Python & Data - Week 11

Web Scraping - Beautiful Soup 及進一步處理

打包下載

Beautiful Soup

1. 使用 select 方法搜尋

`soup.select(css_selector)`

> 與 find_all 類似，但是 select 接受 css_selector 作為參數，例如： `div.news-item-container`
> `div` 為標籤名稱，`.news-item-container` 為 class 名。
> 可同時選取多個 CSS 的 class

使用上週的方法抓取"澳門特區入口網站消息"

網址: https://www.gov.mo/zh-hant/news/

In [ ]:
```
import requests
from bs4 import BeautifulSoup

url = "https://www.gov.mo/zh-hant/news/"
page = requests.get(url)

soup = BeautifulSoup(page.content, 'html.parser')

news_container = soup.select('div.news--item-body')
for news in news_container:
    title = news.find('h')
    if title is not None:
        print(title.text)

# Not yet finished
```

```
交通事務局（DSAT）
"2021/2022學年大專學生學習用品津貼"接受登記
衛生局（SS）
漁船火警
海事局海關持續跟進救援工作
海事及水務局（DSAMA）
內港漁船火警 海事局海關全力救援
社會工作局（IAS）
（圖文包）開設合資格家務工作準外僱或外僱入境限制豁免申請
生產力中心領澳時裝設計參加大灣區巡展
一汗藝術小組-夢之寓言-威尼斯國際藝術雙年展揭幕
4月26日及晨1時起新增及取消對首到內地相關區域人士的防疫措施
藝術節多項節目門票續發售
文化局（IC）
行政長官賀一誠會見紀念"五•四"青年節系列活動籌備委員會一行
行政長官賀一誠會見紀念"五四"青年節系列活動籌備委員會一行
"五·一"迎客：加推酒店購物美食優惠
衛生局（SS）
```

1. 出現問題：同一個網頁內有兩款不同排版的消息

1. 解決方法之一

- 先從 .news--item-head 找尋 h3；
- 如果沒有，則在 .news--item-body 找尋 h3。

如何繼續抓取其他內容

1. 在合理的時間內抓取
2. 將工作排隊進行

在合理的時間內抓取

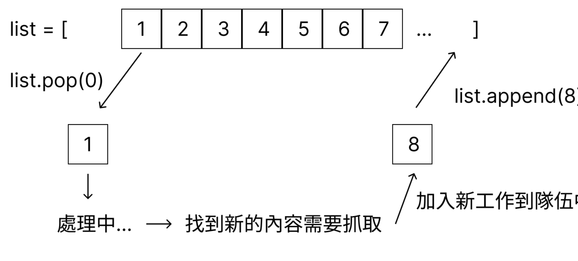
可以使用 `time.sleep(seconds)` 將程式延遲執行程式，等待數秒後繼續執行。

In [ ]:
```
# Sending request with delay
import time

urls = ['http://www.google.com/1.html', 'http://www.google.com/2.html', 'http://www.google.com/3.html']

def request_without_delay():
    startTime = time.time()
    for url in urls:
        print(f"Fetching {url}...")
    executionTime = time.time() - startTime
    print(f'Time used: {executionTime} seconds')


def request_with_delay():
    startTime = time.time()
    for url in urls:
        print(f"Fetching {url}...")
        time.sleep(1) # Sleep for 1 seconds
    executionTime = time.time() - startTime
    print(f'Time used: {executionTime} seconds')


print('Request without delay')
request_without_delay()

print('Request with delay')
request_with_delay()
```

```
Request without delay
Fetching http://www.google.com/1.html...
Fetching http://www.google.com/2.html...
Fetching http://www.google.com/3.html...
Time used: 2.9802322387695312e-05 seconds
Request with delay
Fetching http://www.google.com/1.html...
Fetching http://www.google.com/2.html...
Fetching http://www.google.com/3.html...
Time used: 3.0016019344329834 seconds
```

** 3.3140182495117 19e-05 seconds ~ 0.0000331401824951 seconds

將工作排隊進行



- 使用 `list.pop(0)` 將內容由 list 的開頭抽出；
- 使用 `list.append(x)` 將 x 加入到 list 的最後。

In [ ]:
```
# queue example

count = 0 # keep track of count to stop when count reaches 10
queue = [1]
while len(queue) > 0 and count < 10:
    item = queue.pop(0) # fetch the first item
    print(f'processing {item}')
    new_item = item + 1
    queue.append(new_item)
    print(f'add new item {new_item}')
    count += 1
    if count >= 10:
        print('processing stopped because count >= 10, new item will not be processed')
```

```
processing 1
add new item 2
processing 2
add new item 3
processing 3
add new item 4
processing 4
add new item 5
processing 5
add new item 6
processing 6
add new item 7
processing 7
add new item 8
processing 8
add new item 9
processing 9
add new item 10
processing 10
add new item 11
processing stopped because count >= 10, new item will not be processed
```

👉 將暫停和排隊功能應用到 Web Scraping 的處理上...