R Crawler
Week2

關於教材授權

本教材之智慧財產權，

屬木刻思股份有限公司所有！

如果有朋友，覺得此教材很棒，希望能分享給朋友，或是拿此教材開課。非常歡迎大家來信至 course@agilearning.io 請求教材的使用授權唷！

課程資訊

- 網站 / 論壇 / 粉絲頁 / 廣播 / 共筆

課程廣播

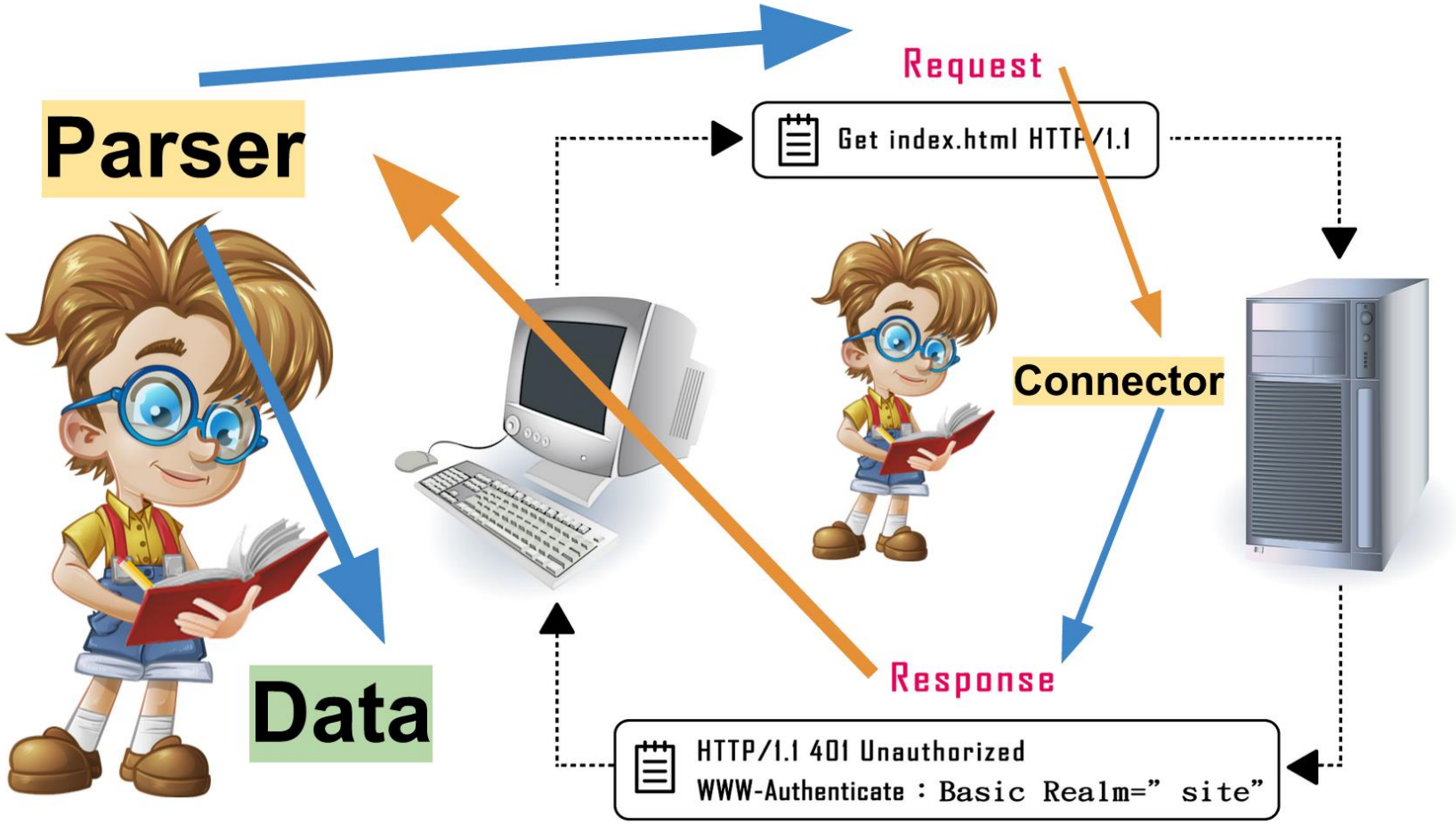
https://goo.gl/A5Rvtx

同學共筆

https://goo.gl/R7rqrg

Content

1. 爬蟲設計流程
2. HTML And JSON
3. GET
4. 中文網址秘密
5. Cookie
6. POST


Request
Parser
Get index.html HTTP/1.1
Connector
Data
Response
HTTP/1.1 401 Unauthorized
WWW-Authenticate : Basic Realm=" site"

