Name : Annapoornima S

Roll NO.: 225229101

Lab.8 : Python Regular Expressions

Qunestion 1 : Using Email Collection file,mbox-short.txt,write a python program for the following queries

In [7]:
```
#Qns :1

import re
email=open("mbox_short.txt")
for line in email:
    line=line.rstrip()
    if re.search('From:',line):
        print(line)
```

```
From: stephen.marquard@uct.ac.za
From: louis@media.berkeley.edu
From: zqian@umich.edu
From: rjlowe@iupui.edu
From: zqian@umich.edu
From: rjlowe@iupui.edu
From: cwen@iupui.edu
From: cwen@iupui.edu
From: gsilver@umich.edu
From: gsilver@umich.edu
From: zqian@umich.edu
From: gsilver@umich.edu
From: wagnermr@iupui.edu
From: zqian@umich.edu
From: antranig@caret.cam.ac.uk
From: gopal.ramasammycook@gmail.com
From: david.horwitz@uct.ac.za
From: david.horwitz@uct.ac.za
From: david.horwitz@uct.ac.za
From: david.horwitz@uct.ac.za
From: stephen.marquard@uct.ac.za
From: louis@media.berkeley.edu
From: louis@media.berkeley.edu
From: ray@media.berkeley.edu
From: cwen@iupui.edu
From: cwen@iupui.edu
From: cwen@iupui.edu
```

In [8]:
```
#Qns :2

import re
mail=open("mbox_short.txt")
for line in mail:
    line=line.rstrip()
    if re.search('^F',line):
        print(line)
```

```
From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
From: stephen.marquard@uct.ac.za
From louis@media.berkeley.edu Fri Jan  4 18:10:48 2008
From: louis@media.berkeley.edu
From zqian@umich.edu Fri Jan  4 16:10:39 2008
From: zqian@umich.edu
From rjlowe@iupui.edu Fri Jan  4 15:46:24 2008
From: rjlowe@iupui.edu
From zqian@umich.edu Fri Jan  4 15:03:18 2008
From: zqian@umich.edu
Files Changed
From rjlowe@iupui.edu Fri Jan  4 14:50:18 2008
From: rjlowe@iupui.edu
From cwen@iupui.edu Fri Jan  4 11:37:30 2008
From: cwen@iupui.edu
From cwen@iupui.edu Fri Jan  4 11:35:08 2008
From: cwen@iupui.edu
From gsilver@umich.edu Fri Jan  4 11:12:37 2008
From: gsilver@umich.edu
From gsilver@umich.edu Fri Jan  4 11:11:52 2008
From: gsilver@umich.edu
From zqian@umich.edu Fri Jan  4 11:11:03 2008
From: zqian@umich.edu
From gsilver@umich.edu Fri Jan  4 11:10:22 2008
From: gsilver@umich.edu
From wagnermr@iupui.edu Fri Jan  4 10:38:42 2008
From: wagnermr@iupui.edu
From zqian@umich.edu Fri Jan  4 10:17:43 2008
From: zqian@umich.edu
From antranig@caret.cam.ac.uk Fri Jan  4 10:04:14 2008
From: antranig@caret.cam.ac.uk
From gopal.ramasammycook@gmail.com Fri Jan  4 09:05:31 2008
From: gopal.ramasammycook@gmail.com
From david.horwitz@uct.ac.za Fri Jan  4 07:02:32 2008
From: david.horwitz@uct.ac.za
From david.horwitz@uct.ac.za Fri Jan  4 06:08:27 2008
From: david.horwitz@uct.ac.za
From david.horwitz@uct.ac.za Fri Jan  4 04:49:08 2008
From: david.horwitz@uct.ac.za
From david.horwitz@uct.ac.za Fri Jan  4 04:33:44 2008
From: david.horwitz@uct.ac.za
From stephen.marquard@uct.ac.za Fri Jan  4 04:07:34 2008
From: stephen.marquard@uct.ac.za
From louis@media.berkeley.edu Thu Jan  3 19:51:21 2008
From: louis@media.berkeley.edu
From louis@media.berkeley.edu Thu Jan  3 17:18:23 2008
From: louis@media.berkeley.edu
```

```
From ray@media.berkeley.edu Thu Jan  3 17:07:00 2008
From: ray@media.berkeley.edu
From cwen@iupui.edu Thu Jan  3 16:34:40 2008
From: cwen@iupui.edu
From cwen@iupui.edu Thu Jan  3 16:29:07 2008
From: cwen@iupui.edu
From cwen@iupui.edu Thu Jan  3 16:23:48 2008
From: cwen@iupui.edu
```

