網頁爬蟲教學

ronnywang @ HTC 2016/6/13, 14

關於 Ronny Wang

http://ronny.tw/data

ronnywang@gmail.com

- 曾任痞客邦產品開發副理
- 現為李慕約公司共同創辦人
- g0v 零時政府新聞小幫手、求職小幫手、開放政治獻金等專案發起人

我的爬蟲作品

- PTT 人氣 http://ptthot.ronny.tw/
- 公司資料 http://company.g0v.ronny.tw/
- newsdiff http://newsdiff.g0v.ronny.tw/
- 中華民國內閣記錄 https://ronnywang.github.io/taiwan-cabinet/
- 每日四大報 http://oldpaper.g0v.ronny.tw/
- 關貿進出口資料 http://portal.g0v.ronny.tw/
- …… http://ronny.tw/data or https://github.com/ronnywang

一個完整的爬蟲包含...

1. 如何跟伺服器要到 HTML
2. 要到並解析資料的列表
3. 要到並解析各別資料的結構性資料

如何跟伺服器要到 HTML

- HTTP GET / POST
- cookie / session
- captcha 驗證碼
- referer / user agent
- 很溫柔，讓對方沒有感覺

要到並解析資料的列表

- 找出完整的列表
  - 一次性爬蟲
- 找出有更新的列表
  - 持續性爬蟲
- 利用搜尋功能
- 利用 API
- 從其他外部集合

要到並解析各別資料的結構性資料

- HTML DOM parser
- Regular Expression

結構性資料

JSON/XML

- 結構彈性較大，可以有樹狀巢狀結構
- 各程式語言都滿好處理的
- 較肥大
- 需要先了解其結構才方便處理
- 編碼固定為 UTF-8

CSV

- 只支援表格結構資料
- 編碼不固定
- 方便直接給 Excel, R 或是各統計軟體使用
- 所佔空間較小

About HTML

Markup LAnguage

```
<同學名單>
  <同學 座號="01" 姓名="王小明">
    <家長 關係="父">王大明</家長>
    <家長 關係="母">林大美</家長>
  </同學>
  <同學 座號="02" 姓名="吳小華">
    <家長 關係="父">吳大中</家長>
    <家長 關係="母">李大蓮</家長>
  </同學>
</同學名單>
```

Tag: <同學>

Attribute: 座號="01"

HTML : HyperTEXT Markup Language

- What is HyperText?
- 資料常出沒 tag

,   用在區塊

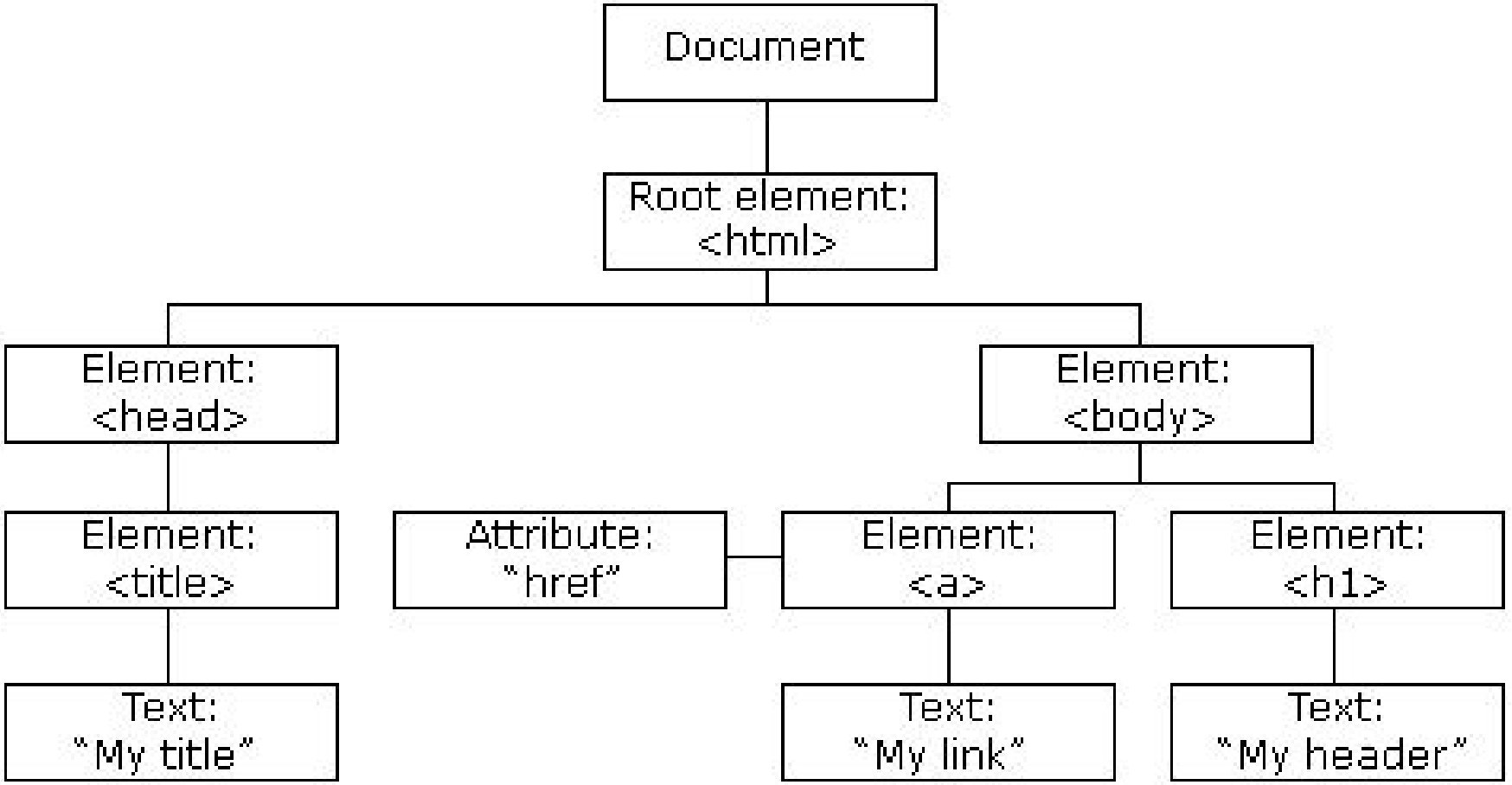
<table>
  <tr> <td>col1</td><td>col2</td> </tr>
  <tr> <td>1</td><td>2</td> </tr>
</table> 用在表格

 val1 val2

 val1 val2 用在列表

bar 用在超連結

HTML DOM Document Object Model



準備來用 PHP 拆解 HTML 吧

PHP 語法與 Java, C 不同事項

- PHP 程式部分是在 <?php … ?> 的之間
  - 如果是一個純粹的 PHP 的話，就檔案以 <?php 開頭，後面沒有 ?> 也沒關係
  - 寫爬蟲通常都是純粹的 PHP
- PHP 的變數一定是以 $ 開頭 (Ex: $a = $a + 1;)
- PHP 變數不需要宣告型別和初始化(建議不要預期他的預設值)
- PHP 有 array 和 object
- array
  - 初始化 $array = array();
  - list array (array key 是照順序的數字)
    - $array[] = $row; // 插入新的 $row 到 $array 中
    - count($array) // 回傳 $array 的大小
  - assoicate array (key 是任意字串，會以 hash 的型式)
    - $array[$key] = $row; // 把 $row 塞進 hash key = $key 的 hash 內
  - associate array 也可以用 $array[] 當 list 用，但不建議混用

PHP 語法與 Java, C 不同事項 (Cont.)

- Object
  - 初始化 $obj = new [SomeClass]; // 把 $obj 宣告成某個型別
  - 初始化 $obj = new StdClass; // 把 $obj 宣告成標準物件
  - $obj->key = 'value'; // 把物件 $obj 塞入 key='key' 的值
  - associate array 跟 stdclass 很像...
    - 一個是 $array['key'] = $value; 一個是 $obj->key = $value
  - $obj->func() // 可以執行 $obj 物件的 func() method
  - PHP Object 沒有 $obj.func() 用法，只有 $obj->func();
- Associate Array 和 object 都可以被 json_encode($obj) or json_encode ($array) 成 object JSON
- list Array 可以用 foreach ($array as $row) { … } 跑每個值
  - 等於 for ($i = 0; $i < count($array); $i ++) { $row = $array[$i]; … }
- associate array 和 object 可以用 foreach ($obj as $key => $value) { … } 跑每個物件
- PHP 沒有 import 或是 #include，PHP 的 extension 裝了之後就直接可以使用

PHP DOM function

- $doc = new DOMDocument; // 產生空的 DOM 物件
- $doc->loadHTML($html); // 載入 HTML進 $doc
- $div_doms = $node->getElementsByTagName('div'); // 取得 $node 下的所有 DIV
  - $div_doms->length // 取得 $div_doms 有幾個物件
  - $div_dom_3 = $div_doms->item(3); // 取得 $div_doms 的第4個物件(from 0)
  - foreach ($div_doms as $div_dom) { ... } // foreach $div_doms 每一個個別的 DOM

PHP DOM function

- $title_dom = $node->getElementById('title'); // 取得 $node 下面的 id="title" 的 dom
- $node->getAttribute('href'); // 取得 $node 的 href="xxx" 的值
- $node->nodeValue; // 取得 $node 裡面的值
- $node->childNodes // 取得 $node 下面包含哪些子 DOM
- $node->nextSibling; // 取得 $node 的下一個兄弟
- $node->nodeName // 取得這個 node 是甚麼種類, Ex: <a> => 'a', => 'div' 或是 #text 是純文字
- $doc->saveHTML($node) // 可以回傳 HTML

練習1 從 HTML 取出資料

檔名 : 1.php

PTT 人氣: https://www.ptt.cc/hotboard.html

輸出 [英文板名],[數字人氣],[中文板名] 的 CSV

- 輸出 CSV
  - $output = fopen('php://output', 'w');
  - fputcsv($output, array(v1, v2, v3 …));
- $html = file_get_contents('網址'); 可以把網址的內容會傳給 $html 變數
  - 建議可以先用 curl 指令把 HTML 抓下來，在開發時用靜態檔案，等到開發完成再改成用線上網址
  - curl https://www.ptt.cc/hotboard.html > ptt.html
- 網頁上看到第一個是 Gossiping，從原始檔找找看 Gossiping 在哪裡
- 除了用 foreach ($td_doms as $td_dom) { … } 以外，可以用 $td_doms->item ($i) 取得某個特定的 <td> DOM
- explode("a:b:c", ":") => [a,b,c]

練習2 斧頭幫 Level 1

檔名：2.php

http://axe.g0v.tw/ 練習看看吧

輸出: 如網頁要求的 JSON 格式

- 用 echo json_encode($array) 輸出結果
  - $array = array()
  - $array[] = array('name' => 'Ronny', 'grades' => array('國語' => 90, '英語' => 30'));
  - echo json_encode($array);
- 數字記得加上 intval($str) 確保他變成數字型別
- 中文字變成「\u9673\u653f\u61b2」沒關係，這是合法的，如果真的有潔癖想要看到正確中文字，可以用 json_encode($array, JSON_UNESCAPED_UNICODE)

練習3 抓 PTT 的推文

檔名 : 3.php

https://www.ptt.cc/bbs/MobileComm/M.1465283968.A.EA3.html

輸出 [推or→or噓],[ID],[說的內容],[時間] 的 CSV

- $class = $dom->getAttribute('class') # 取得 class="xxx" 的值
- $classes = explode(' ', $class); // 如果有多個 class 以空格分開，把他變成陣列
- if (in_array('str', $classes)) // 可以檢查 str 是否在 $classes array 中
- trim($str) 可以清除字串前後的空白或跳行

練習 - 選些題目來練吧

- 從 https://www.ptt.cc/bbs/mobilecomm/index.html 抓這一頁的文章列表
  - 檔名 4-1.php
  - 輸出 [文章網址],[文章標題],[帳號],[時間] 的 CSV
  - 如果整個 ... 內很確定只有一個 <a> tag，可以直接用 $div_dom->getElementsTagName('a')->item(0) 把他抓出來
- 從 http://www.mobile01.com/category.php?id=4 抓出最新文章列表
  - 檔名 4-2.php
  - 輸出 [新聞網址],[新聞手機種類],[新聞標題] 的 CSV
  - 如果有發現資料在 id="foo" 區塊內就可以直接用 $doc->getElementById('foo') 取得該 dom，比 class 快超多
  - mobile01 有做簡單的擋機器人，因此直接 curl 或是 file_get_contents 會抓不到資料
    - 可以用「curl --user-agent 'Chrome' 'http://www.mobile01.com/category.php?id=4' > 4-2.html」，讓 mobile01 以為這是來自 chrome 瀏覽器...
    - 後面會再教到程式中怎麼處理

有些網站
沒那麼好
直接抓...

以 PTT 八卦板為例...

https://www.ptt.cc/bbs/Gossiping/index.html

第一次連入會問是否滿 18 歲...

不能直接用 file_get_contents 了

HTTP 簡介

- HyperText Transfer Protocol
- Server
  - Apache, nginx, IIS …
- Client
  - Chrome, Firefox, IE, Safari, curl ...
- Protocol
  - REQUEST: Client 對 Server 送出 Method + 網址 以及 request header (或者有些 method 可能會有 request body)
  - RESPONSE: Server 回傳該網址應該回應的內容，包含 response code 、response header 和 response body

HTTP 簡介

例如瀏覽器打開 http://ronny.tw/index.html?name=ronny&value=blabla

1. 瀏覽器連上 ronny.tw port 80
2. 瀏覽器送出 GET /index.html?name=ronny&value=blabla 的 request 並加上 header (例如宣稱自己是 Chrome 瀏覽器, 支援哪些語言...)
3. ronny.tw server 回傳結果的 200 OK, header 和 body

可以用 Chrome 開發者工具來看看

HTTP 簡介

Request Method

- GET - 只透過網址取得內容
- POST - 除網址以外，還可以額外讓 client 送多點資訊給 server

Response Code

- 2xx - 一切正常，給你內容
  - 200 OK
- 3xx - 一切正常，不過沒內容可給你
  - 301 東西永久搬到其他地方了
  - 302 東西暫時搬到其他地方了
  - 304 內容跟你上次讀時沒變，不需要再給你了
- 4xx - 不正常，出在客戶端身上的問題
  - 403 你要看的網址 你沒權限看
  - 404 你要看的網址東西不存在
- 5xx - 不正常，出在伺服器端身上的問題
  - 500 Server 出問題了

HTTP 簡介

Request Header

- Cookie - 之前 Server 透過 Set-Cookie 存下來的東西
- User-Agent - 宣稱自己是什麼客戶端
- Referer - 宣稱自己是從哪個網頁過來的

Response Header

- Set-Cookie - 告訴 Client 之後 request 時給我這個 cookie
- Content-Type - 回傳的內容是什麼格式的文件

curl library

```
$curl = curl_init($url);
curl_setopt($curl, CURLOPT_RETURNTRANSFER, true);
$contnet = curl_exec($curl);
curl_close($curl);
```

等同於

```
$content = file_get_contents($url);
```

所以遇到八卦板的 case 怎麼辦

1. 把「已經按下滿18歲」的 cookie 複製到程式中
2. 在程式端實作「我按下我已滿18歲」的動作

複製 cookie 法

方法:

- 透過 Chrome 開發者工具將已經成功可以讀到內容的 cookie 複製下來
- curl_setopt($curl, CURLOPT_HTTPHEADER, array("Cookie:xxx"));

使用情況:

1. 比較省事
2. 這個狀況只能以人工做到, 難以用程式做到時(Ex: 有驗證碼)
3. 狀況要可以被複製

練習 - cookie 複製法

抓出 https://www.ptt.cc/bbs/Gossiping/M.1465540420.A.CA6.html 推文列表

檔名: 5.php

輸出 [推or→or噓],[ID],[說的內容],[時間] 的 CSV

把剛剛前面 3.php 改寫 (cp 3.php 5.php)

1. 先把 file_get_contents 改寫成 curl 用法
2. 從 Chrome 開發者工具取得 cookie 現值
3. curl_setopt($curl, CURLOPT_COOKIE, 'xxx'); 貼進來

實作多步驟

方法

- curl 本身會保存 cookie，所以只要 curl_init 一次取得 $curl 物件，然後把每個動作做進去
- 先 POST 送出滿十八歲
- 再用同一個 $curl 去要資料看看

使用情境

- 多步驟比較複雜，不能直接複製 cookie 的情況

練習 抓PTT改用多步驟法

抓出 https://www.ptt.cc/bbs/Gossiping/M.1465540420.A.CA6.html 推文列表

檔名: 6.php

輸出 [推or→or噓],[ID],[說的內容],[時間] 的 CSV

把剛剛前面 3.php 改寫 (cp 3.php 6.php)

1. 先把 file_get_contents 改寫成 curl 用法
2. 需要加上 curl_setopt($curl, CURLOPT_COOKIEFILE, ""); ，這樣之後的 $curl 都會延續之前的 session
3. 在抓資料之前先做出 POST 送出滿十八歲的動作
   a. 用 Chrome 開發者工具，找出滿十八 歲是對哪個網址以及送了什麼 內容
4. curl_setopt($curl, CURLOPT_POSTFIELDS, 'aaa=bbb&ccc=ddd');
5. curl_setopt($curl, CURLOPT_URL, ' 新的網址');
6. 沿用 $curl 物件，來抓資料看看

如何跟伺服器要到 HTML

- HTTP GET / POST
- cookie / session
- captcha 驗證碼
- referer / user agent
- 很溫柔，讓對方沒有感覺

要到並解析資料的列表

- 找出完整的列表
  - 一次性爬蟲
- 找出有更新的列表
  - 持續性爬蟲
- 利用搜尋功能
- 利用 API
- 從其他外部集合

要到並解析各別資料的結構性資料

- HTML DOM parser
- Regular Expression

今天就到
這邊囉!
明天繼續!

練習 - 一次抓很多頁資料

檔案: 7.php

http://axe.g0v.tw/level/2

斧頭幫 Lv2 抓有大量列表

- 把原來的 2.php 改寫一下，加上 for 迴圈
  - (找一下他每一頁的網址有什麼規則，用 for 迴圈把每一頁都跑一次吧)
- 可以在 file_get_contents($url) 之前，加上 error_log($url)
  - 這樣可以在 stderr 看到目前的執行進度，如果爬的頁數太多至少不會心裡不踏實 ...

練習 - 需要使用多步驟

檔案: 8.php

http://axe.g0v.tw/level/3

- 會需要用到一個 $curl 物件重覆使用
  - curl_setopt($curl, CURLOPT_COOKIEFILE, '');
- 把原來的 2.php 改寫一下吧
- 可以先只抓個兩頁就把迴圈給 break 掉，然後人工看看輸出結果是否正確，不正確的話還可以再改程式(以免跑完 76 頁才發現錯了就哭哭)
  - 如果只是要看看的話，可以用 json_encode($obj, JSON_UNESCAPED_UNICODE | JSON_PRETTY_PRINT) 會比較好看一點

有些網站
擋機器人

網站常見擋機器人方式

1. 會認 User Agent，必須是常見瀏覽器才給內容
   a. IE, Firefox, Chrome, Safari ...
2. 會認 Referer，沒給 referer 或是外部來的就不給內容
3. 以驗證碼阻擋
4. 短時間內大量存取就會阻擋

練習 對付會擋機器人的網站

檔案: 9.php

http://axe.g0v.tw/level/4

這邊有用到兩種擋機器人的判斷法

- curl_setopt($curl, CURLOPT_USERAGENT, 'xxx'); // 可以指定 User Agent
- curl_setopt($curl, CURLOPT_REFERER, 'xxx'); // 可以指定 Referer
  - Referer 網址不一定要乖乖的用上一頁的網址，大部分時候 referer 網址用你正要抓的網址就可以了
  - 不過還真的有少部份網站龜毛到真的要不同網址才能 referer……

擋機器人?

- 如果你的程式的行為模式跟一般瀏覽器相同, 誰能擋的了你?
  - 擋了你就等於也擋了正常的人了...

歷史的傷痕
Big5

處理 Big5, UTF-8

- 建議都轉成 UTF-8 處理
- iconv($from, $to, $str);
- DOM 會處理 Big5 轉 UTF-8，但是有些情況可能會失敗
  - 網頁內含有不合法的 Big5 字元
    - 解法：用 iconv 把 HTML 轉成 utf-8，再把 <meta http-equiv="Content-Type" content="text/html; charset=big5"> 改成 utf-8
  - 網頁沒說清楚自己的編碼
    - 解法 $html = str_replace('<head>', '<head><meta http-equiv="Content-Type" content="text/html; charset=big5">', $html); 硬幫他加上去編碼
- 伺服器端只支援 Big5 時，記得 GET 和 POST 參數也要轉成 Big5 再傳

練習 抓 Big5 網站, PTT 人氣

檔名 : 10.php

PTT 人氣: https://www.ptt.cc/hotboard.html

輸出 [英文板名],[數字人氣],[中文板名] 的 CSV

- 從 1.php 改寫 (cp 1.php 10.php)
- 用 iconv('big5', 'utf-8//IGNORE', $str) 把 Big5 轉成 UTF-8
- 用 str_replace('charset=big5', 'charset=utf-8', $str); 把 HTML 宣告編碼改成 UTF-8

很溫柔
讓對方
沒感覺

這是我的溫柔....

- 如果是政府網站的話
  - 有的時候是伺服器端本身效能就不夠，就算開個十台分散式抓資料對方也只有一台可以處理，這種時候還是溫柔點別抓太快吧
- 如果是民間網站的話
  - 刑法360條 無故以電腦程式或其他電磁方式干擾他人電腦或其相關設備，致生損害於公眾或他人者，處三年以下有期徒刑、拘役或科或併科十萬元以下罰金。
- 所以還是溫柔一點吧...

睡吧...

- 每一次 query 前睡個 1 秒鐘吧...
  - sleep(1);

從 POST
拉資料

練習 抓搜尋結果

檔案: 11.php

http://tgos.nat.gov.tw/tgos/Web/Address/TGOS_Address.aspx

寫出一個程式, 可以抓出 $road 變數的門牌列表 (Ex: 臺北市羅斯福路, 臺北市市府路)

1. 實際上去搜尋一次, 看看他對哪個網址送了什麼 POST 內容
2. curl_setopt($curl, CURLOPT_POSTFIELDS, $post);
    a. $params = array();
    b. $params[] = 'name=' . urlencode($name);
    c. $params[] = 'value1=' . urlencode($value1);
    d. $post = implode('&', $params);
3. $params = array();
4. $params['name'] = $name;
5. $params['value1'] = $value1;
6. $post = http_build_query($params);

練習 Big5 + post

檔案 12.php

到 http://jirs.judicial.gov.tw/FJUD/FJUDQRY01_1.aspx 抓出法院名稱為「臺灣臺北地方法院」, 類型為刑事判決, 全文檢索包含「宏達國際」的判決書

- urlencode 之前要把值轉成 big5
  - $params[] = 'key=' . urlencode(iconv('utf-8', 'big5', $key));
- 這個網站有擋機器人..把 9.php 斧頭幫 lv4 的技巧拿來用吧

抓吧抓吧
大量抓吧

問自己想抓到什麼程度?

- 只要一次性就好
- 只要從現在起抓未來資料就好
- 從古至今的資料必需要抓光光
- 只要抓到一定數量就好, 不一定要抓完

如何抓完資料...

1. 流水號暴力掃完
   a. 臺灣公司資料 http://company.g0v.ronny.tw/ 我是從 00000000 - 99999999 把所有統一編號組合都跑過一次爬完的，爬蟲跑了三個月
   b. 有頁碼的話就可以把每一頁掃完
      i. http://www.mobile01.com/topiclist.php?f=566
   c. 有流水號 ID 的也可以用流水號 ID 來跑
      i. http://newtalk.tw/news/view/2016-05-10/73000
2. 利用搜尋功能
   a. 想辦法找出可以搜尋出全部條件
   b. http://prtr.epa.gov.tw/FacilityInfo/Data
   c. http://jirs.judicial.gov.tw/Index.htm

遇到驗證碼
怎麼辦?

captcha 如何對付？

1. 有的網站 captcha 只要輸入成功一次，這個 session 就一直可以抓到內容了，這種就用 cookie 複製法解決就好
2. 花錢用工人智慧解決：
   a. http://www.deathbycaptcha.com/user/login
   b. http://decaptcher.com/
   c. http://www.bypasscaptcha.com/

REGULAR EXPRESSION
超好用
的工具!!!

