Beginning Python Programming Lesson 6

#### **Essential CSS Selectors**

Selectors are used to match elements in the HTML content tree.

### tag

```
without any marks. e.g. h1, a, p, main
```

#### .classname

```
prefix with . e.g. .content-title
```

### #id

```
prefix with # e.g. #logo
```

#### A B

Anything matches B within A. e.g. main a

#### A > B

```
Anything matches B under A, one level-only. e.g. ul.menu > li
```

### :first-of-type

The first encountered of matched element. e.g. main p:first-of-type

More selector exercise on CSS Diner: https://flukeout.github.io

# **BeautifulSoup Essential Functions**

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc, 'html.parser')

# Select elements with CSS selectors
# It always returns a list
soup.select(".main a")

# Select first matched element with CSS selectors
# It always returns one element, or None
soup.select_one("a")

# We can further select elements within a scope
```

```
h2 = soup.select_one("h2")
h2.select("a")

# Getting text value from element
soup.select_one("h1").text

# Getting attribute value from element
# e.g. getting href from a from first h1.
soup.select_one("h1 a")['href']

More usages in https://mak.la/bs4
```

# **Possible HTML parser options**

- html.parser
- lxml
- html5lib

# Code example: Fetching news title from news.gov.mo

```
from bs4 import BeautifulSoup
import requests

res = requests.get("https://news.gov.mo/home/zh-hant")
soup = BeautifulSoup(res.text, "html.parser")

for h5 in soup.select("h5"):
    print(h5.text.strip())
```

### **Catching network error**

```
try:
    res = requests.get("https://news.gov.mo/home/zh-
hant")
except requests.exceptions.ConnectionError:
    print("Error: Invalid URL or Connection Lost.")
    exit()

soup = BeautifulSoup(res.text, "html.parser")

for h5 in soup.select("h5"):
    print(h5.text.strip())
```

Beginning Python Programming Lesson 6

## Code Example: Fetching detail page per found link

```
res = requests.get("https://news.gov.mo/home/zh-hant")
soup = BeautifulSoup(res.text, "html.parser")
for h5 in soup.select("h5")[:5]:
   print(h5.getText().strip())
   # Fetch the content
   href = h5.select one("a")["href"]
    res = requests.get("https://news.gov.mo/" + href)
   soup2 = BeautifulSoup(res.text, "html.parser")
    content = soup2.select one(".asideBody p:first-of-
type")
    print(content.text)
   print("---")
```

# **Code example: Fetching Macao Daily news**

```
from bs4 import BeautifulSoup
import requests
import datetime
today = datetime.date.today()
vear = todav.vear
month = today.month
day = today.day
month = str(month).zfill(2)
day = str(day).zfill(2)
res = requests.get(f"http://www.macaodaily.com/html/
{vear}-{month}/{day}/node 1.htm")
res_encoding = "utf-8"
soup = BeautifulSoup(res.text, "html.parser") # Be aware
that you may need a different parser if "lxml" not found.
links = soup.select("#all_article_list a")
for link in links[:40]:
    print(link.text)
```

## Code example: What is next holiday in Macao?

```
import datetime
```

```
url = f"https://www.gov.mo/zh-hant/public-holidays/year-
{datetime.date.today().year}/"
response = requests.get(url)
soup = BeautifulSoup(response.text, "html.parser")
month = soup.select("#public-holidays .month")[0].text
day = soup.select("#public-holidays .day")[0].text
weekday = soup.select("#public-holidays .weekday")
[0].text
description = soup.select("#next-holiday-description
strong")[0].text
print(f"接下來的公眾假期:{description}, {month}{day}日
```

{weekday}") • • • DevTools - www.gov.mo/zh-hant/public-holidavs/vear-2020/ □ Elements Console Sources Network Performance Memory Application Security » 🛕 2 🗏 24 🕴 🚼 ▼<aside class="sidebar">

```
Styles Computed Event Listeners >>>
             ▼<div id="public-holidays"> == $0
                                                                                                         :hov .cls +
                ▼ <div class="card card-underline">
                   ::before
                                                                                  element.style {
                  ▶ <div class="card-head">...</div>
                  ▼<div class="card-body">
                                                                                                    bootstrap.min.css:1
                    ::before
                                                                                  -webkit-box-sizing: border-box;
                    ▼<div class="row">
                                                                                      -moz-hox-sizing: border-box:
                      ::before
                                                                                  box-sizing: border-box;
                      ▼<div class="col-xs-12 col-sm-4 date">
                       ▼<div class="valign-wranner">
                                                                                  div {
                                                                                                  user agent stylesheet
                         ▼<div class="valign">
                                                                                     display: block;
                            <span class="weekday">星期四</span>
                            <span class="day text-danger"> 1</span>
                                                                                  Inherited from aside.sidebar
                            <br>>
                                                                                  <span class="month">+ Fl</span>
                                                                                    font-size: 1em:
                           </div>
                         </div>
                                                                                  Inherited from body.header-...
                       </div>
                      ▼ <div class="col-xs-12 col-sm-8 description">
                                                                                  body:lang(zh govmo.min.c...?20200124:1
                       ▼<div class="valign-wrapper">
                         ▼<div class="valign" id="next-holiday-description">
                                                                                    font-family: "Open
                           ▼class="no-margin">
                                                                                          Sans", "Roboto", "Microsoft
                                                                                       JhengHei UI","Microsoft
JhengHei",Helvetica,Arial,sans-
                             <strong>中華人民共和國國慶日</strong>
                              "強制性公眾假期"
                                                                                     font-size: 1.7em:
                            </div>
                                                                                  body {
                                                                                                 materialadmin.min.css:1
                         </div>
                                                                                    background: ▶ □#fff;
                                                                                     background-size: 100% 100%;
... div#public-holidays div.card.card-underline div.card-body div.row div div div span
```

color: #313534: