Name: Atharva Rewatkar

Section: E

Roll Number: 32

Batch: E2

TASK A

Consider the text file with following URL & perform the following operation using Regular Expression

url = 'https://www.gutenberg.org/files/2638/2638-0.txt%27;

1. Find the number of the pronoun "the" in the corpus. Hint: Use the len() function.

2. Try to convert every single stand-alone instance of 'i' to 'I' in the corpus. Make sure not to change the 'i' occurring within a word:

3. Find the number of times anyone was quoted ("") in the corpus.

4. What are the words connected by '--' in the corpus?

5. Find the numbers available in the text.

6. Return all words of a string those starts with vowel.

7. Return all the roman numbers available in the file.

In [1]:
```
import re
import requests
```

In [2]:
```
url="https://www.gutenberg.org/files/2638/2638-0.txt"
path=r'https://www.gutenberg.org/files/2638/2638-0.txt'
response=requests.get(path)
data=response.text
```

In [3]:
```
the=re.compile("the")
the
re.compile(r'the',re.UNICODE)
print(len(data))
```

```
1427675
```

In [4]:
```
print("Number of times 'the' appeared: ",len(re.findall(the,data)))
```

```
Number of times 'the' appeared:  14424
```

In [5]:
```
s="i in it i am mica miles apart i am me"
s=re.sub(r"i","I",s)
print("The new string is: ",s)
```

```
The new string is:  I In It I am mIca mIles apart I am me
```

In [6]:
```
f=re.sub(r"i","I",data)
```

In [7]:
```
f[:150]
```

Out[7]:
```
'ï»¿The Project Gutenberg eBook of The IdIot, by Fyodor Dostoyevsky\r\n\r\nThIs eBook Is for the use of anyone anywhere In the UnIted States and\r\nmost othe'
```

In [8]:
```
quoted=(re.findall('"([^"]*)"',data))
print(len(quoted))
```

```
11
```

In [9]:
```
c=re.findall('\s[a-zA-Z]*.--.[a-zA-Z]*\s',data)
c
```

Out[9]:
```
[' one--the ', ' away--you ']
```

In [10]:
```
numbers=re.findall('\s[0-9]+\s',data)
print(numbers)
print("Total: ",len(numbers))
```

```
[' 2001 ', ' 1812 ', ' 60 ', ' 30 ', ' 90 ', ' 3 ', ' 90 ', ' 3 ', ' 4 ', ' 809 ', ' 1500 ', ' 50 ']
Total:  12
```

In [11]:
```
vow=re.findall('\s[AEIOUaeiou]+[a-z]*',data)
print("Total: ",len(vow))
```

```
Total:  67166
```

In [12]:
```
roman=re.findall(r"^M{0,3}(CM|CD|D?C{0,3})(XC|XL|L?X{0,3})(IX|IV|V?I{0,3})$",data)
print(roman)
```

```
[]
```

TASK B

i) Phone Number Verification

Problem Statement – The need to easily verify phone numbers in any relevant scenario. Consider the following Phone numbers:

444-122-1234

123-122-78999

111-123-23

67-7890-2019

The general format of a phone number is as follows:

Starts with 3 digits and '-' sign

3 middle digits and '-' sign

4 digits in the end

In [13]:
```
import re
phn = ["412-555-1212","123-122-78999","111-123-23","67-7890-2019"]
for i in phn:
    print(i)
    if re.search("\w{3}-\w{3}-\w{4}", i):
        print("Valid Phone Number")
    else:
        print("Invalid")
    print('\n')
```

```
412-555-1212
Valid Phone Number


123-122-78999
Valid Phone Number


111-123-23
Invalid


67-7890-2019
Invalid
```

ii) Email Verification

Problem statement – To verify the validity of an E-mail address in any scenario. Consider the following examples of email addresses:

Anirudh@gmail.com

Anirudh @ com

AC .com

123 @.com

All E-mail addresses should include:

1 to 20 lowercase and/or uppercase letters, numbers, plus . _ % + An @ symbol

2 to 20 lowercase and uppercase letters, numbers and plus

A period symbol

2 to 3 lowercase and uppercase letters

In [14]:
```
import re
email = [" db@.com"," @seo.com"," pm@.com","mp@xyz.com","Anirudh@gmail.com","Anirudh@com"]
x=[]
for i in email:
    x.append(re.findall("[\w._%+-]{1,20}@[\w.-]{2,20}.[A-Za-z]{2,3}",i))
print('Valid Emails are:')
for i in x:
    if(len(i)>0):
        print(''.join(i))
    else:
        pass
```

```
Valid Emails are:
mp@xyz.com
Anirudh@gmail.com
```

iii) Password Verification:

Write a Python program to check the validity of a password using Regular expression. Validation Rules:

At least 1 letter between [a-z A-Z]. At least 1 number between [0-9]. At least 1 character from [&#@]. Minimum length 6 characters.

In [15]:
```
import re
p= input("Input your password: ")
x = True
while x:
    if (len(p)<6):
        break
    elif not re.search("[a-zA-Z]",p):
        break
    elif not re.search("[0-9]",p):
        break
    elif not re.search("[&#@]",p):
        break
    elif re.search("\s",p):
        break
    else:
        print("Valid Password")
        x=False
        break
if x:
    print("Not a Valid Password")
```

```
Input your password: Atharva_4433#
Valid Password
```

TASK C

Problem Statement – Scrapping all of the phone numbers from a website for a requirement by making use of Python Regular Expressions & save it in CSV/ list Website URL:

http://www.summet.com/dmsi/html/codesamples/addresses.html

In [16]:
```
import urllib.request
from re import findall
url = "http://www.summet.com/dmsi/html/codesamples/addresses.html"
response = urllib.request.urlopen(url)
html = response.read()
htmlStr = html.decode()
pdata = findall("\(\d{3}\) \w{3}-\d{4}", htmlStr)
for item in pdata:
    print(item)
```

```
(257) 563-7401
(372) 587-2335
(786) 713-8616
(793) 151-6230
(492) 709-6392
(654) 393-5734
(404) 960-3807
(314) 244-6306
(947) 278-5929
(684) 579-1879
(389) 737-2852
(660) 663-4518
(608) 265-2215
(959) 119-8364
(468) 353-2641
(248) 675-4007
(939) 353-1107
(570) 873-7090
(302) 259-2375
(717) 450-4729
(453) 391-4650
(559) 104-5475
(387) 142-9434
(516) 745-4496
(326) 677-3419
(746) 679-2470
(455) 430-0989
(490) 936-4694
(985) 834-8285
(662) 661-1446
(802) 668-8240
(477) 768-9247
(791) 239-9057
(832) 109-0213
(837) 196-3274
(268) 442-2428
(850) 676-5117
(861) 546-5032
(176) 805-4108
(715) 912-6931
(993) 554-0563
(357) 616-5411
(121) 347-0086
(304) 506-6314
(425) 288-2332
(145) 987-4962
(187) 582-9707
(750) 558-3965
(492) 467-3131
(774) 914-2510
(888) 106-8550
(539) 567-3573
(693) 337-2849
(545) 604-9386
(221) 156-5026
(414) 876-0865
(932) 726-8645
(726) 710-9826
(622) 594-1662
(948) 600-8503
(605) 900-7508
(716) 977-5775
(368) 239-8275
(725) 342-0650
(711) 993-5187
(882) 399-5084
(287) 755-9948
(659) 551-3389
(275) 730-6868
(725) 757-4047
(314) 882-1496
(639) 360-7590
(168) 222-1592
(896) 303-1164
(203) 982-6130
(906) 217-1470
(614) 514-1269
(763) 409-5446
(836) 292-5324
(926) 709-3295
(963) 356-9268
(736) 522-8584
(410) 483-0352
(252) 204-1434
(874) 886-4174
(581) 379-7573
(983) 632-8597
(295) 983-3476
(873) 392-8802
(360) 669-3923
(840) 987-9449
(422) 517-6053
(126) 940-2753
(427) 930-5255
(689) 721-5145
(676) 334-2174
(437) 994-5270
(564) 908-6970
(577) 333-6244
(655) 840-6139
```