網頁爬蟲教學 (PHP crawler)

網頁爬蟲教學 (PHP crawler)
ronnywang @ HTC 2016/6/13, 14

==== ==========

關於 Ronny Wang
- Email: ronnywang@gmail.com
- 經歷
  - 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
- 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
  - https://github.com/ronnywang

==== ==========

一個完整的爬蟲包含…
- 如何跟伺服器要到 HTML
- 解析結構性資料 (網頁, JSON/XML, CSV)
- 找出需要的資料網頁 (範圍 / 頻率)

如何跟伺服器要到 HTML
- 產生 HTML 查詢
  - HTTP GET - 查詢參數接在網址之後
  - HTTP POST - 查詢參數藏在查詢通知內
- 網站擋機器人
  - 會認瀏覽器，必須是常見瀏覽器才給內容
    - User Agent
  - 會認來源 (從那個網址來)，沒有來源或外部就不給內容
    - Referer
  - 有驗證動作 (e.g. 驗證碼、18 禁)
    - cookie / session, 模擬行為
    - captcha 驗證碼
  - 短時間內大量存取就會阻擋
    - 很溫柔，讓對方沒有感覺
  - 絕招: 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...

解析結構性資料 (網頁, JSON/XML, CSV)
- HTML DOM parser (DOM, Document Object Model)
  PHP DOM manual: http://php.net/manual/en/book.dom.php
- Regular Expression

找出需要的資料網頁 (範圍 / 頻率)
- 範圍: 全部 / 時段
  - 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
  - 找出有更新的列表
- 頻率: 一次性 / 定期更新資料爬蟲
- 流水號暴力掃完
  - 臺灣公司資料 http://company.g0v.ronny.tw/
    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
  - 有頁碼的話就可以把每一頁掃完
    http://www.mobile01.com/topiclist.php?f=566
  - 有流水號 ID 的也可以用流水號 ID 來跑
    http://newtalk.tw/news/view/2016-05-10/73000
- 利用搜尋功能
  - 想辦法找出可以搜尋出全部條件
    http://prtr.epa.gov.tw/FacilityInfo/Data (縣市)
    http://jirs.judicial.gov.tw/Index.htm (法院名稱 + 全文檢索: (什麼關鍵式是所有資料都有))
  - 從其他外部集合 (Wiki、縣市組合)
- 利用 API (To be added)

==== 解析結構性資料 ==========

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

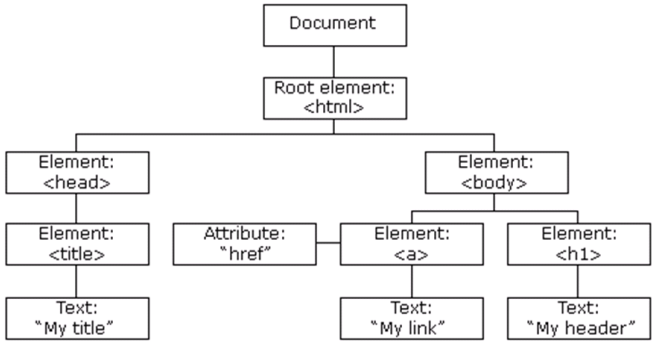
About HTML

- 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 安裝 (Ubuntu)
  sudo apt-get install php5
  sudo apt-get install php5-curl
- 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
- 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
  - TagName 必須為小寫
- $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
- $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01: 從 HTML 取出資料
- 抓出 PTT 人氣看板列表: https://www.ptt.cc/hotboard.html
- 檔名：1.php
- 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 ⋯));
  - 擷取網址內容: $html = file_get_contents('網址');
    - 建議可以先用 curl 指令把 HTML 抓下來，在開發時用靜態檔案，等到開發完成再改成用線上網址
    - '網址' 可換為檔案
    - curl https://www.ptt.cc/hotboard.html > ptt.html
  - 搜尋資料區塊，找出規律
    - 網頁上看到第一個是 Gossiping ，從原始檔找找看 Gossiping 在哪裡
  - 利用 getElementsBy ... 擷取區塊
  - 除了用 foreach ($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02: 斧頭幫 Level 1
- 斧頭幫 Level 1 - http://axe.g0v.tw/
- 檔名： 2.php
- 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03: 抓 PTT 的推文
- https://www.ptt.cc/bbs/MobileComm/M.1465283968.A.EA3.html
- 檔名：3.php
- 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04-1:
- 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

