Python & Data - Week 13

Web Scraping - 總結

打包下載

抓取"澳門特區入口網站消息" - 我的解答

💡 `next_page_element[0].get('href')` 和 `next_page_element[0]['href']` 相同，但是使用 `get` 方法時，當找不到 Key，會返回 `None`；而 `next_page_element[0]['href']` 則會出現 `KeyError`

In [ ]:

```
import requests
import time
from bs4 import BeautifulSoup

count = 0
limit = 3 # we set a limit to 3 to limit number of get requests
job_queue = ["https://www.gov.mo/zh-hant/news/"]
news_headings = []

def process_news_list(url):
    print(f"processing {url}...")

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

    # we find the headings in news
    news_container = soup.select('div.news--item')
    for news in news_container:
        title = news.select('h3')
        if len(title) > 0:
            news_headings.append(title[0].text.strip())

    # then, we find the next page to process
    next_page_element = soup.select('nav.pagination a.next.page-numbers')
    if len(next_page_element) > 0 and next_page_element[0].get('href') is n
        job_queue.append(next_page_element[0]['href'])


while len(job_queue) > 0 and count < limit:

    if len(job_queue) == 0:
        print('Processing ended')
        break

    url = job_queue.pop(0)
    process_news_list(url)
    time.sleep(3)

    count = count + 1

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

```
processing https://www.gov.mo/zh-hant/news/...
processing https://www.gov.mo/zh-hant/news/page/2/...
processing https://www.gov.mo/zh-hant/news/page/3/...
['政府數據中心周末進行網絡維護', '5張圖了解申請【菲律賓籍家務工作外僱】步驟', '5月8日凌晨1時起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07日08H00至16H00)', '第四晚無人機"冬日澳門"明（8）日壓軸登場', '「新型冠狀病毒感染社區治療中心」第一次跨部門演練順利進行', '社區治療中心第一次跨部門演練', '社會文化司司長歐陽瑜出席社區治療中心演練', '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06日08H00至2022年05月07日08H00)', '5月7日至9日繼續舉辦"學童接種日"', '"電消優惠"推行前加緊巡查促商戶守規則穩物價', '5月7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', '文化傳播月深入社區拓展文化傳承傳播', '街坊鄰里—陳顯耀紀實攝影展 傳承本土歷史文化及城市記憶', '整治下水道氹仔菜園路部份路段5月9日至23日封閉交通', '演藝音校青春的旋律音樂會展示教學成果', '教育及青年發展局黃嘉祺副局長就職', '中葡論壇常設秘書處參觀國家安全展', '仁伯爵綜合醫院模擬為新冠病人進行手術作實地演練', '衛生局發佈本年首季自殺死亡監測結果 \r\n市民如受情緒困擾應尋求專業協助及輔導', '澳大舉辦首個高管人員工商管理碩士工作坊', '關閘廣場東側行人天橋及周邊街道優化工程下周二展開', '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06日08H00至16H00)', '中院裁定騙取持續進修發展計劃資助的主謀上訴理由不成立', '消委會調查冰鮮及急凍豬肉價格', '《華夏小靈精家國情懷繪本叢書》及《動畫》隆重推出', '2022年1月至3月住宅樓價指數', '美副將大馬路斜行過路設施11日啟用', '第三晚無人機"金秋澳門"明（7）日閃耀夜空', '新增1例輸入性新冠病毒無症狀感染個案 為由加拿大回澳的澳門居民', '澳門理工大學視覺藝術畢業作品聯展勉大學生融國家發展大局', '新冠疫苗接種站、核酸檢測站於佛誕節假期的服務安排', '仁伯爵綜合醫院、衛生中心/站和捐血中心於佛誕節假期的服務安排', '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05日08H00至2022年05月06日08H00)', '"電消優惠"5.10登記最新8條短片易看易明', '2022年第1季支付卡和移動支付統計', '5月6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', '新冠肺炎疫情下當局對特殊照顧需要人士提供不同支援', '婦女及兒童事務委員會2022年度第一次全體會議', '文化局演藝學院舞校四學生獲佳績', '亞美娛樂訴金沙集團索要賠償被初級法院裁定敗訴', '澳大隊包攬中銀盃澳門區創業賽冠亞季軍', '內地勞動節假期澳門客況佳\r\n入境旅客及酒店入住率皆升', '澳門健康碼持續優化 現已具有多項功能', '衛生局持續推廣抗原檢測使用方法', '澳門理工大學語言著作獲劍橋大學出版社出版', '當局收到773名在外地就讀學生計劃回澳', '理工-貝爾英語中心各類英語課程現正接受報名', 'mRNA兒童劑型疫苗預計5月中到貨', '對具特殊照顧需要人士的支援圖文包', '醫學觀察酒店將於5月6日中午起開放接收預訂', '已獲確認、於措施生效入住醫學觀察酒店的預訂安排', '醫學觀察酒店最新安排', '公佈《人才引進制度》公開諮詢意見總結報告', '疫情下對具特殊照顧需要人士的支援', '出入境總體情況', '549人於酒店接受醫學觀察', '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05日08H00至16H00)', '菲律賓籍家務外僱豁免入境限制先導計劃第二階段', '2022年第1季公司統計']
total news items: 60
```

抓取"培正網站消息" - 我的解答

💡 只需要作少量修改就能對其他網站做 Web Scraping

In [ ]:

```
import requests
import time
from bs4 import BeautifulSoup

count = 0
limit = 3 # we set a limit to 3 to limit number of get requests
url_prefix = "http://www.puiching.edu.mo/"
job_queue = ["news/category/"]
news_headings = []

def process_news_list(url):
    print(f"processing {url}...")

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

    # we find the headings in news
    news_container = soup.select('div.news-item')
    for news in news_container:
        title = news.select('.news-title a')
        if len(title) > 0:
            news_headings.append(title[0].text.strip())

    # then, we find the next page to process
    next_page_element = soup.select('.pagination a.pagination-next')
    if len(next_page_element) > 0 and next_page_element[0].get('href') is n
        job_queue.append(next_page_element[0]['href'])


while len(job_queue) > 0 and count < limit:

    if len(job_queue) == 0:
        print('Processing ended')
        break

    url = job_queue.pop(0)
    process_news_list(url_prefix + url)
    time.sleep(3)

    count = count + 1

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

```
processing http://www.puiching.edu.mo/news/category/...
processing http://www.puiching.edu.mo/news/category/?start=5...
processing http://www.puiching.edu.mo/news/category/?start=10...
['斯坦福大學數學錦標賽獲傲人成績', '學界游泳比賽奪23金；破7項大會紀錄', '培正同學會主辦乒乓球聯誼賽其樂融融', '澳門文學獎暨讀後感徵文獲佳績', '本校教師參與誠信教案設計取優異成績', '望社級社成立典禮順利舉行', '「青年法律工作者走進校園」辦培正專場', '茶文化導師開設春茶茶會', '初一同學參觀全民國家安全教育展', '中文科老師於南方傳媒杯作文大賽獲獎', '無煙無毒繪畫比賽得季軍', '校際辯論比賽季軍獎座頒發', '本校同學參與無人機科普活動', '國際青少年科學家大會獲銀獎佳績', '本年度小學Dress Up Day揭開序幕']
total news items: 15
```