Advanced Programming: Week 5

Meta characters

- Recap of meta-characters:

  * - match zero or more

  + - match one or more

  [hbla] – character class

  ^ - start of string

  $ - end of string

  (subexpression)

  {4} – match 4 times

Application!

```
import urllib2
import re

# a function which gets the area of a district by extracting the value from
# a wikipedia page
def finddistrictarea(name):
    name = name.lower()
    districtwikipediapage = 'http://en.wikipedia.org/wiki/%s_District' %
(name)
    try:
        for line in urllib2.urlopen(districtwikipediapage):
            if ('<sup>2</sup>' in line) and ('sq mi' in line) :

                res=re.search('<td>([0-9,]*)',line)
                area_string = res.group(1);
                area_string = area_string.replace(',','');
                return int(area_string)
    except urllib2.HTTPError:
        return 'District not found'

# now call the above function to find some country areas
for district in ['Yumbe','Bukwo','Jinja','Gulu']:
    area = finddistrictarea(district)
    print "%15s: %5d" % (district,area)
```

Download from webpage or my laptop...

Application!

Can you alter the program to return the elevation of the district instead?

Download from webpage or my laptop...

More meta characters

- Other meta-characters:

  {3,5} - match 3, 4 or 5 times

  . - match any character

  [^ab] - don't match a or b

  ([abc]*)\1 - back references