32-regular

April 29, 2016

1 Regular Expressions

- fairly standard implementation
- compile and match objects

```
In [1]: import re

        pat = 'x[0-9]+y'
        s = 'zxcvx9784843845ysdfx234yzxcv234'

In [2]: # compile the regular pattern for speed

        rec=re.compile(pat)
        rec

Out[2]: re.compile(r'x[0-9]+y', re.UNICODE)

In [3]: # find all substrings that match the pattern

        rec.findall(s)

Out[3]: ['x9784843845y', 'x234y']

In [4]: # replace the pattern with a string

        re.sub(pat, 'FOOBAR', s)

Out[4]: 'zxcvFOOBARsdfFOOBARzxcv234'

In [5]: # split on the pattern

        re.split(pat, s)

Out[5]: ['zxcv', 'sdf', 'zxcv234']

In [6]: # another way to analyse hamlet

        import urllib.request
        import collections

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

        def actors(url):
            # orginal pattern
```

```python
    pat = '<A NAME=speech[0-9]+><b>(.+)</b></a>'
    comp = re.compile(pat)
    ef = urllib.request.urlopen(url)
    cd = collections.defaultdict(int)
    lcnt=0
    for ba in ef:
        s = ba.decode('utf-8')
        lcnt += 1
        m = comp.match(s)
        if m != None:
            name = m.group(1)
            cd[name] += 1
    speeches = sum(cd.values())
    return([lcnt,speeches, cd])

actors(url)
```

```
Out[6]: [8881,
         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})]
```

```
In [ ]:
```