Pattern matching and Web Scraping

It is used to describe a searh pattern,

when we want to extract any pattern from the given raw data, for eg: mobile No., email id, client id etc.

In [14]:
```
#Searching a pattern in string 'findall'
#findall returns list of matches
import re
Nameage = 'David is 25 and Smith is 30 /n Michael is 28 and Wayne is 35'
ages = re.findall('\d{1,2}',Nameage)
print (ages)
names = re.findall('[A-Z][a-z]*',Nameage)
print (names)
agedict ={}
x=0
for i in names:
    agedict[i]=ages[x]
    x+=1
print (agedict)
```

```
['25', '30', '28', '35']
['David', 'Smith', 'Michael', 'Wayne']
{'David': '25', 'Smith': '30', 'Michael': '28', 'Wayne': '35'}
```

In [7]:
```
#we can directly search for a string in a given string using 'search'
#search returns a single match
if re.search('match','I want to match the string'):
    print ('match captured')
```

```
match captured
```

In [9]:
```
# #to get the index range of the pattern match
str_ = 'This is a demo regex prog for a regex understanding'
for i in re.finditer('regex',str_):
    print (i)
    index = i.span()
    print (index)
```

```
<re.Match object; span=(15, 20), match='regex'>
(15, 20)
<re.Match object; span=(32, 37), match='regex'>
(32, 37)
```

In [15]:
```
#compile method which catches patterns and provide method to substitute
demo = 'Java html c++ ruby html'
object_ = re.compile('html') #matching objects with compile
sub_ = object_.sub('python',demo)
sub_
```

Out[15]:
```
'Java python c++ ruby python'
```

In [17]:
```
num = '123 1234 12345 123456 1234567 87654321'
print ('Matches :', len(re.findall(r'\d{5,7}' , num))) #use len to give the co
unt of the number of matches
```

```
Matches : 4
```

Web Scraping

Scrap useful data from web and store it in csv format or excel format.

In [58]:
```
#Zomato customer care India
import urllib.request
import re
url = 'http://www.talkingtrendo.com/zomato-customer-care-number-address-contac
t-number/'
# url = 'http://www.arrl.org/list-all-products'
response=urllib.request.urlopen(url)
html = response.read()
htmlstr=html.decode()
data=re.findall('\d{3} - \d{8}',htmlstr)
for i in data:
    print (i)
```

```
079 - 60601010
080 - 60601010
044 - 60601010
011 - 60601010
040 - 60601010
030 - 60601010
022 - 60601010
020 - 60601010
141 - 60601010
079 - 60601010
080 - 60601010
044 - 60601010
011 - 60601010
040 - 60601010
030 - 60601010
022 - 60601010
020 - 60601010
141 - 60601010
```

In [21]:
```
import re
x = 'my name is Michael and my age is 25 , Michael'
name = re.findall('[A-Z][a-z]*',x)
age = re.findall ('\d{1,2}',x)
nameage={}
j=0
for i in name:
    nameage[i]=age[j]
print (nameage)

if re.search('Abbas',x):
    print ('match captured')

obj = re.compile('25')
obj1 = obj.sub('26',x)
print (obj1)

for i in re.finditer('Michael',x):
    index = i.span()
    print (index)
```

```
{'Michael': '25'}
my name is Michael and my age is 26 , Michael
(11, 18)
(38, 45)
```

In [ ]:
```
```