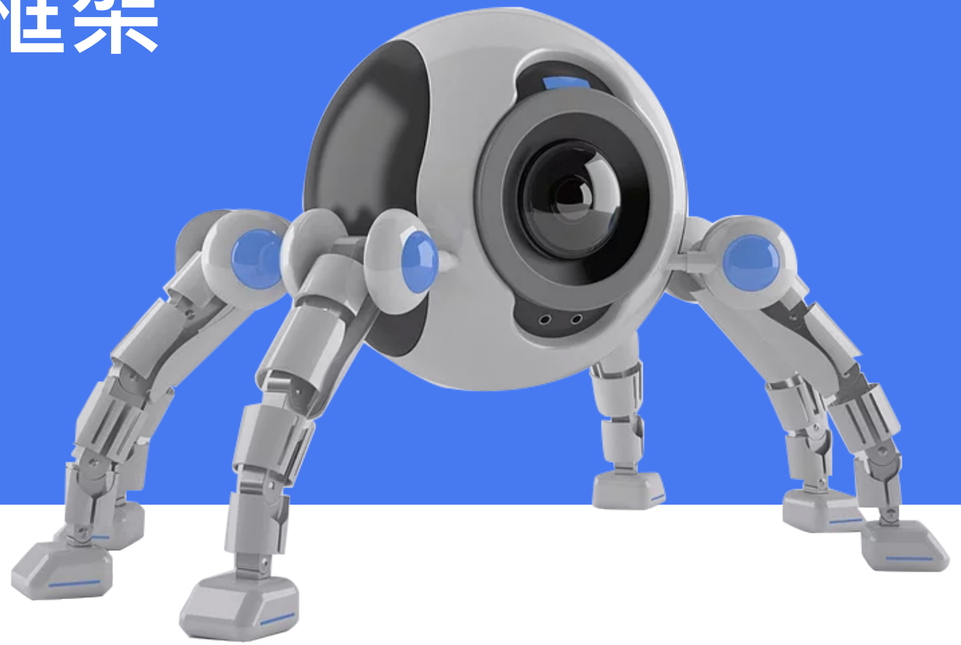
Day 25 SCRAPY 網頁爬蟲框架

多網頁爬蟲實作策略介紹



出題教練：楊鎮銘

python

本日知識點目標

- 了解跨網頁與跨網站的爬蟲概念
- 多網頁爬蟲的注意事項與文件

多網頁爬蟲概念

CUPOY

多網頁爬蟲基本上就是逐一對**網址清單**上的網址爬蟲

而根據**網址清單**型式的不同會有額外的策略

- 自訂清單列表：透過文件或是 List 紀錄目標網頁網址
- HTML 清單列表： 等 tag 紀錄目標網頁網址

圖片來源： 批踢踢實業坊

單一網站多網頁爬蟲概念

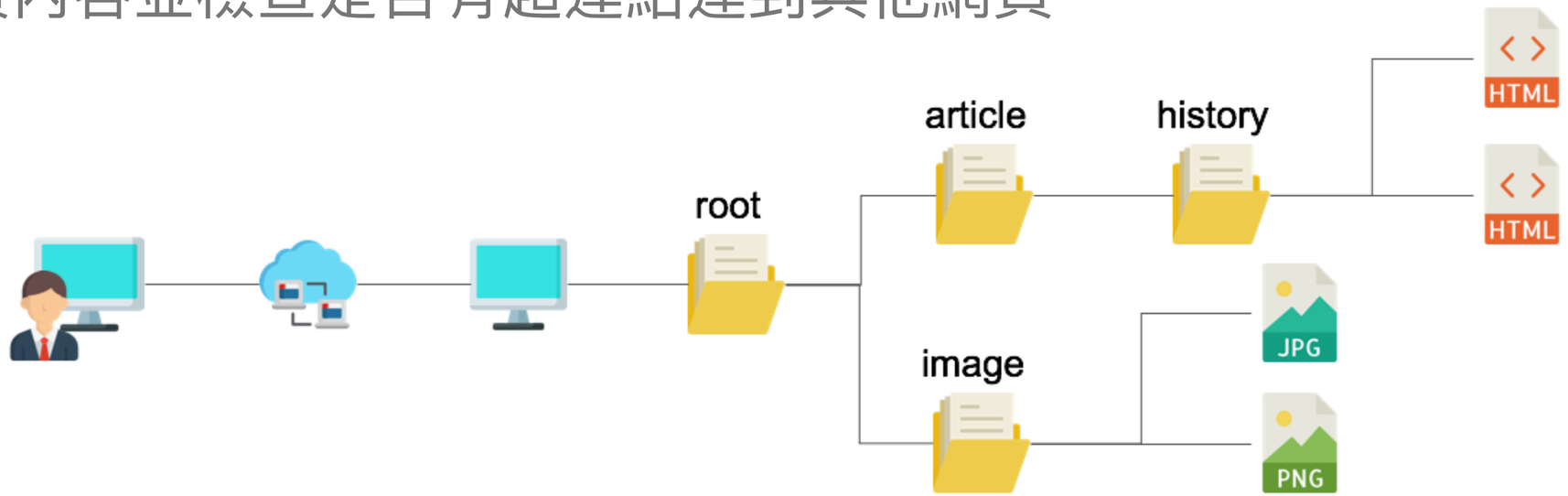
CUPOY

如果是要爬取**單一網站下的多個網頁**，我們不太可能拿到所有目標網頁的清單，此時比較適合的策略是**階層式搜尋網頁並爬取**

1.了解網站文件結構

2.從網站首頁逐一根據超連結 (e.g. <a> tag) 找到其他網頁網址

3.爬取網頁內容並檢查是否有超連結連到其他網頁



跨網站多網頁爬蟲概念

CUPOY

概念上跟**多網頁爬蟲**一樣可以列出多個目標網站的網址清單
再根據**單一網站多網頁爬蟲**的概念逐一爬取
過程中有可能超連結會連到其他網站上的網頁

隨著需要搜索的次數愈多，不確定性愈高
需要更謹慎的檢查每次目標網址的合法性

階層式搜尋的注意事項

CUPOY

- **紀錄已搜尋過的網址**，避免重複爬蟲與無窮回圈
- 合法超連結格式為**絕對路徑**
  - 相對路徑建議可以透過 urllib.parse.urljoin 轉換

建議在處理網址問題時要先了解每個片段的意義

```
scheme://netloc/path;params?query#fragment
```

其中比較重要的是判斷網域的 netloc 與路徑的 path

超連結的注意事項

CUPOY

- 超連結可以不是網址格式
  - <a> 可以是其他非網址格式的型式

超連結屬性	範例	備註
絕對路徑	https://www.cupoy.com	**合法網址格式**
相對路徑	/ex1/index1.html	**合法網址格式**
其他 tag	#top	我們應關注網頁而不是 tag
其他協定	mailto://example@gmail.com	**無法對這種格式送 request**
程式碼	javascript:console.log("Hello")	**無法對這種格式送 request**

網址網域的注意事項

CUPOY

- 超連結網址可以是任何網路位置
  - 網域建議可以透過 urllib.parse.urlparse 判斷
  - 子網域建議可以透過 tldextract.extract 判斷

根據需求有時候需要更詳細的判斷

	netloc
台大首頁	www.ntu.edu.tw
台大中文系首頁	www.cl.ntu.edu.tw

可以再細分為：子網域 + 網域 + 後綴

網域都是 ntu 但是子網域不同

爬蟲的禮貌運動

CUPOY

網站擁有者有時候會**限制爬蟲行為** (e.g. 搜尋引擎的爬蟲，可以爬全網站網頁；一般爬蟲，只能爬首頁的內容)

這些規則通常會放在首頁底下的 **robots.txt**

e.g. https://www.facebook.com/robots.txt

```
# Notice: Crawling Facebook is prohibited unless you have express written
# permission. See: http://www.facebook.com/apps/site_scraping_tos_terms.php

User-agent: Applebot
Disallow: /ajax/
Disallow: /album.php
Disallow: /checkpoint/
Disallow: /contact_importer/
Disallow: /feeds/
Disallow: /file_download.php
Disallow: /hashtag/
Disallow: /l.php
Disallow: /live/
Disallow: /moments_app/
Disallow: /p.php
Disallow: /photo.php
Disallow: /photos.php
Disallow: /sharer/
```

建議開發者根據這些不允許存取的路徑，讓爬蟲直接忽略

可以簡單把 Disallow 的路徑列一個名單，或是參考 google/robotstxt

重要知識點複習

CUPOY

- 根據取得網址清單的方式與是否跨網站可以分為
  - 多網頁爬蟲
  - 單一網站多網頁爬蟲
  - 跨網站多網頁爬蟲
- **網址是有意義的片段組合**，建議根據套件去找出合法的網址
- 爬蟲前參照 robots.txt 去設計，不要造成惡意爬蟲

延伸閱讀

CUPOY

- URL wkik
  - 中文版的 wiki 詳細講解網址 URL 的文法與其意義
- W3C 超連結屬性
  - 本篇有提到 <a> tag 超連結的注意事項，並非所有值都適合送請求做爬蟲，W3C 文件中有定義及範例可以幫助理解
- 關於 robots.txt
  - Google 對於 robots.txt 的解釋，包含用途與限制
- google/robotstxt
  - Google 在 2019.07 開放 robots.txt 的解析器，該程式一直被用於 Google engine 服務，後來甚至發展成 Robots Exclusion Protocol (REP) 標準

解題時間
LET'S CRACK IT

請跳出 PDF 至官網 Sample Code & 作業

完成本日知識點

START