

Thread 程式作業

資電二 109504502 陳紓嫻

作業主題

利用 `python` 結合多執行緒 `multithread`，爬取【Meme 梗圖倉庫】的梗圖及 `gif`，使下載圖片的速度加速，一開始原本使用生產者-消費者的方法跑，但之後讓下載每一張圖片都用一條 `thread` 跑，速度變得更快了。（抓到的圖片→）



使用介紹

執行程式後，可以選擇想要下載的圖片，預設會下載 **4000** 張圖片，或可自行選擇頁數，並將圖片儲存於 **images** 的資料夾，如果沒有資料夾會自動創建。



```
which kind of image do you like, 梗圖(´▽`ʃ♡Ʒ) or 動圖(ಠ_ಠ)`ಠ_ಠ')ಠ_ಠ'???
```

```
Your answer: 梗圖
```

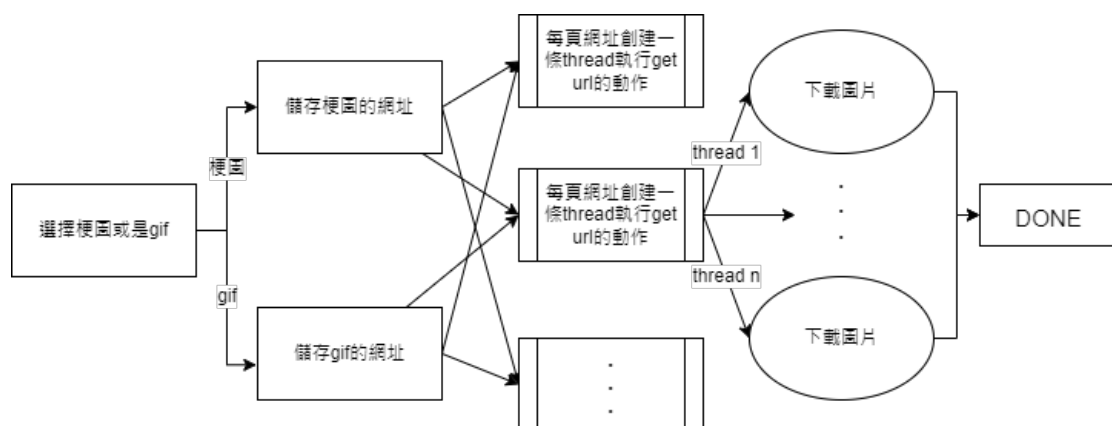
```
How many images do you want to download, 預設為4000張? default or choose: choose
```

```
Number of images: 20(張) * 請輸入頁數(MAX 250): 87
```

```
1649737029845.jpg is downloading.
```

```
1649736355380.jpg is downloading.
```

流程圖



函式介紹

```
17     # 提取圖片的URL
18     def get_url():
19         # 鎖住
20         gLock.acquire()
21         # 使用完後要釋放
22         if len(page_URL_list) == 0:
23             gLock.release()
24         else:
25             page_url = page_URL_list.pop()
26             gLock.release()
27             user_agent = ua.random
28             headers = {'user-agent': user_agent}
29             response = requests.get(page_url, headers=headers)
30             soup = BeautifulSoup(response.content, 'lxml')
31             img_list = soup.find_all('img', class_='img-fluid lazy')
32             time.sleep(0.5)
33
34         for image in img_list:
35             src = image['data-src']
36             if not src.startswith('http'):
37                 src = 'http:' + src
38             # 爬到url就開thread去下載
39             download_thread.append(threading.Thread(target=download, args=(src,)))
40             download_thread[len(download_thread) - 1].start()
```

【get_url】

每次從 `page_url_list` 提取一個網址下載當頁圖片，為防止 `thread` 同時 `pop` 掉 `page_url_list`，使用 `gLock.acquire()` 鎖住，`pop` 完即解鎖，因為一次 `get` 很多 `page` 的 `url`，因此使用 `fake_useragent` 裡的 `UserAgent()`，每次皆產生不同的 `header`，防止被網站 `ban` 掉，使用 `BeautifulSoup` 抓取及解析網頁資料，再利用 `find_all` 抓取所有的圖片 `url`，儲存在 `img_list`，每次執行完皆暫停 `0.5` 秒，每張圖片皆創建一條新的 `thread` 來執行下載的動作。

```

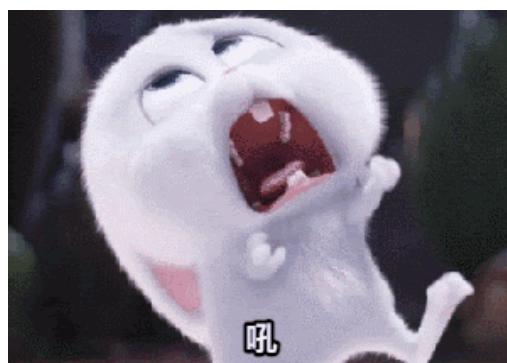
45 def download(picture_url):
46     if not os.path.isdir('images'):
47         os.mkdir('images')
48     spilt = picture_url.split('/')
49     filename = spilt.pop()
50     print('%s is downloading.' % filename)
51     path = os.path.join('images', filename)
52     urlretrieve(picture_url, filename=path)

```

【download】

將圖片 url 的最後一部份網址當作圖片名稱，利用 urlretrieve 將圖片下載下來，存到 images 的資料夾。

(抓到的圖片→)



引入函式庫

函式庫	使用函式
request	request.get()
Bs4	BeautifulSoup
os	os.path.isdir() os.mkdir() os.path.join()
urllib.request	urllib.request.urlretrieve()
threading	threading.Lock() threading.Lock().acquire() threading.Lock().release()
fake_useragent	UserAgent() UserAgent().random