(lam)
    cd=collections.defaultdict(int)
    lcnt=0
    for a in al:
        name = a.string
        cd[name] += 1
    speeches = sum(cd.values())
    # no line count
    return([len(al), speeches, cd])

hamlet(url)
```

Out[1]:
```
[1150,
 1150,
 defaultdict(int,
             {'All': 4,
              'BERNARDO': 23,
              'CORNELIUS': 1,
              'Captain': 7,
              'Danes': 3,
              'FRANCISCO': 8,
              'First Ambassador': 1,
```

```
 'First Clown': 33,
 'First Player': 8,
 'First Priest': 2,
 'First Sailor': 2,
 'GUILDENSTERN': 33,
 'Gentleman': 3,
 'Ghost': 14,
 'HAMLET': 359,
 'HORATIO': 112,
 'KING CLAUDIUS': 102,
 'LAERTES': 62,
 'LORD POLONIUS': 86,
 'LUCIANUS': 1,
 'Lord': 3,
 'MARCELLUS': 36,
 'Messenger': 2,
 'OPHELIA': 58,
 'OSRIC': 25,
 'PRINCE FORTINBRAS': 6,
 'Player King': 4,
 'Player Queen': 5,
 'Prologue': 1,
 'QUEEN GERTRUDE': 69,
 'REYNALDO': 13,
 'ROSENCRANTZ': 49,
 'Second Clown': 12,
 'Servant': 1,
 'VOLTIMAND': 2})]
```

2 Want to find all the headlines on the front page of the New York Times

- html structure is quite complex
- would be very difficult to do with string.find() or regular expressions

In [2]:
```
# 'lxml' is a XML parser
# must tell soup what encoding to use

from bs4 import BeautifulSoup

nf2 = urllib.request.urlopen('http://nyt.com')
sp = BeautifulSoup(nf2, 'lxml', from_encoding='utf-8')
```

In [3]:
```
# headlines seem to be contained in 'h2' elements

sp.findAll('h2')[10:20]
```

Out[3]:
```
[<h2 class="story-heading"><a href="http://www.nytimes.com/2016/04/30/us/politics/indiana-republ
 <h2 class="story-heading"><a href="http://www.nytimes.com/2016/04/29/us/politics/out-of-office-
 <h2 class="story-heading"><i class="icon"></i><a href="http://www.nytimes.com/2016/04/30/us/pol
 <h2 class="story-heading"><i class="icon"></i><a href="http://www.nytimes.com/2016/04/30/us/pol
 <h2 class="story-heading"><i class="icon"></i><a href="http://www.nytimes.com/2016/04/30/world/
 <h2 class="story-heading"><a href="http://www.nytimes.com/2016/04/30/books/walt-whitman-promote
 <h2 class="story-heading"><a href="http://www.nytimes.com/2016/04/30/us/small-colleges-losing-m
```

```
 <h2 class="story-heading"><a href="http://www.nytimes.com/2016/04/29/us/koch-brothers-antonin-s
 <h2 class="story-heading"><a href="http://www.nytimes.com/interactive/2016/04/29/upshot/money-r
 <h2 class="story-heading"><a href="http://www.nytimes.com/2016/04/30/world/middleeast/surge-in-
```

In [4]:
```
# first 'h2' element

h2 = sp.h2
h2
```

Out[4]:
```
<h2 class="branding"><a href="http://www.nytimes.com/">
<svg aria-label="The New York Times" class="nyt-logo" height="64" role="img" width="379">
<image alt="The New York Times" border="0" height="64" src="https://a1.nyt.com/assets/homepage/2
</svg>
</a></h2>
```

In [5]:
```
# can pull 'a' element out of 'h2'
# this 'a' element is a picture

a=h2.find('a')
a
```

Out[5]:
```
<a href="http://www.nytimes.com/">
<svg aria-label="The New York Times" class="nyt-logo" height="64" role="img" width="379">
<image alt="The New York Times" border="0" height="64" src="https://a1.nyt.com/assets/homepage/2
</svg>
</a>
```

In [6]:
```
# try pulling the 'a' out of all 'h2' elements
# looks like we get mostly headlines

al=[h2.find('a') for h2 in sp.findAll("h2")]
al[:20]
```

Out[6]:
```
[<a href="http://www.nytimes.com/">
 <svg aria-label="The New York Times" class="nyt-logo" height="64" role="img" width="379">
 <image alt="The New York Times" border="0" height="64" src="https://a1.nyt.com/assets/homepage/
 </svg>
 </a>,
 None,
 None,
 None,
 None,
 None,
 <a href="http://www.nytimes.com/2016/04/30/world/asia/afghanistan-doctors-without-borders-hospi
 <a href="http://www.nytimes.com/interactive/2015/11/25/world/asia/errors-us-airstrike-afghan-ku
 <a href="http://www.nytimes.com/2016/04/29/us/politics/hillary-clinton-donald-trump-women.html"
 <a href="http://www.nytimes.com/2016/04/30/us/donald-trump-rally-protest-costa-mesa.html">Proto
 <a href="http://www.nytimes.com/2016/04/30/us/politics/indiana-republican-transgender-rights-ba
 <a href="http://www.nytimes.com/2016/04/29/us/politics/out-of-office-ex-speaker-john-boehner-gi
 <a href="http://www.nytimes.com/2016/04/30/us/politics/obama-puts-his-weight-behind-smart-gun-t
 <a href="http://www.nytimes.com/2016/04/30/us/politics/justices-leave-texas-voter-id-law-intact
 <a href="http://www.nytimes.com/2016/04/30/world/asia/north-korea-kim-dong-chul-sentence.html"?
 <a href="http://www.nytimes.com/2016/04/30/books/walt-whitman-promoted-a-paleo-diet-who-knew.ht
 <a href="http://www.nytimes.com/2016/04/30/us/small-colleges-losing-market-share-struggle-to-ke
 <a href="http://www.nytimes.com/2016/04/29/us/koch-brothers-antonin-scalia-george-mason-law-sch
 <a href="http://www.nytimes.com/interactive/2016/04/29/upshot/money-race-and-success-how-your-s
 <a href="http://www.nytimes.com/2016/04/30/world/middleeast/surge-in-palestinian-youths-in-pris
```

In [7]:
```
# pull out the 'a' link text

