網路爬蟲

網路爬蟲 / 網路蜘蛛

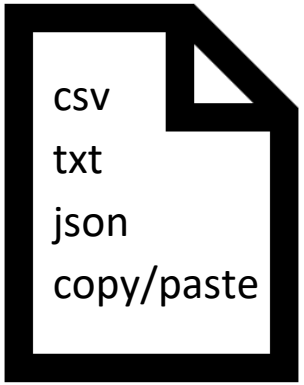
藉由網頁鏈接（URL / API）對伺服器進行 HTTP 請求，並解析其回傳內容，以獲取所需資料。

執行流程 - 用戶請求流程

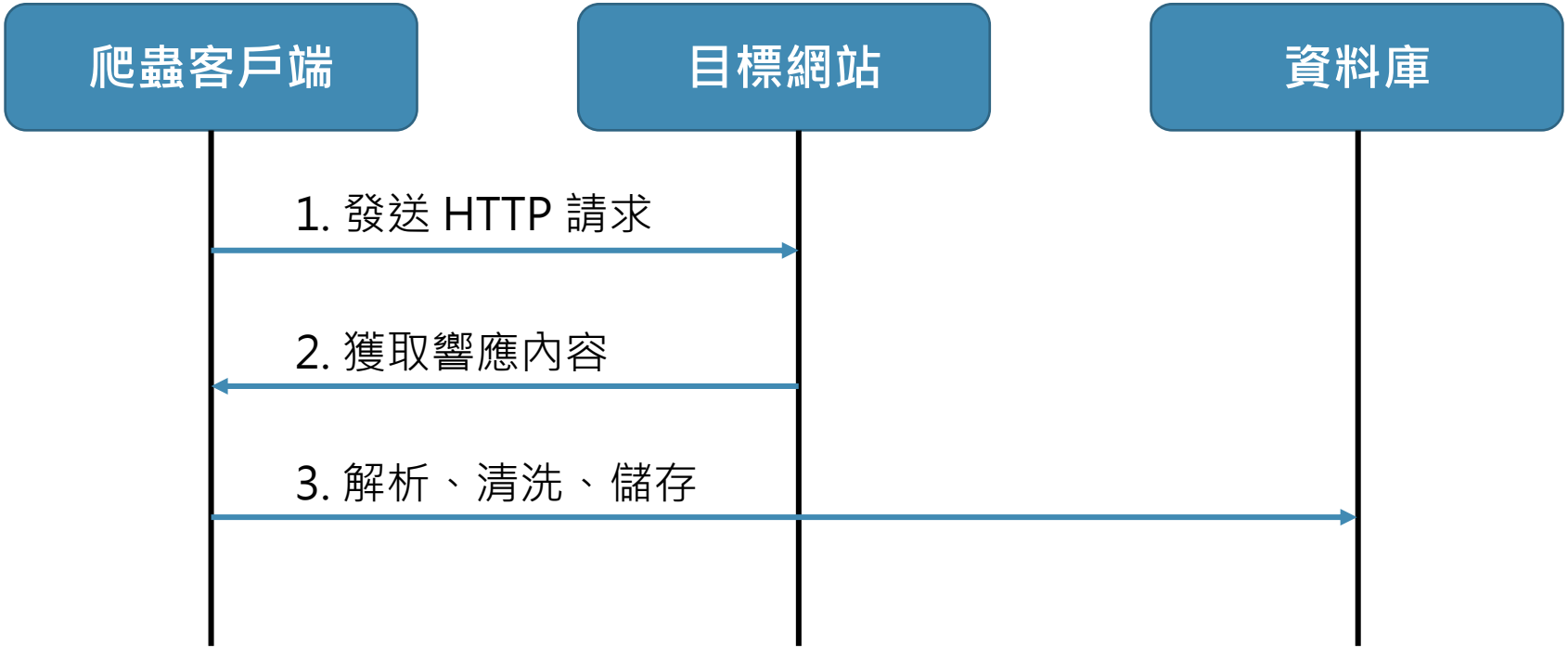
登入網頁

搜尋資料

下載資料



執行流程 - 程式請求流程



程式請求流程

第 1 階段 (程式請求流程) – 我的資料在哪?

- 哪個網站有我想要的資料？
- 網站是否需要登錄或身份驗證？
- 內容品質是否可靠？
- 網站內容的更新頻率為何？
- IPO 與資料結構設計

在這個資訊大爆炸的時代，若進行網路爬蟲時，須嚴謹考量目標網站之內容重複程度及完整性，以利於提升分析效率。

第 1 階段 (程式請求流程) – 我的資料在哪?

- 哪個網站有我想要的資料？
- 網站是否需要登錄或身份驗證？
- 內容品質是否可靠？
- 網站內容的更新頻率為何？
- IPO 與資料結構設計

在進行爬蟲時，除了了解目標網站外，須確保在獲取資料前，是否進行登錄行為或任何身份驗證行為，其行為可能為：

- reCAPTCHA
- CAPTCHA
- Session
- Cookies 等

由於上述行為將提升爬蟲難度，因此在撰寫爬蟲前，須將該因素納入考量。

第 1 階段 (程式請求流程) – 我的資料在哪?

- 哪個網站有我想要的資料？
- 網站是否需要登錄或身份驗證？
- 內容品質是否可靠？
- 網站內容的更新頻率為何？
- IPO 與資料結構設計

在確認目標網站後，需確保爬取內容之品質具可靠性。反之，資料複雜高或來源不明的網站將造成分析結果不準確。

第 1 階段 (程式請求流程) – 我的資料在哪?

- 哪個網站有我想要的資料？
- 網站是否需要登錄或身份驗證？
- 內容品質是否可靠？
- 網站內容的更新頻率為何？
- IPO 與資料結構設計

在了解資料可靠性及目標網站後，對於即時性資料需要掌握其更新頻率，除了能夠掌握資料的流動性，還有助於自動化的機制設定。

第 1 階段 (程式請求流程) – 我的資料在哪?

- 哪個網站有我想要的資料？
- 網站是否需要登錄或身份驗證？
- 內容品質是否可靠？
- 網站內容的更新頻率為何？
- IPO 與資料結構設計

Input-Process-Output Model 的應用有助於建立更完整的程式導向流程及程式碼本身的再利用性及易讀性。

在確認規劃儲存資料後，需針對須儲存的資料進行結構設計。除了能有效優化儲存空間，還能夠在完成爬取資料後進行更複雜的操作。

第 2 階段 (程式請求流程) – 尋找消失的 API / URL, 我要請求咯!

URL / API	溝通方式	常用工具

- 你的鏈接不是你要的鏈接？
- 是否有現成套件或 API 提供？
- 是否提供手機版頁面？

在尋找目標網站時，由於使用的 HTTP 溝通方式不同或網頁設計不同，導致搜尋引擎所顯示的網址不完全是目標網站的網址。

因此可以透過使用開發人員工具檢視網頁原始碼來驗證其相關資料的溝通來源及其溝通方式。

第 2 階段 (程式請求流程) – 尋找消失的 API / URL, 我要請求咯!

URL / API	溝通方式	常用工具

- 你的鏈接不是你要的鏈接？
- 是否有現成套件或 API 提供？
- 是否提供手機版頁面？

在爬蟲前，若相關網站有提供 API 或現成套件獲取資料，在限制條件允許下，則可不考慮使用爬蟲形式獲取資料。

第 2 階段 (程式請求流程) – 尋找消失的 API / URL, 我要請求咯!

URL / API	溝通方式	常用工具

- 你的鏈接不是你要的鏈接？
- 是否有現成套件或 API 提供？
- 是否提供手機版頁面？

在檢視 URL 過程中，若原有目標網站在嘗試數次無法破解後，可嘗試尋找該網站之手機版（不是所有網站皆有手機版），並嘗試尋找手機版網站的目標網址。

第 2 階段 (程式請求流程) – 尋找消失的 API / URL, 我要請求咯!

URL / API	溝通方式	常用工具

OPTIONS

GET

HEAD

POST

PUT

DELETE

TRACE

CONNECT

***以上溝通方式（HTTP Method）定義於 HTTP 1.1 版本

第 2 階段 (程式請求流程) – 尋找消失的 API / URL, 我要請求咯!

URL / API	溝通方式	常用工具

GET vs POST

	GET	POST
網址差異	網址會帶有 HTML Form 表單的參數與資料。	資料傳遞時，網址並不會改變。
資料傳遞量	由於是透過 URL 帶資料，所以有長度限制。	由於不透過 URL 帶參數，所以不受限於 URL 長度限制。
安全性	表單參數與填寫內容可在 URL 看到。	透過 HTTP Request 方式，故參數與填寫內容不會顯示於 URL。

第 2 階段 (程式請求流程) – 尋找消失的 API / URL, 我要請求咯!

URL / API	溝通方式	常用工具

GET vs

```
GET /?id=010101 HTTP/1.1
Host: xxx.toright.com
User-Agent: Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-TW; rv:1.9.2.13) Gecko
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: zh-tw,en-us;q=0.7,en;q=0.3
Accept-Encoding: gzip,deflate
Accept-Charset: UTF-8,*
Keep-Alive: 115
Connection: keep-alive
```

第 2 階段 (程式請求流程) – 尋找消失的 API / URL, 我要請求咯!

URL / API	溝通方式	常用工具

vs POST

```
POST / HTTP/1.1
Host: xxx.toright.com
User-Agent: Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-TW; rv:1.9.2.13) Gecko
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: zh-tw,en-us;q=0.7,en;q=0.3
Accept-Encoding: gzip,deflate
Accept-Charset: UTF-8,*
Keep-Alive: 115
Connection: keep-alive

Content-Type: application/x-www-form-urlencoded
</code><code>Content-Length: 9
id=010101
```

第 2 階段 (程式請求流程) – 尋找消失的 API / URL, 我要請求咯!

URL / API	溝通方式	常用工具

HTTP 狀態碼 (Status Codes)

狀態碼	說明 (英)	說明 (中)
200	OK	成功
301	Moved Permanently	目標網頁移到新網址(永久轉址)
302	Found(Moved Temporarily)	暫時轉址
304	Not Modified	已讀取過的圖片或網頁，由瀏覽器緩存中讀取
400	Bad Request	伺服器無法處理這個 Request
401	Unauthorized	需身分驗證，如 SSL key or htaccess pasword

第 2 階段 (程式請求流程) – 尋找消失的 API / URL, 我要請求咯!

URL / API	溝通方式	常用工具

HTTP 狀態碼 (Status Codes)

狀態碼	說明 (英)	說明 (中)
403	Forbidden	無讀取權限，可能是 IP 被阻檔或是伺服器限制
404	Not Found	伺服器未找到目標網址，檔案不存在
500	Internal Server Error	伺服器發生錯誤 : 可能是 htaccess 有錯
502	Bad Gateway	伺服器的某個服務沒有正確執行
503	Service Unavailable	伺服器當掉
504	Gateway Timeout	伺服器上的服務沒有回應

第 2 階段 (程式請求流程) – 尋找消失的 API / URL, 我要請求咯!

URL / API	溝通方式	常用工具

回傳資料格式

HTML XML JSON 特定項目格式

第 2 階段 (程式請求流程) – 尋找消失的 API / URL, 我要請求咯!

URL / API	溝通方式	常用工具

程式請求

urllib3 / requests

模擬請求

selenium

第 3 階段 (程式請求流程) – 是的，我要斷開一切的牽連~


```
[{'title': '最高法院 109 年度 台上 字第 1719 號刑事判決（29K）',
  'date': '109.04.15',
  'judge': '違反證券交易法等罪',
  'content': '最高法院刑事判決\u3000\u3000\u3000\u3000\u3000\u3000\u3000\u3000\u3000 108年度台上字第4056號\n\u3000\u3000\u3000\u3000\u3000\u
3000\u3000\u3000\u3000\u3000\u3000\u3000\u3000\u3000\u3000\u3000 109年度台上字第1719號\n上\u3000訴\u3000人\u3000臺灣高等檢察署檢察官黃
騰耀\n上\u3000訴\u3000人\u3000\n（被\u3000告）\u3000王宗立\n選任辯護人\u3000張簡勵如律師\n\u3000\u3000\u3000\u3000\u3000\u3000林宗憲律師\n上\u
3000訴\u3000人\u3000王安石（原名王宗貞）\n（被\u3000告）\u3000\u3000\u3000\u3000\n原審辯護人\u3000王寶菈律師\n上\u3000訴\u3000人\u3000\n（被\u3
000告）\u3000劉玉增\n選任辯護人\u3000丁中原律師\n\u3000\u3000\u3000\u3000\u3000\u3000林俊吉律師\n被\u3000\u3000\u3000告\u3000楊善博（原名楊偉
祥）\n\u3000\u3000\u3000\u3000\u3000\u3000林憶慧（原名林月桂）\n選任辯護人\u3000王福民律師\n上列上訴人等因被告等違反證券交易法等罪案件，不服臺灣高
等\n法院中華民國108年8月1日第二審判決（106年度金上重訴字第11\n號、106 年度上訴字第1052號，起訴案號：臺灣臺北地方檢察署\n90年度偵字第21111、23982
號，91年度偵字第9844、10256、25\n181號，92年度偵字第1501、20496、24888號，93年度偵字第500\n、10184、10344、13340號，94年度偵字第840、841、842、7
661\n、14564 號；追加起訴案號：105年度蒞追字第4號），提起上訴\n（王安石部分由其原審辯護人代為提起上訴），本院判決如下：\n    主   文\n原判決關於王宗
立犯其附表一㈠編號一、四部分；王安石犯其附\n表一㈡編號一部分；及劉玉增有罪部分，均撤銷，發回臺灣高等\n法院。\n其他上訴駁回。\n    理   由\n壹、撤銷發
回部分\n一、本件原判決認定上訴人即被告王宗立有其事實欄（下稱事實\n    欄）一、二㈠至㈢如其附表（下稱附表）一㈠編號一及事實\n    欄六如附表一㈠編號四
犯行；上訴人即被告王安石有事實欄\n    一、二㈡如附表一㈡編號一犯行；上訴人即被告劉玉增有事\n    實欄二㈡犯行，因而撤銷第一審對其等3 人關於此部分之科
\n    刑判決，經比較新舊法規定後，改判就王宗立附表一㈠編號\n    一依修正前刑法想像競合犯規定從一重論處王宗立如附表一\n    ㈠編號一所示修正前證券交易
法第179條、第171條第1項第1\n    款共同犯法人行為負責人詐偽出售有價證券罪刑、就附表一\n    ㈠編號四依集合犯論處王宗立如附表一㈠編號四所示修正前\n
銀行法第125條第1項前段、第3 項法人行為負責人，犯銀行\n    法第125條第1項前段之非法經營收受存款業務罪刑；就王安\n    石附表一㈡編號一依修正前刑法想像
競合犯規定從一重論處\n    王安石如附表一㈡編號一所示修正前證券交易法第179 條、\n    第171條第1項第1 款共同犯法人行為負責人詐偽出售有價證\n    券罪
刑；就劉玉增依修正前刑法想像競合犯規定從一重論處\n    修正前證券交易法第179條、第171條第1項第1款共同犯法人\n    行為負責人詐偽出售有價證券罪刑，及均
諭知相關沒收。固\n    非無見。\n二、惟按：\n    ㈠證券交易法制訂後歷經多次修正，其中於民國89年7 月19日\n    修正前之證券交易法第2條、第6條規定，有價證
券之募集、\n    發行、買賣，其管理、監督依證券交易法之規定，所稱有價\n    證券，謂政府債券及「公開募集、發行」之公司股票、公司\n    債券及經財政部核
定之其他有價證券。但89年7 月19日修正\n    公布施行後之證券交易法第6 條則規定，該法所稱之有價證\n    券，謂政府債券、公司股票、公司債券及經財政部核定
之其\n    他有價證券，已將該法條「有價證券」之定義，刪除「公開\n    募集、發行」等文字。二者間有所不同，即證券交易法修正\n    前「有價證券」所稱之公
司股票，限於「公開募集、發行之\n    公司股票」，與修正後「有價證券」所稱之公司股票未加限\n    制不同，依其修正規範目的，在於杜絕未公開募集、發行之\n
公司股票的買賣行為，俾免經濟交易秩序目的失衡；又公開\n    募集、發行之公司股票，係指依公司法第133條、第268條及\n    證券交易法相關規定申請公開募集、
發行之股票。是基於罪\n    刑法定之溯及既往之禁止及明確性原則，89年7 月19日修正\n    公布前證券交易法所規範之對象，自不包括未公開發行之公\n    司股
票。本件原判決於事實欄二㈡記載「王宗立、王安石及\n    劉玉增（下稱王宗立等3 人）共同基於非法以詐偽出售有價\n    證券之犯意聯絡.. 使附表三所示之人陷於
錯誤，購買『 富\n    貴專案』，或進而購買台基開發投資股份有限公司（下稱台\n    基開發公司）股票，以詐偽方式向不特定人販售專案或台基\n    開發公司股票
（購買專案之被害人、時間、單位、被害金額\n    如附表三各編號所示），合計詐得新臺幣（下同）2,309萬5\n    ,995 元。」等情（見原判決第6至7頁）。於理由
欄載敘「.\n    . 如附表三編號6、10、11至19『證據欄』所示之證據.....\n    足認此部分被害人除購買「富貴專案」外，亦購買台基開發\n    公司股票。... 」
（見原判決第52、53頁），並說明：「..\n    . 此觀諸證券交易法第6條第1項原規定：『本法所稱有價證\n    券，謂政府債券及公開募集、發行之公司股票、公司債
券及\n    經財政部核定之其他有價證券。』89年7 月修法時，曾有提\n    案將公開募集、發行等語刪除，理由係本條乃屬對證券定義\n    之條文，公開募集、發行
字樣實與證券之定義無關，而係是\n    否屬豁免交易時應考慮之問題，而此觀之證券交易法第7 條\n    、第22條即可明之，爰將此贅文刪除。亦即，證券交易法所\n
規範之公司股票原以公開募集發行者為限，有關證券詐欺之\n    民刑事責任，原僅適用於公開募集發行之有價證券。修正之\n    後，刪除「公開募集、發行」等語，則
即使行為人以未公開\n    發行的證券為詐欺買賣的工具，仍受證券交易法的規範。..\n    ... 」（見原判決第53至54頁），復於附表三富貴專案記載\n    編號10施
盈州購買台基開發公司股票時間為88年間、編號11\n    魏開智為89年1月18日、編號12陳世斌為89年3月上旬、編號\n    13曾慶中為89年3月10日、編號18洪瑞隆 (1)
為89年5月間、\n    編號53李瑞芳為89年5月等情（見原判決第281至285、296頁\n    ）。上情倘若無訛，原判決似亦認89年7 月19日修正公布前\n    證券交易法所
規範之對象，並不包括未公開發行之公司股票\n    ，則附表三編號10、11、12、13、18(1)及53，王宗立等3人\n    銷售台基開發公司股票之時間均在89年7 月19日
修正公布施\n    行之前，能否以89年7 月19日修正公布施行之證卷交易法相\n    關規定處罰之，即值研求。原判決未予究明上情，亦未為必\n    要之釐清及說明，
```