REGEX: Regular expression

- /.../, !...!, #...#, ,..., , REGEX 可以自由選擇 delimiter 當作開頭
- * 表示吻合 0 ~ N 筆, ? 表示吻合 0 or 1 筆
  - x* 吻合 "", "x", "xx", "xxx" …
  - x? 吻合 "", "x" ，不吻合 "xx", "Y"
  - x+ 吻合 "x", "xx" … 不吻合 ""
- [abc] 表示吻合 a, b, c
  - /b[ao]y/ 吻合 "boy", "bay" 不吻合 "by", "bey" …
  - [a-z] 表示 a ~ z 的小寫英文字母
  - [A-Z] 表示 A ~ Z 的大寫英文字母
  - [0-9] 表示 0 ~ 9 的數字
  - [a-zA-Z0-9] 表示英文大小寫字母或是數字都吻合
- [^abc] 表示不吻合 abc
  - /href="[^"]*"/ 表示吻合 href="..." 之間任何不是 " 的情況
  - /<div[^>]*>/ 表示吻合 , … 等各種情況
- ^xxx 表示 xxx 開頭, xxx$ 表示 xxx 結尾

REGEX: Regular expression on PHP

- 用括號 ( ) 包起來區塊表示希望能夠回傳的部分
  - /<a href="([^"]*)">/
    - <a href="http://foo.com"> 回傳 ['<a href="http://foo.com">', ' http://foo.com']
  - /([0-9]+) \+ ([0-9]+)/
    - 123 + 456 回傳 ["123 + 456", "123", "456"]
- preg_match($regex, $str, $matches);
  - preg_match('/I am (.*)/', 'Hi! I am Ronny', $matches) // $matches => ['I am Ronny', 'Ronny']
- preg_match_all($regex, $str, $matches);

練習

檔案 13.php

用 regex 抓出 https://www.ptt.cc/hotboard.html 裡面的 最後更新時間是幾點，以及

- 不需要用到 DOM 了，直接用 preg_match 來抓
- 還要要轉成 UTF-8 再抓喔

在什麼環境
跑爬蟲
比較好?

UNIX 環境

- 一次性爬蟲
  - 用 screen 跑爬蟲不間斷
- 定期新資料爬蟲
  - 用 crontab 跑 (every 1 minute, 5 minutes, 1 hours, 1 days …)
  - 如果是高頻率的爬蟲，例如五分鐘一次的檢 查，請確定五分鐘前那爬蟲是否已經跑完
    - 可以在爬蟲開跑時 touch('/tmp/crawling'); 跑完後用 unlink('/tmp/crawling'); 刪掉他，這樣只要 /tmp/crawling 存在就表示上一次的還沒跑完，那可能需要警告
    - 更嚴謹作法可以用 flock (http://php.net/manual/en/function.flock.php)
- 可以用 Amazon Web Service 架一個 proxy，讓爬蟲透過 proxy 抓資料，假如 IP 被擋了就換個 IP 再抓
  - curl_setopt($curl, CURLOPT_PROXY, '123.123.123.123:3128');
  - 連一些限使用數量的 API 也可以用這招...

最後總結

如何跟伺服器要到 HTML

- HTTP GET / POST
- cookie / session
- captcha 驗證碼
- referer / user agent
- 很溫柔，讓對方沒有感覺

要到並解析資料的列表

- 找出完整的列表
  - 一次性爬蟲
- 找出有更新的列表
  - 持續性爬蟲
- 利用搜尋功能
- 利用 API
- 從其他外部集合

要到並解析各別資料的結構性資料

- HTML DOM parser
- Regular Expression

我的爬蟲作品

- PTT 人氣 http://ptthot.ronny.tw/
  - https://github.com/ronnywang/ptthot
- 公司資料 http://company.g0v.ronny.tw/
  - https://github.com/ronnywang/twcompany/blob/master/webdata/models/Updater.php
- newsdiff http://newsdiff.g0v.ronny.tw/
  a. https://github.com/ronnywang/newsdiff/tree/master/webdata/models/Crawler
- 中華民國內閣記錄 https://ronnywang.github.io/taiwan-cabinet/
  a.
- 每日四大報 http://oldpaper.g0v.ronny.tw/
- 關貿進出口資料 http://portal.g0v.ronny.tw/
- …… http://ronny.tw/data or https://github.com/ronnywang

我想要知道這網頁怎麼爬

- https://tw.stock.yahoo.com/q/bc?s=0050
- https://play.google.com/store/apps/details?id=com.htc.launcher&hl=zh_TW
-