Forward School

Program Code: J620-002-4:2020

Program Name: FRONT-END SOFTWARE DEVELOPMENT

Title : Regular Expressions - Part 2

Name: Chuay Xiang Ze

IC Number: 021224070255

Date : 6/7/2023

Introduction : Learning how to extract data using regex

Conclusion : Managed to complete tasks relating to the topic.

P12 - Regular Expressions - Part 2

Extract Email using Regex

```python
import requests
import re

url='https://www.selangor.gov.my/index.php/pages/view/339'

# get the data
data = requests.get(url, verify=False)


emails = re.findall(r'([\d\w\.]+@[\d\w\.\-]+\.\w+)', data.text)

print(emails)
```

D:\Anaconda3\envs\python-dscourse\lib\site-packages\urllib3\connectionpoo
l.py:981: InsecureRequestWarning: Unverified HTTPS request is being made t
o host 'www.selangor.gov.my'. Adding certificate verification is strongly
advised. See: https://urllib3.readthedocs.io/en/latest/advanced-usage.html
#ssl-warnings (https://urllib3.readthedocs.io/en/latest/advanced-usage.htm
l#ssl-warnings)
  warnings.warn(

['cccsel@rmp.gov.my', 'pertanyaan@icu.gov.my', 'jpn.selangor@moe.gov.my',
'cprc_sel@moh.gov.my', 'hqweb@jupem.gov.my', 'p.selangor@jpj.gov.my', 'pin
sel@imi.gov.my', 'pppselangor@jpn.gov.my', 'pensel@inform.gov.my', 'jbssel
angor@gmail.com', 'kamarudin@hasil.gov.my', 'sel@sprm.gov.my', 'prn_selang
or@moha.gov.my', 'webmaster@jpph.gov.my', 'penduduk@lppkn.gov.my', 'ppn_se
langor@dof.gov.my', 'ppnselangor@perpaduan.gov.my', 'kemas_selangor2@kema
s.gov.my', 'jbaudit@audit.gov.my', 'pneg_sgr@anm.gov.my', 'pjlwt@marine.go
v.my', 'pro@jakoa.gov.my', 'wpkl@prison.gov.my', 'zikril@kpdnhep.gov.my',
'korporat@bomba.gov.my', 'webmaster@spr.gov.my', 'selangor@wildlife.gov.m
y', 'webmaster@mkn.gov.my', 'jmgselwp@jmg.gov.my', 'jhekssgor@mohr.gov.m
y', 'pmssubang@met.gov.my', 'ccc@customs.gov.my', 'jtknselangor@mohr.gov.m
y', 'jppmsel@mohr.gov.my', 'info@jkkn.gov.my', 'jpselangor@dosm.gov.my',
'pengarah_selangor@adk.gov.my', 'jkkpsl@mohr.gov.my', 'ppwnselangor@jpw.go
v.my', 'pro@civildefence.gov.my']

Extract Phone Number using Regex

```python
import requests
import re

url='https://www.selangor.gov.my/index.php/pages/view/339'

# get the data
data = requests.get(url, verify=False)

phones = re.findall(r'[0-9]{2}\-[0-9]{8}', data.text)
print(phones)
```

```
['03-55145004', '03-55195175', '03-55213600', '03-55213700', '03-5518650
0', '03-55102133', '03-51237333', '03-51237209', '03-55144000', '03-551326
13', '03-55669555', '03-55432202', '03-55190653', '03-55107255', '03-55117
355', '03-55136755', '03-55192196', '03-55121411', '03-55192326', '03-5519
2231', '03-55215200', '03-55103500', '03-55256500', '03-55256514', '03-552
56515', '03-55103436', '03-55103436', '03-55184603', '03-55197825', '03-55
107397', '03-55110915', '03-55100575', '03-55190169', '03-55190690', '03-5
5132655', '03-55184617', '03-55184618', '03-55195114', '03-55195319', '03-
55102376', '03-55193044', '03-55193175', '03-55199533', '03-55147400', '03
-55147404', '03-55147405', '03-55205200', '03-55105049', '03-31695100', '0
3-31695190', '03-55190375', '03-55111063', '03-87328299', '03-55144300',
'03-55195255', '03-78464444', '03-78469892', '03-55194273', '03-55400717',
'03-55193915', '03-55101830', '03-55218790', '03-55218794', '03-55218791',
'03-55447828', '03-55109705', '03-55101833', '03-55101918', '03-55193233',
'03-55193551', '03-55193457', '03-55199059', '03-78463114', '03-78464982',
'03-31693888', '03-31693600', '03-56328800', '03-56361605', '03-56361625',
'03-56361573', '03-56361534', '03-56501600', '03-56361534', '03-55102664',
'03-55102791', '03-55102839', '03-55150200', '03-55180408', '03-79568512',
'03-79576396', '03-56236400', '03-56389159', '03-55118891', '03-55118706',
'03-33411031', '03-33410506', '03-33410443', '03-33411894', '03-55447000']

D:\Anaconda3\envs\python-dscourse\lib\site-packages\urllib3\connectionpoo
l.py:981: InsecureRequestWarning: Unverified HTTPS request is being made t
o host 'www.selangor.gov.my'. Adding certificate verification is strongly
advised. See: https://urllib3.readthedocs.io/en/latest/advanced-usage.html
#ssl-warnings (https://urllib3.readthedocs.io/en/latest/advanced-usage.htm
l#ssl-warnings)
  warnings.warn(
```