	年度	字號	案號	類型	裁判日期	裁判案由	金額
0	109	台上	1719	刑事判決	109.04.15	違反證券交易法等罪	61785382
1	108	台上	4056	刑事判決	109.04.15	違反證券交易法等罪	573105930
2	108	台上	16	刑事判決	109.01.16	違反證券交易法等罪	372933046
3	109	台抗	46	刑事裁定	109.01.09	違反證券交易法等罪不服再執行羈押	898763655
4	107	台上	846	民事判決	108.03.27	請求損害賠償	75615095

資料解析：清整、提取的過程

第 3 階段 (程式請求流程) - 是的，我要斷開一切的牽連~

資料解析工具

工具語言	CSS 選擇器	路徑語言
BeautifulSoup / PyQuery / JSON / Scrapy	selenium	selenium / lxml

第 3 階段 (程式請求流程) - 是的，我要斷開一切的牽連~

資料儲存

Excel 儲存

pandas / built-in write function

資料庫 儲存

Mysql-connector / API

TXT 儲存

built-in write function

第 4 階段 (程式請求流程) – 不是完了嗎? 怎麼還要繼續呀!

優化程式

完成一個爬取的流程僅僅只是最小可行性（MVP）的實現，
我們還需按照對於資料的需求進行一系列的規劃與設計：

- 模組化程式設計（Design Pattern）
- 定時定期的爬取觸發機制（Cron job）
- 異常處理設計（Error Handling）
- 。。。

其它

爬蟲的禮貌

1. 不要太快進行下個請求，請善用休息函式（time.sleep(休息秒數)）

2. 不要把不需要的資料也爬下來（若請求允許彈性查詢）

3. 不要違反 Robots 協議和網站規則，請在請求時附上該有的參數

爬蟲的應用場域

項目預訂　交易策略　資料分析　應用程序

法律看爬蟲!

- 你的爬蟲會送你進監獄嗎？

https://www.mdeditor.tw/pl/2qTa/zh-tw

- 靠挖掘別人家的資料數據來賺錢，「網路爬蟲」這個行為合法嗎？

https://www.techbang.com/posts/75284-is-the-internet-crawler-legal-china-and-the-united-states-have-different-views#top

實作時間