Python & Data - Week 12

Web Scraping - 進階處理

打包下載

抓取"澳門特區入口網站消息" - 解答之一

In [ ]:

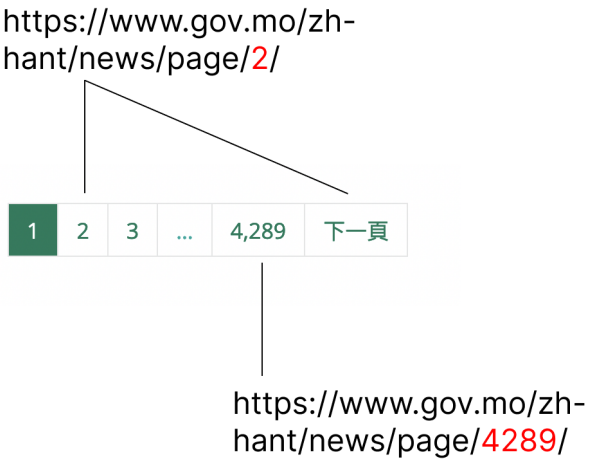
```
import requests
from bs4 import BeautifulSoup

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

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

news_headings = []
news_container = soup.select('div.news--item')
for news in news_container:
    title_in_head = news.select('div.news--item-head h3')
    if len(title_in_head) > 0:
        # print(title_in_head[0].text)
        news_headings.append(title_in_head[0].text)
    else:
        title_in_body = news.select('div.news--item-body h3')
        if len(title_in_body) > 0:
            news_headings.append(title_in_body[0].text)


print(news_headings)
print(f'total news items: {len(news_headings)}')
```

```
[' 5月3日凌晨1時起新增及取消對曾到內地相關區域人士的防疫措施 ', ' 新型冠狀病毒感染應變協調中心查詢熱線統計數字 (由2022年05月02日08H00至16H00) ', ' 新型冠狀病毒感染應變協調中心查詢熱線統計數字 (由2022年05月01日08H00至2022年05月02日08H00) ', ' 新增1例新冠病毒復陽個案　為由英國回澳的澳門居民 ', '“2022翱遊澳門無人機表演盛會”明晚精彩上演 首晚主題“春臨澳門” 寓意萬物復甦', '風季將至\u3000海事局協調海上工程防颱避風保安全', '5月2日凌晨1時起新增及取消對曾到內地相關區域人士的防疫措施', ' 新型冠狀病毒感染應變協調中心查詢熱線統計數字 (由2022年05月1日08H00至16H00) ', ' 氹仔柯維納馬路重型客車停車場5月4日起開放 ', ' 新增2例輸入性新冠病毒無症狀感染個案　分別為由泰國及香港回澳的澳門居民 ', ' 新型冠狀病毒感染應變協調中心查詢熱線統計數字\r\n (由2022年04月30日08H00至2022年5月1日08H00) ', '澳門入境旅客連日上升 昨（30日）錄得逾4萬人次新高', '法務局續推網上法律知識闖關遊戲', '2022“五一黑沙海灘迎夏日康體活動”因天氣影響取消', '勞工局與法務局合辦“網上法律知識闖關遊戲 — 勞動權益”特別大賽', '5月1日凌晨1時起新增及取消對曾到內地相關區域人士的防疫措施', ' 衛生局接獲1宗流感樣疾病群集性感染報告 ', '勞工局跟進工作意外', '建校十載樹中葡賢才\u3000鄭觀應公立學校迎校慶', '受天氣影響5月1日及2日的澳門無人機表演延後至5月3日及4日']
total news items: 20
```

將以上程式改寫，令它可以動態處理其他頁面

請參考 Week 11：

1. 使用 time.delay，安排在合理的時間間隔發出 request；
2. 使用排隊的方式，使程式可以順序處理發出 request。



甚麼時候停止 Web Scraping?

1. 明確的結束條件 (4289頁)
2. 不明確的結束條件 (使用其他方式判段停止時機)

Web Scraping 模式

第一種 –

1. 使用 request.get 取得起始頁，並使用 Beautiful Soup 找到最後一頁 (4289)
2. `for i in range(1, 4289): request.get(i)`
3. 使用 Beautiful Soup 處理

第二種 –

1. `request.get` 取得起始頁
2. 使用 Beautiful Soup 處理，並檢查 "下一頁" 是否存在
3. 如果"下一頁" 存在，則重覆第 1 步；否則，結束。

使用多個隊伍 (Queue) 處理不同的工作

- 將 Web Scraping 的一同步驟分拆成數個工作 (可以以 function 表示)
- 將處理列表頁的工作和處理詳細頁的工作分開兩個隊伍，在處理資料時加入新的工作
- 當隊伍為空時，即所有工作處理完成，結束程式