Quiz

Quiz 1

Extract the Emails and Phone numbers from this page using RegEx

https://www.ksn.gov.my/mengenai-pejabat/direktori-pejabat (https://www.ksn.gov.my/mengenai-pejabat/direktori-pejabat)

In [34]:

```python
import requests
import re

url='https://www.ksn.gov.my/mengenai-pejabat/direktori-pejabat'

# get the data
data = requests.get(url, verify=False)

phones = re.findall(r'[0-9]{3}\-[0-9 ]{9}', data.text)
emails = re.findall(r'([[\d\w\.]+[at]]+[\d\w\.\-]+\.\w+)', data.text)

print(phones)
print(emails)
```

C:\Anaconda\envs\python-dscourse\lib\site-packages\urllib3\connectionpool.
py:1056: InsecureRequestWarning: Unverified HTTPS request is being made to
host 'www.ksn.gov.my'. Adding certificate verification is strongly advise
d. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-w
arnings (https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-
warnings)
  warnings.warn(

['603-8872 7321', '603-8872 7329', '603-8872 7327', '603-8872 7223', '603-
8872 7200', '603-8872 7216', '603-8872 7224', '603-8872 7218', '603-8872 7
211', '603-8872 7321', '603-8872 7212']
['zuki[at]jpm.gov.my', 'zuki[at]jpm.gov.my', 'dr.haniff[at]jpm.gov.my', 'h
asnah[at]jpm.gov.my', 'akram[at]jpm.gov.my', 'zulamri[at]jpm.gov.my', 'ahm
adnizam[at]jpm.gov.my', 'nor.izuan[at]jpm.gov.my', 'mohd.abduh[at]jpm.gov.
my', 'hasif[at]jpm.gov.my', 'hanisah[at]jpm.gov.my', 'faizah.bastamam[at]j
pm.gov.my']

C:\Users\Xiang Ze\AppData\Local\Temp\ipykernel_21844\1649087138.py:10: Fut
ureWarning: Possible nested set at position 2
  emails = re.findall(r'([[\d\w\.]+[at]]+[\d\w\.\-]+\.\w+)', data.text)

Quiz 2

Extract the Emails and Phone numbers from this page using RegEx

https://www.jpn.gov.my/my/hubungi-kami/direktori-kakitangan/bahagian-kewarganegaraan
(https://www.jpn.gov.my/my/hubungi-kami/direktori-kakitangan/bahagian-kewarganegaraan)

```python
import requests
import re

url='https://www.jpn.gov.my/my/hubungi-kami/direktori-kakitangan/bahagian-kewarganegaraa

# get the data
data = requests.get(url, verify=False)

phones = re.findall(r'[0-9]{2}\-[0-9 ]{9}', data.text)

print(phones)
```

```
C:\Anaconda\envs\python-dscourse\lib\site-packages\urllib3\connectionpool.
py:1056: InsecureRequestWarning: Unverified HTTPS request is being made to
host 'www.jpn.gov.my'. Adding certificate verification is strongly advise
d. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-w
arnings (https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-
warnings)
  warnings.warn(

['03-8880 7023', '03-8880 8185', '03-8880 8179', '03-8880 8185', '03-8880
8143', '03-8880 8185', '03-8880 7042', '03-8880 8185', '03-8880 8117', '03
-8880 8185', '03-8000 8000', '03-8880 8288']
```

Quiz 3

Extract the Emails and Phone Numbers from this page using RegEx

https://www.ptptn.gov.my/hubungi-ptptn (https://www.ptptn.gov.my/hubungi-ptptn)

```python
import requests
import re

url='https://www.ptptn.gov.my/hubungi-ptptn'

# get the data
data = requests.get(url, verify=False)

phones = re.findall(r'[0-9]{2}\-[0-9 ]{9}|[0-9]{2}\-[0-9 ]{8}|[0-9]{3}\-[0-9 ]{8}|[0-9]'
print(phones)
```

C:\Anaconda\envs\python-dscourse\lib\site-packages\urllib3\connectionpool.
py:1056: InsecureRequestWarning: Unverified HTTPS request is being made to
host 'www.ptptn.gov.my'. Adding certificate verification is strongly advis
ed. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-
warnings (https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl
-warnings)
  warnings.warn(
C:\Anaconda\envs\python-dscourse\lib\site-packages\urllib3\connectionpool.
py:1056: InsecureRequestWarning: Unverified HTTPS request is being made to
host 'www.ptptn.gov.my'. Adding certificate verification is strongly advis
ed. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-
warnings (https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl
-warnings)
  warnings.warn(

['03-21931177', '03-21931197', '011-51813747', '03-21931179', '03-2193118
4', '011-21326228', '011-12327988', '011-10952566', '011-11567607', '011-3
3518922', '011-17911236', '011-14201982', '011-40224241', '011-14201829',
'011-62016403', '011-14201832', '011-23718675', '011-12834055', '011-14202
056', '011-33178276', '011-1140 576', '011-1420 191', '011-11181518', '011
-27765184', '011-28932548', '03-55231630', '011-73449057', '011-14202017',
'011-11913055', '011-27282150', '011-21178900', '011-33091677', '011-25489
744', '011-14202063', '011-23637492', '011-37030877', '011-11294543', '011
-14202084', '011-31715313', '011-21106861', '011-12294739', '011-1237989
9', '011-37711539', '011-35065884', '011-28013600', '011-31929187', '011-1
2849088', '019-218 9132', '011-23625105', '011-31749310', '07-237 1088',
'011-26421223', '011-26384286', '011-28767644', '011-21349402', '012-265 7
306', '011-74120538', '011-10707204', '011-17471051', '011-24047600', '011
-51488755', '011-11318943', '05-801 2398', '012- 4690670', '011-35134245',
'04-226 2430', '011-14709963', '011-40737367', '011-36306991', '011-542514
70', '011-37846617', '011-12350288', '011-88880690', '011-19909159', '011-
14201934', '011-25519411', '011-16121236', '09-960 2800', '011-37533717',
'011-31212003', '011-6363 300', '011-12171847', '011-20668682', '013-365 2
998', '011-23749448', '011-11163836', '011-25779475', '011-10069030', '011
-31735788', '011-29993920', '011-16037479', '011-31449144', '011-3332949
2', '011-36016122', '011-31725631', '011-31236318']

In [38]:

```python
# Method 1

import requests
import re

url='https://www.ksn.gov.my/mengenai-pejabat/direktori-pejabat'

# get the data
```

In [40]:

```python
# Method 2

import requests
import re

url='https://www.ksn.gov.my/mengenai-pejabat/direktori-pejabat'

# get the data
```

In [41]:

```python
# Method 3

import requests
import re

url='https://www.ksn.gov.my/mengenai-pejabat/direktori-pejabat'

# get the data
```

In [39]:

```python
# Answer to Exercise 2

import requests
import re

user_agent = 'user-agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (

headers={'User-Agent':user_agent}

url='https://www.jpn.gov.my/my/hubungi-kami/direktori-kakitangan/bahagian-kewarganegaraa

# get the data
# disable SSL cert verification
```

In [9]:

```
data.text
```

Out[9]:

```
'<!DOCTYPE html>\n<html lang="ms-my" dir="ltr" vocab="http://schema.or
g/">\n    <head>\n          <meta http-equiv="X-UA-Compatible" content="I
E=edge">\n      <meta name="viewport" content="width=device-width, in
itial-scale=1">\n        <link rel="shortcut icon" href="/images/logojp
n40px.png">\n        <link rel="apple-touch-icon" href="/images/logojpn
40px.png">\n        <meta charset="utf-8" />\n\t<base href="https://ww
w.jpn.gov.my/my/hubungi-kami/direktori-kakitangan/bahagian-kewarganegar
aan" />\n\t<meta name="description" content="Portal Rasmi JPN" />\n\t<m
eta name="generator" content="XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX"
/>\n\t<title>Portal JPN - Bahagian Kewarganegaraan</title>\n\t<link hre
f="/my/hubungi-kami/direktori-kakitangan/bahagian-kewarganegaraan?forma
t=feed&amp;type=rss" rel="alternate" type="application/rss+xml" title
="RSS 2.0" />\n\t<link href="/my/hubungi-kami/direktori-kakitangan/baha
gian-kewarganegaraan?format=feed&amp;type=atom" rel="alternate" type="a
pplication/atom+xml" title="Atom 1.0" />\n\t<link href="https://www.jp
n.gov.my/my/hubungi-kami/direktori-kakitangan/bahagian-kewarganegaraan"
rel="alternate" hreflang="ms-MY" />\n\t<link href="https://www.jpn.gov.
my/en/contact-us/direktori-kakitangan/citizenship-division" rel="altern
```

In [37]:

```
# Answer to Exercise 3

import requests
import re

url='https://www.ptptn.gov.my/hubungi-ptptn'

# get the data
```

```
D:\Anaconda3\envs\python-dscourse\lib\site-packages\urllib3\connectionp
ool.py:981: InsecureRequestWarning: Unverified HTTPS request is being m
ade to host 'www.ptptn.gov.my'. Adding certificate verification is stro
ngly advised. See: https://urllib3.readthedocs.io/en/latest/advanced-us
age.html#ssl-warnings (https://urllib3.readthedocs.io/en/latest/advance
d-usage.html#ssl-warnings)
  warnings.warn(

['ahmadshahril@ptptn.gov.my', 'sarah@ptptn.gov.my', 'umminadira@ptptn.g
ov.my', 'nabillah@ptptn.gov.my', 'noorshaheera@ptptn.gov.my', 'fadzlina
@ptptn.gov.my', 'nurulhusna@ptptn.gov.my', 'nurul_atiqah@ptptn.gov.my',
'nur_atikah@ptptn.gov.my', 'nadhirah@ptptn.gov.my', 'noorehan@ptptn.go
v.my', 'ismail.m@ptptn.gov.my', 'norzaidah@ptptn.gov.my', 'farahanim_a@
ptptn.gov.my', 'masliana@ptptn.gov.my', 'arif_h@ptptn.gov.my', 'kamarul
_a@ptptn.gov.my', 'fazaitum@ptptn.gov.my', 'wan_norasyikin@ptptn.gov.m
y', 'karthinee@ptptn.gov.my', 'izzasuria@ptptn.gov.my', 's.nurul@ptptn.
gov.my', 'nursyafira@ptptn.gov.my', 'hafsoh@ptptn.gov.my', 'nuraishah@p
tptn.gov.my', 'norhidayati@ptptn.gov.my', 's.nurhaishah@ptptn.gov.my',
'roslan@ptptn.gov.my', 'haslida@ptptn.gov.my', 'md_ismail@ptptn.gov.m
y', 'amani@ptptn.gov.my', 'malissa@ptptn.gov.my', 'm_shafik@ptptn.gov.m
```

```
data.text
```

Need to use Selenium because of the Javascript on the page.

Discussion

Is regex good for scraping non regular texts from Web pages?

Eg. Look for all text between
* and

can work? yes but how usable?

Eg. Look for all names starting with Mr. and then extract the name. How?

We need a better way to extract the meta data and elements about the web page.

r'

[https://docs.python.org/3/library/re.html (https://docs.python.org/3/library/re.html)](https://docs.python.org/3/library/re.html)

> Regular expressions use the backslash character ('\') to indicate special forms or to allow special characters to be used without invoking their special meaning. This collides with Python's usage of the same character for the same purpose in string literals; for example, to match a literal backslash, one might have to write '\\\\' as the pattern string, because the regular expression must be \\, and each backslash must be expressed as \\ inside a regular Python string literal. Also, please note that any invalid escape sequences in Python's usage of the backslash in string literals now generate a DeprecationWarning and in the future this will become a SyntaxError. This behaviour will happen even if it is a valid escape sequence for a regular expression.
>
> The solution is to use Python's raw string notation for regular expression patterns; backslashes are not handled in any special way in a string literal prefixed with 'r'. So r"\n" is a two-character string containing '\' and 'n', while "\n" is a one-character string containing a newline. Usually patterns will be expressed in Python code using this raw string notation.

Note:

Alternative 3rd party regex implementation

[https://pypi.org/project/regex/ (https://pypi.org/project/regex/)](https://pypi.org/project/regex/)

Try it out own your own

In [8]:

```
# Example

teststring = 'this is \n a test'

print(teststring)
```

this is
 a test

In [16]:

```
# Example escape character \ is now a raw string not an escape char

teststring = r'this is \n a test'

print(teststring)
```

this is \n a test

In [18]:

```
# \b Word boundary, allow to perform "whole words only" search
import re
re.findall('\btest\b', 'test this is a test') # the backslash gets consumed by the pytho
```

Out[18]:

[]

In [19]:

```
re.findall('\\btest\\b', 'test this is a test') # backslash is explicitly escaped and is
```

Out[19]:

['test', 'test']

In [14]:

```
re.findall(r'\btest\b', 'test this is a test') # often this syntax is easier
```

Out[14]:

['test', 'test']

In [21]:

```python
# some example using regex to extract the URL

import re

x="""<!DOCTYPE html>

<html itemscope itemtype="http://schema.org/QAPage">

<head>
"""

matching = re.findall(r"[a-zA-Z0-9\-\.]+\.(?:com|org|net|mil|edu|COM|ORG|NET|MIL|EDU)+(?

print(matching)
```

['schema.org/QAPage']