Web Scraping Trilogy

1. Communicating with **API**.
2. Fetching web content with **BeautifulSoup**.
3. Controlling real web browser to access content via **Selenium**.

Downloading file

```
from urllib.request import urlretrieve
url = "https://cm540-example.netlify.app/kitten.jpeg"
urlretrieve(url, “kitten.jpeg")
```

Finding patterns in URL

Example patterns

```
https://docs.python.org/3/search.html?
q=namedtuple&check_keywords=yes&area=default
```

```
https://duckduckgo.com/?q=python+doc
```

```
https://www.google.com/maps/search/Libraries/
@22.1612464,113.5303786,13z
```

```
http://macaodaily.com/html/2020-05/04/node_2.htm
```

```
http://www.dicj.gov.mo/web/cn/information/
DadosEstat_mensal/2020/index.html
```

```
https://bis.dsat.gov.mo:37812/macauweb/routeLine.html?
routeName=3&direction=0&language=zh-tw&ver=3.5.12
```

Code example: searching Google map:

```
import webbrowser

query = input("Please input search query to search near-by
Macao. ")

# A map search in Macao.
url = f"https://www.google.com/maps/search/{query}/
@22.1612464,113.5303786,13z"

webbrowser.open(url)
```

API

API stands for Application Programming Interface.

Example of XML data

https://xml.smg.gov.mo/c_actual_brief.xml

```
<?xml version="1.0" encoding="UTF-8" ?>
<ActualWeatherBrief>
    <System>
        <SysAuthor>DINF</SysAuthor>
        <SysPubdate>2020-08-17 15:55</SysPubdate>
        <SysLanguage>0</SysLanguage>
    </System>
    <Custom>
        <ValidFor>2020-08-17 16:00</ValidFor>
        <Temperature>
            <MeasureUnit>°C</MeasureUnit>
            <Type>3</Type>
            <Value>28</Value>
        </Temperature>
        <Humidity>
            <MeasureUnit>%</MeasureUnit>
            <Type>3</Type>
            <Value>86</Value>
        </Humidity>
        <WindSpeed>
            <MeasureUnit>km/hr</MeasureUnit>
            <Type>3</Type>
            <Value>6</Value>
            <WindSpeedDescription>二級</WindSpeedDescription>
        </WindSpeed>
        <WindDirection>
            <MeasureUnit>°</MeasureUnit>
            <Type>3</Type>
            <Value>ESE</Value>
            <WindDescription>東南偏東</WindDescription>
        </WindDirection>
        <Icon>
            <IconName>ww-c03.gif</IconName>
            <IconURL>http://www.smg.gov.mo/icons/weatherIcon/
ww-c03.gif</IconURL>
        </Icon>
        <WeatherStatus>03</WeatherStatus>
        <humanAT>32</humanAT>
        <comfK>76</comfK>
```

```
        <comfK_desc>2</comfK_desc>
    </Custom>
</ActualWeatherBrief>
```

Example of JSON data

https://api.exchangeratesapi.io/latest?symbols=HKD,EUR&base=CNY

```
{
  "rates": {
    "EUR": 0.1218026797,
    "HKD": 1.1151644336
  },
  "base": "CNY",
  "date": "2020-08-14"
}
```

Code Example: Current Macao weather

Please install untangle via pip.

```
import untangle
import datetime

obj = untangle.parse('https://xml.smg.gov.mo/
c_actual_brief.xml')

humidity = obj.ActualWeatherBrief.Custom.Humidity.Value.cdata

if type(obj.ActualWeatherBrief.Custom.Temperature) == list:
    temperature =
obj.ActualWeatherBrief.Custom.Temperature[0].Value.cdata
else:
    temperature =
obj.ActualWeatherBrief.Custom.Temperature.Value.cdata


print(f"現時澳門氣溫 {temperature} 度，濕度 {humidity}%。")
```

Code Example: JSON API for Exchange rate

```
import json
import requests

url = "https://api.exchangeratesapi.io/latest?
symbols=HKD&base=CNY"

response = requests.get(url)
data = json.loads(response.text)
print(data)

print(data['rates']['HKD'])
```

Finding URLs and Inspecting HTML elements

1. In web browser, press F12, or right click and select "Inspect Elements".
2. This will open the Developer Panel.
3. We may check the "Elements" tabs for the HTML elements tree.
4. We may check "Network" and "XHR" for the JavaScript controlled network requests.

Noted for Safari user on macOS. Please enable "Show Developer menu in menu bar" in advanced preferences.