In [9]:

```
#Qns :3

import re
mail=open("mbox_short.txt")
for line in mail:
    line=line.rstrip()
    if re.search('F..m:',line):
        print(line)
```

```
From: stephen.marquard@uct.ac.za
From: louis@media.berkeley.edu
From: zqian@umich.edu
From: rjlowe@iupui.edu
From: zqian@umich.edu
From: rjlowe@iupui.edu
From: cwen@iupui.edu
From: cwen@iupui.edu
From: gsilver@umich.edu
From: gsilver@umich.edu
From: zqian@umich.edu
From: gsilver@umich.edu
From: wagnermr@iupui.edu
From: zqian@umich.edu
From: antranig@caret.cam.ac.uk
From: gopal.ramasammycook@gmail.com
From: david.horwitz@uct.ac.za
From: david.horwitz@uct.ac.za
From: david.horwitz@uct.ac.za
From: david.horwitz@uct.ac.za
From: stephen.marquard@uct.ac.za
From: louis@media.berkeley.edu
From: louis@media.berkeley.edu
From: ray@media.berkeley.edu
From: cwen@iupui.edu
From: cwen@iupui.edu
From: cwen@iupui.edu
```

In [10]:
```
#Qns :4

import re
mail=open("mbox_short.txt")
for line in mail:
    line=line.rstrip()
    if re.search('From.+@',line):
        print(line)
```

```
From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
From: stephen.marquard@uct.ac.za
From louis@media.berkeley.edu Fri Jan  4 18:10:48 2008
From: louis@media.berkeley.edu
From zqian@umich.edu Fri Jan  4 16:10:39 2008
From: zqian@umich.edu
From rjlowe@iupui.edu Fri Jan  4 15:46:24 2008
From: rjlowe@iupui.edu
From zqian@umich.edu Fri Jan  4 15:03:18 2008
From: zqian@umich.edu
From rjlowe@iupui.edu Fri Jan  4 14:50:18 2008
From: rjlowe@iupui.edu
From cwen@iupui.edu Fri Jan  4 11:37:30 2008
From: cwen@iupui.edu
From cwen@iupui.edu Fri Jan  4 11:35:08 2008
From: cwen@iupui.edu
From gsilver@umich.edu Fri Jan  4 11:12:37 2008
From: gsilver@umich.edu
From gsilver@umich.edu Fri Jan  4 11:11:52 2008
From: gsilver@umich.edu
From zqian@umich.edu Fri Jan  4 11:11:03 2008
From: zqian@umich.edu
From gsilver@umich.edu Fri Jan  4 11:10:22 2008
From: gsilver@umich.edu
From wagnermr@iupui.edu Fri Jan  4 10:38:42 2008
From: wagnermr@iupui.edu
From zqian@umich.edu Fri Jan  4 10:17:43 2008
From: zqian@umich.edu
From antranig@caret.cam.ac.uk Fri Jan  4 10:04:14 2008
From: antranig@caret.cam.ac.uk
From gopal.ramasammycook@gmail.com Fri Jan  4 09:05:31 2008
From: gopal.ramasammycook@gmail.com
From david.horwitz@uct.ac.za Fri Jan  4 07:02:32 2008
From: david.horwitz@uct.ac.za
From david.horwitz@uct.ac.za Fri Jan  4 06:08:27 2008
From: david.horwitz@uct.ac.za
From david.horwitz@uct.ac.za Fri Jan  4 04:49:08 2008
From: david.horwitz@uct.ac.za
From david.horwitz@uct.ac.za Fri Jan  4 04:33:44 2008
From: david.horwitz@uct.ac.za
From stephen.marquard@uct.ac.za Fri Jan  4 04:07:34 2008
From: stephen.marquard@uct.ac.za
From louis@media.berkeley.edu Thu Jan  3 19:51:21 2008
From: louis@media.berkeley.edu
From louis@media.berkeley.edu Thu Jan  3 17:18:23 2008
From: louis@media.berkeley.edu
From ray@media.berkeley.edu Thu Jan  3 17:07:00 2008
```

```
From: ray@media.berkeley.edu
From cwen@iupui.edu Thu Jan  3 16:34:40 2008
From: cwen@iupui.edu
From cwen@iupui.edu Thu Jan  3 16:29:07 2008
From: cwen@iupui.edu
From cwen@iupui.edu Thu Jan  3 16:23:48 2008
From: cwen@iupui.edu
```

In [11]:

```
#Qns :5

s="Swathi send a message from the address swathi@caret.cam.ac.uk to Jmbo addr
fm=re.findall('\S+@\S+',s)
print("The Mail Adress :")
for i in fm:
    print(i)
```

```
The Mail Adress :
swathi@caret.cam.ac.uk
jumbo1924@gmail.com
```