[a.contents for a in al if a != None][:30]
```

Out[7]:
```
[['\n',
  <svg aria-label="The New York Times" class="nyt-logo" height="64" role="img" width="379">
  <image alt="The New York Times" border="0" height="64" src="https://a1.nyt.com/assets/homepag
  </svg>,
  '\n'],
 ['Punishing 16, Pentagon Says Mistakes Led to Hospital Attack'],
 ['A Step-by-Step Look at the Errors Behind the Strike'],
 ['Trump and Clinton Gear Up for a Race Defined by Gender'],
 ['Protest Turns Violent at Trump Rally in Southern California'],
 ['Cruz Seizes on Transgender Issue in Attacks on Trump'],
 ['Boehner, Unbound and Speaking Freely'],
 ['Obama Puts Weight Behind Smart Gun Technology'],
 ['Justices Leave Texas Voter ID Law Intact'],
 ['North Korea Sentences American to 10 Years for Spying'],
 ['Buried in Microfilm, Whitman's Health Tips'],
 ['At Small Colleges, Harsh Lessons About Cash Flow'],
 ['University in Turmoil Over Scalia Tribute and Koch Role'],
 ['Money, Race and Success: How Your School District Compares'],
 ['Surge in Palestinian Youths in Prison Tests Israel'],
 ['Notes From Aleppo: Glimpses of War-Ravaged Syria'],
 ['Study Finds Fewer Homeless on Streets of New York'],
 ['The Perks of Monotasking, a.k.a. 'Paying Attention''],
 ['Review: In Key & Peele's 'Keanu,' Guns and Kittens'],
 ['The Day Prince's Guitar Wept the Loudest'],
 ['The Opinion Pages'],
 ['There's No Such Thing as a Free Rolex'],
 ['Editorial: The Racist Roots of a Way to Sell Homes'],
 ['Brooks: If Not Trump, What?'],
 ['Krugman: Wrath of the Conned'],
 ['Egan: Working-Class Fraud'],
 ['Op-Ed: What It's Like to Write Jokes for President Obama'],
 ['Taking Note: The Cruz Campaign and Execution'],
 ['Sunday Review'],
 ['The Upside to Overt Racism']]
```

In [8]:
```
# filter out images

[a.contents for a in al if a != None and len(a)==1][:30]
```

Out[8]:
```
[['Punishing 16, Pentagon Says Mistakes Led to Hospital Attack'],
 ['A Step-by-Step Look at the Errors Behind the Strike'],
 ['Trump and Clinton Gear Up for a Race Defined by Gender'],
 ['Protest Turns Violent at Trump Rally in Southern California'],
 ['Cruz Seizes on Transgender Issue in Attacks on Trump'],
 ['Boehner, Unbound and Speaking Freely'],
 ['Obama Puts Weight Behind Smart Gun Technology'],
 ['Justices Leave Texas Voter ID Law Intact'],
 ['North Korea Sentences American to 10 Years for Spying'],
 ['Buried in Microfilm, Whitman's Health Tips'],
 ['At Small Colleges, Harsh Lessons About Cash Flow'],
 ['University in Turmoil Over Scalia Tribute and Koch Role'],
```

```
['Money, Race and Success: How Your School District Compares'],
['Surge in Palestinian Youths in Prison Tests Israel'],
['Notes From Aleppo: Glimpses of War-Ravaged Syria'],
['Study Finds Fewer Homeless on Streets of New York'],
['The Perks of Monotasking, a.k.a. 'Paying Attention''],
['Review: In Key & Peele's 'Keanu,' Guns and Kittens'],
['The Day Prince's Guitar Wept the Loudest'],
['The Opinion Pages'],
['There's No Such Thing as a Free Rolex'],
['Editorial: The Racist Roots of a Way to Sell Homes'],
['Brooks: If Not Trump, What?'],
['Krugman: Wrath of the Conned'],
['Egan: Working-Class Fraud'],
['Op-Ed: What It's Like to Write Jokes for President Obama'],
['Taking Note: The Cruz Campaign and Execution'],
['Sunday Review'],
['The Upside to Overt Racism'],
['Better Aging Through Practice, Practice, Practice']]
```

In [ ]: