期中作業-爬蟲

概念:

網路爬蟲始於一張被稱作種子的統一資源位址列表。當網路爬蟲存取這些統一資源定位器時，它們會甄別出頁面上所有的超連結，並將它們寫入一張「待訪列表」，即所謂爬行疆域。此疆域上的 URL 將會被按照一套策略迴圈來存取。如果爬蟲在執行的過程中複製歸檔和儲存網站上的資訊，這些檔案通常儲存，使他們可以較容易的被檢視。閱讀和瀏覽他們儲存的網站上並即時更新的資訊，這些被儲存的網頁又被稱為「快照」。越大容量的網頁意味著網路爬蟲只能在給予的時間內下載越少部分的網頁，所以要優先考慮其下載。高變化率意味著網頁可能已經被更新或者被取代。一些伺服器端軟體生成的 URL 也使得網路爬蟲很難避免檢索到重複內容。

組成:

- 指定頁面下載的選擇策略
- 檢測頁面是否改變的重新存取策略
- 定義如何避免網站過度存取的約定性策略
- 如何部署分散式網路爬蟲的並列策略

程式碼:

```
import requests
from bs4 import BeautifulSoup

url = 'https://www.ettoday.net/news/news-list.htm'
r = requests.get(url)

soup = BeautifulSoup(r.text, 'html5lib')
for d in soup.find(class_ = 'part_list_2').find_all('h3'):
#     print(d.find(class_ = 'date').text, d.find_all('a')[-
1].text)
print(d.find(class_ = 'date').text, d.find('a').text)
```

參考資料:

1.

https://zh.wikipedia.org/wiki/%E7%B6%B2%E8%B7%AF%E7%88%AC%E8%9F%B2
2. https://tammy-discovery.com/best-webcrawler-resources/
3. https://www.cupoy.com/event/pycrawler/missions/intro