練習04-2:
- 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

==== 如何跟伺服器要到 HTML ==========

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
    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**
- 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
- 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

- https://curl.haxx.se/
- PHP cURL functions: http://php.net/manual/en/ref.curl.php
- curl is an open source command line tool and library for transferring data with URL syntax
- $curl = curl_init($url);
- curl_setopt($curl, **CURLOPT_RETURNTRANSFER**, true); // return the value of curl_exec() as string instead of outputting it out directly
  http://php.net/manual/en/function.curl-setopt.php => **CURLOPT_RETURNTRANSFER**
  (MUST HAVE - else cannot get the content)
- $contnet = curl_exec($curl);
- curl_close($curl);
- 等同於 $content = file_get_contents($url);

所以遇到八卦板的 case 怎麼辦

- 把「已經按下滿18歲」的 cookie 複製到程式中 => 複製 cookie 法
- 在程式端實作「我按下我已滿18歲」的動作 => 模擬行為法

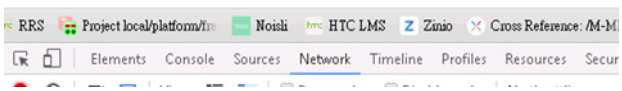
複製 cookie 法

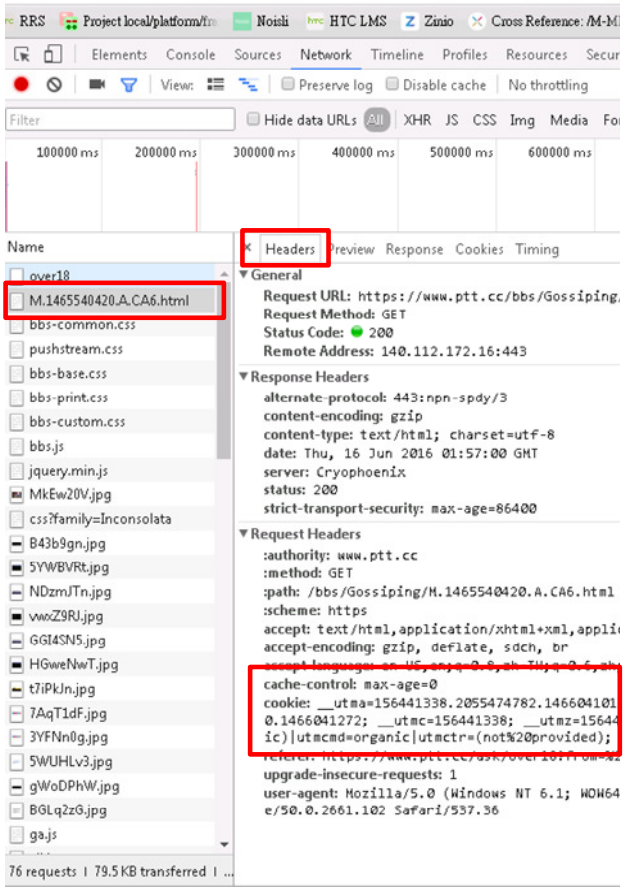
- 方法：
    - 透過 **Chrome 開發者工具**將已經成功可以讀到內容的 cookie 複製下來
      1. Ctrl+Alt+I to enable Google devTools
      https://developers.google.com/web/tools/chrome-devtools/?hl=en
    - 建議在 Chrome 無痕模式 (incognito mode) 下操作，瀏覽器關閉後 cookie 就會被清除，便於重複操作
    - curl_setopt($curl, **CURLOPT_HTTPHEADER**, array("Cookie:xxx"));
      http://php.net/manual/en/function.curl-setopt.php => **CURLOPT_HTTPHEADER**
      // An array of HTTP header fields to set, e.g. array('Content-type: text/plain', 'Content-length: 100', ...)
- 使用情況：
    - 比較省事
    - 這個狀況只能以人工做到，難以用程式做到時(Ex: 有驗證碼）
    - 狀況要可以被複製

練習05: 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)
    - 先把 file_get_contents 改寫成 curl 用法
    - 從 Chrome 開發者工具取得 cookie 現值
    - curl_setopt($curl, **CURLOPT_COOKIE**, 'xxx'); 貼進來
      // The contents of the *"Cookie: "* header to be used in the HTTP request
- cookie 取得步驟
  1. 在 Chrome 無痕模式下開啟網頁
  2. 開啟 Chrome 開發者工具 (Ctrl+Alt+I)
  3. 開啟 Network tag, 清除之前的資料