In [12]:

```
#Qns :6

import re
mail=open("mbox_short.txt")
for line in mail:
    line=line.rstrip()
    txt=re.findall('\S+@\S+',line)
    if len(txt)>0:
        print(txt)
```

```
['stephen.marquard@uct.ac.za']
['<postmaster@collab.sakaiproject.org>']
['<200801051412.m05ECIaH010327@nakamura.uits.iupui.edu>']
['<source@collab.sakaiproject.org>;']
['<source@collab.sakaiproject.org>;']
['<source@collab.sakaiproject.org>;']
['apache@localhost)']
['source@collab.sakaiproject.org;']
['stephen.marquard@uct.ac.za']
['source@collab.sakaiproject.org']
['stephen.marquard@uct.ac.za']
['stephen.marquard@uct.ac.za']
['louis@media.berkeley.edu']
['<postmaster@collab.sakaiproject.org>']
['<200801042308.m04N8v6O008125@nakamura.uits.iupui.edu>']
['<source@collab.sakaiproject.org>;']
['<source@collab.sakaiproject.org>;']
['<source@collab.sakaiproject.org>;']
['apache@localhost)']
```

In [13]:
```
#Qns :7

import re
mail=open("mbox_short.txt")
for line in mail:
    line=line.rstrip()
    txt=re.findall('[a-zA-Z0-9]\S*@\S*[a-zA-Z0-9]',line)
    if len(txt)>0:
        print(txt)
```

```
['stephen.marquard@uct.ac.za']
['postmaster@collab.sakaiproject.org']
['200801051412.m05ECIaH010327@nakamura.uits.iupui.edu']
['source@collab.sakaiproject.org']
['source@collab.sakaiproject.org']
['source@collab.sakaiproject.org']
['apache@localhost']
['source@collab.sakaiproject.org']
['stephen.marquard@uct.ac.za']
['source@collab.sakaiproject.org']
['stephen.marquard@uct.ac.za']
['stephen.marquard@uct.ac.za']
['louis@media.berkeley.edu']
['postmaster@collab.sakaiproject.org']
['200801042308.m04N8v6O008125@nakamura.uits.iupui.edu']
['source@collab.sakaiproject.org']
['source@collab.sakaiproject.org']
['source@collab.sakaiproject.org']
['apache@localhost']
```

In [14]:
```
#Qns :8

import re
mail=open("mbox_short.txt")
for line in mail:
    line=line.rstrip()
    if re.search('X\S*: [0-9]+',line):
        print(line)
```

```
X-DSPAM-Confidence: 0.8475
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.6178
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.6961
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.7565
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.7626
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.7556
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.7002
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.7615
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.7601
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.7605
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.6959
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.7606
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.7559
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.7605
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.6932
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.7558
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.6526
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.6948
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.6528
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.7002
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.7554
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.6956
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.6959
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.7556
```

```
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.9846
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.8509
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.9907
X-DSPAM-Probability: 0.0000
```

In [15]:
```
#Qns :9

import re
mail=open("mbox_short.txt")
for line in mail:
    line=line.rstrip()
    txt=re.findall('Details:,*rev=([0-9]+)',line)
    if len(txt)>0:
        print(txt)
```

In [16]:
```
#Qns :10

import re
mail=open("mbox_short.txt")
for line in mail:
    line=line.rstrip()
    txt=re.findall('From.*([0-9][0-9])',line)
    if len(txt)>0:
        print(txt)
```

```
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
['08']
```

Qunestion 2 : Baby Names Popularrity Analysis

In [18]:
```
import re
import sys

def extract_names(filename):
    names=[]

    f=open(filename,'r')
    txt=f.read()
    year_match=re.search(r'Popularity\sin\s(\d\d\d\d)',txt)
    if not year_match:
        sys.stderr.write('Couldn\'\t find the year!\n\n')
        sys.exit(1)
    year=year_match.group(1)
    names.append(year)
    tuples=re.findall(r'<td>(\d+)</td><td>(\w+)</td>\<td>(\w+)</td>',txt)
    names_to_rank={}
    for rank_tuple in tuples:
        (rank,boyname,girlname)=rank_tuple
        if boyname not in names_to_rank:
            names_to_rank[boyname]=rank
        if girlname not in names_to_rank:
            names_to_rank[girlname]=rank
    sorted_name=sorted(names_to_rank.keys())
    for name in sorted_name:
        names.append(name+" "+names_to_rank[name])
    return (names)
```

In [19]:
```
extract_names("baby1990.html")
```

Out[19]:
```
['1990',
 'Aaron 34',
 'Abbey 482',
 'Abbie 685',
 'Abby 222',
 'Abdul 934',
 'Abel 384',
 'Abigail 90',
 'Abraham 246',
 'Abram 920',
 'Adam 32',
 'Adan 548',
 'Addison 645',
 'Adolfo 649',
 'Adrian 94',
 'Adriana 144',
 'Adrianna 325',
 'Adrianne 783',
 'Adrienne 233',
```

In [ ]: