Shri Ramdeobaba College of Engineering and Management, Nagpur
Department of Computer Science and Engineering
Session: 2022-2023

Compiler Design Lab **V Semester [AIML]**

PRACTICAL No. 8

Name : Shantanu Mane

Roll No: E63

Topic Covered: Regular Expression (Python re package)

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'

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.

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

ii) E-mail 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

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.

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

```
import re
import requests
```

```
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
```

```
the=re.compile("the")
the
re.compile(r'the',re.UNICODE)
print(len(data))
```

```
1427675
```

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

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

```
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
```

```
f=re.sub(r"i","I",data)

f[:150]
```

```
'ï»¿The Project Gutenberg eBook of The IdIot, by Fyodor Dostoyevsky\r\n\r\nThIs eBook
```

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

```
11
```

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

```
[' one--the ', ' away--you ']
```

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

```
[' 2001 ', ' 1812 ', ' 60 ', ' 30 ', ' 90 ', ' 3 ', ' 90 ', ' 3 ', ' 4 ', ' 809 ', ' 1
Total:  12
```

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

```
Total:  67166
```

```
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)
```

```
[]
```

```
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
```

```
import re
email = [" db@.com"," @seo.com"," pm@.com","mp@xyz.com","Anirudh@gmail.com","Anirud"]
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
```

```
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
```

```
Input your password:  Shantanu@2002
Valid Password
```

```
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
```