4. 點選 "我同意 ..."
5. 於清單中，點選目標網頁 (e.g. M.1465540420.A.CA6.html), 查詢 headers 內容
6. 於 Request Headers 內取得 cookie 內容





實作多步驟 (模擬行為法)

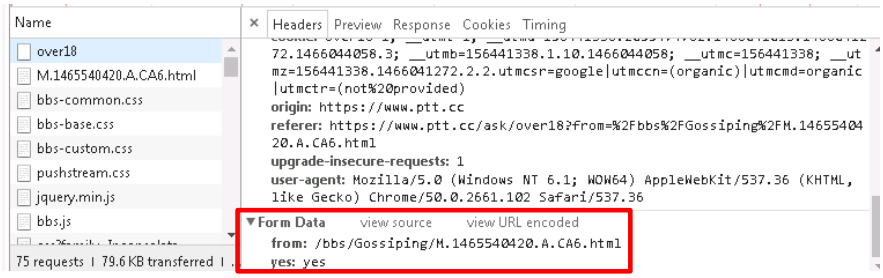
- 方法
  - curl 本身會保存 cookie ，所以只要 curl_init 一次取得 $curl 物件，然後把每個動作做進去
  - 先 POST 送出滿十八歲
  - 再用同一個 $curl 去要資料看看
- 使用情境
  - 多步驟比較複雜，不能直接複製 cookie 的情況

練習06: 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)
  Concept: 模擬 "over18" 的檢查流程取得 cookie 之後，在連線至目標網頁
  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 $curl 都會延續之前的 session
  // The name of the file containing the cookie data. If the name is empty, no cookies are loaded, but cookie handling is still enabled.
  3. 在抓資料之前先做出 POST 送出滿十八歲的動作
  4. curl_setopt($curl, **CURLOPT_POSTFIELDS**, 'aaa=bbb&ccc=ddd');
  // The full data to post in a HTTP "POST" operation
  5. curl_setopt($curl, **CURLOPT_URL**, ' 新的網址');
  6. 沿用 $curl 物件，來抓資料看看
- 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
  1~4. 同練習05 - cookie 取得步驟
  5. 於判別網頁的原始碼中, 得知點選 "滿十八歲" 之後，將會送到 /ask/over18 這個位址

```
<div class="bbs-screen bbs-content center clear">
    <form action="/ask/over18" method="post">
        <input type="hidden" name="from" value="/bbs/Gossiping/M.1465540420.A.CA6.html">
        <div class="over18-button-container">
            <button class="btn-big" type="submit" name="yes" value="yes">我同意，我已年滿十八歲<br><small>進入</small></button>
        </div>
        <div class="over18-button-container">
            <button class="btn-big" type="submit" name="no" value="no">未滿十八歲或不同意本條款<br><small>離開</small></button>
        </div>
    </form>
</div>
```

  6. 實際點選 "滿十八歲", 在 devTools 中找到 over18 的網址, 查詢 Headers 內容
  7. 於 form data 內可查詢 post 所傳送的資料



8. 於 form data 內, 點選 view source 取得傳送格式為
   from=%2Fbbs%2FGossiping%2FM.1465540420.A.CA6.html&yes=yes

練習07: 一次抓很多頁資料

- 斧頭幫 Lv2 抓有大量列表: http://axe.g0v.tw/level/2
- 檔案: 7.php
  - 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08: 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

網站常見擋機器人方式

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

練習09: 對付會擋機器人的網站

- http://axe.g0v.tw/level/4
- 檔案: 9.php
- 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…...

歷史的傷痕 Big5 : 處理 Big5 to 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10: 抓 Big5 網站, PTT 人氣

- PTT 人氣: https://www.ptt.cc/hotboard.html
- 檔名：10.php
- 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

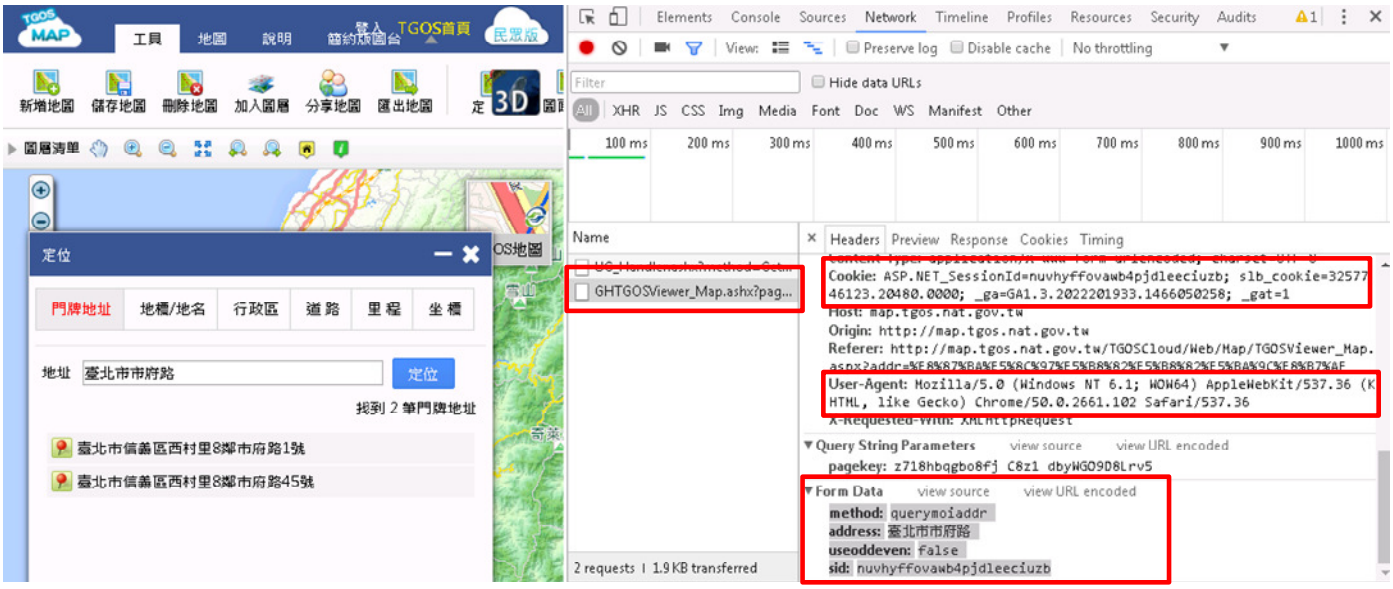
練習11: 透過 POST 送出搜尋條件，抓搜尋結果

- 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);

```
$params = array();
$params[] = 'name=' . urlencode($name);
$params[] = 'value1=' . urlencode($value1);
$post = implode('&', $params);
or
$params = array();
$params['name'] = $name;
$params['value1'] = $value1;
$post = http_build_query($params);
```

- 有檔機器人
- 用 Chrome 開發者工具，找出相關資訊 (URL, Cookie, User-Agent, Form Data)
  1. 在 Chrome 無痕模式下開啟網頁, 開啟開發者工具
  2. 開啟 Network tag, 清除之前的資料
  3. 因無法直接由 source code 取得執行頁資訊, 可透過執行查詢後, 檢視每一個可能網頁的 response, 發現是由 GHTGOSViewer_Map.ashx 傳回查詢結果

4. 再由 Headers 內, 取得
   - Request URL
   - Cookie
   - User-Agent
   - POST 內容, 包含 method, address, useroddeven, sid

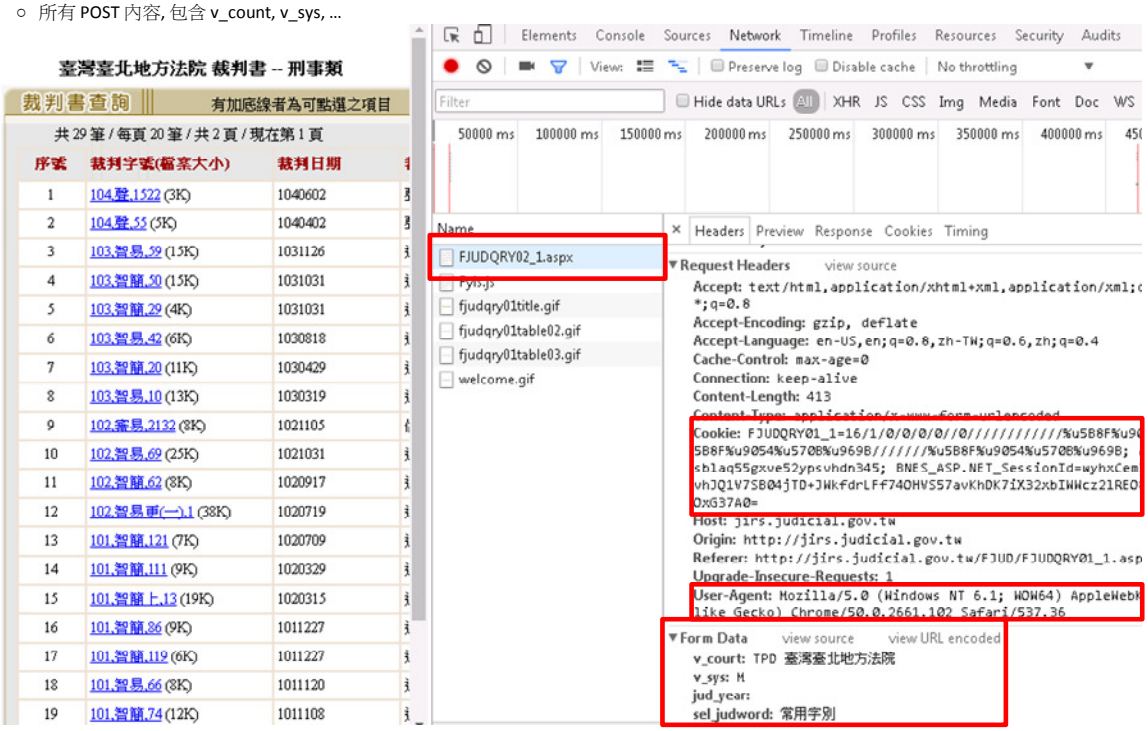


練習12: Big5 + post

- 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));

    => 若使用 http_build_query, 因此函式已包含 URL-encoded 機制, 故可移除 urlencode
  - 這個網站有擋機器人，把 9.php 斧頭幫 lv4 的技巧拿來用吧
- 用 Chrome 開發者工具，找出相關資訊 (URL, Cookie, User-Agent, Form Data)

1. 在 Chrome 無痕模式下開啟網頁, 開啟開發者工具
2. 開啟 Network tag, 清除之前的資料
3. 點選查詢結果網頁 (FJUDQRY02_1.aspx)
4. 再由 Headers 內, 取得
   - Request URL
   - Cookie
   - User-Agent
   - 所有 POST 內容, 包含 v_count, v_sys, ...



遇到驗證碼，怎麼辦？

captcha 如何對付？

- 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
- 花錢用工人智慧解決：
  - http://www.deathbycaptcha.com/user/login
  - http://decaptcher.com/
  - http://www.bypasscaptcha.com/

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);

**==== REGULAR EXPRESSION ========== **

REGEX: REGULAR EXPRESSION - 超好用的工具!!!
- /.../, !...!, #...#, ,..., , 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
- 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"]

REGEX: Regular expression on PHP
- 透過 preg_match 取出符合條件的字串
- preg_match($regex, $str, $matches);
  Will stop when get the first match
  http://php.net/manual/en/function.preg-match.php
  preg_match('/I am (.*)/', 'Hi! I am Ronny', $matches) // $matches => ['I am Ronny', 'Ronny']
- preg_match_all($regex, $str, $matches);
  Search all matches
  http://php.net/manual/en/function.preg-match-all.php

練習13: Regex
- 可先在 https://regex101.com/ 驗證 regular expression 的寫法
- 用 regex 抓出 https://www.ptt.cc/hotboard.html 裡面的 最後更新時間是幾點，以及
  - 配合 \s (http://www.w3schools.com/jsref/jsref_regexp_whitespace.asp)
  - 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...

Reference
- Sheethub: https://sheethub.com/
- 資料視覺化 李慕約公司 / Muyueh Data Visualization: https://www.facebook.com/%E8%B3%87%E6%96%99%E8%A6%96%E8%A6%BA%E5%8C%96-%E6%9D%8E%E6%85%95%E7%B4%84%E5%85%AC%E5%8F%B8-Muyueh-Data-Visualization-1711904639056487/
- Ronnywong's github: https://github.com/ronnywang/