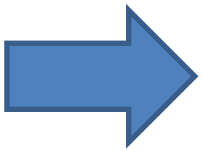
爬蟲 = 不斷重複connector and parser

- Step 1: Connector
  - GET or POST
- Step 2: Data Parser
  - 使用CSS selector or Xpath selector
  - 使用Regular Expression
  - json to data.frame

JSON

- 物件: 大括弧(Key:Value pair)
- 陣列: 中括弧(用“,”分割)

id	name
c4h4	XXX
c3h3	OOO



```
[
   { “id” : “c4h4” ,
     “name” : “XXX”
   },
   { “id” : “c3h3” ,
     “name” : “OOO”
   }
]
```

JSON

```
{
      "firstName": "John",
      "lastName": "Smith",
      "sex": "third",
      "age": 25,
      "address":
      {
          "streetAddress": "21 2nd Street",
          "city": "New York",
          "state": "NY",
          "postalCode": "10021"
      },
      "phoneNumber":
      [
          {
            "type": "home",
            "number": "212 555-1234"
          },
          {
            "type": "fax",
            "number": "646 555-4567"
          }
      ]
 }
```

YAHOO! 奇摩 新聞

新聞搜尋

首頁 | 政治 | 財經 | 影劇 | 運動 | 社會 | 地方 | 國際 | 生活 | 文教 | 健康

棒球 | 籃球 | 高爾夫 | 網球 | 綜合 | 運動熱門 | 運動影音 | 2014世足

黃蜂找前鋒 豪哥將添新同學

根據美媒報導，球隊即將簽下白人前鋒Tyler Hansbrough，盼成第6位新同學，雙方都已達成協議，只差球團正式宣布 ...

精采完全打擊 破亞洲人紀錄

瓊斯盃女籃 全新中華隊誕生

- 曹錦輝西裝筆挺 球迷讚型男
- 皇家拉尾盤 擊倒海盜強投Cole
- 陽岱鋼回穩 連6場敲安
- 哈登力壓柯瑞 獲球員票選MVP
- 19歲女籃世錦賽 中華首勝阿根廷
- 世大運張凱貞晉16強 李亞軒傷退
- 曾仁和奪勝投 羅國華完美防禦率破功
- 曹錦輝安打跑回追平分 卻無緣勝投
- 光芒投手開轟1比0 大聯盟新紀錄
- 小牛再被放鴿子？傑弗森被轉投騎士
- 鵜鶘老將底薪 簽老中鋒帕金斯
- 砸重金邀梅西露臉 加彭共和國否認

腦中想要結果(Yahoo新聞)

Filter

	newsText	newsTitle
1	效力於美國職棒大聯盟道奇隊的台灣投手曹錦輝今天再度登板...	曹錦輝西裝筆挺 球迷讚型男
2	記者張耀中／綜合報導明星外野手Alex Gordon受傷之後，...	MLB／皇家拉尾盤 擊倒海盜強投Cole
3	中國時報【吳政紘/綜合報導】日職火腿隊「台灣一哥」陽岱...	陽岱鋼回穩 連6場敲安
4	NBA球員工會舉辦第一屆NBPA票選活動。火箭當家球星哈登...	NBA》哈登力壓柯瑞　獲球員票選MVP
5	在俄羅斯舉行的19歲級世界女籃錦標賽，中華隊在預賽最後...	19歲女籃世錦賽 中華首勝阿根廷
6	剛在光州世大運奪下網球女單金牌台將張凱貞，一回台立刻轉...	鄭州網賽》世大運金牌張凱貞送蛋晉16強　李亞軒傷退

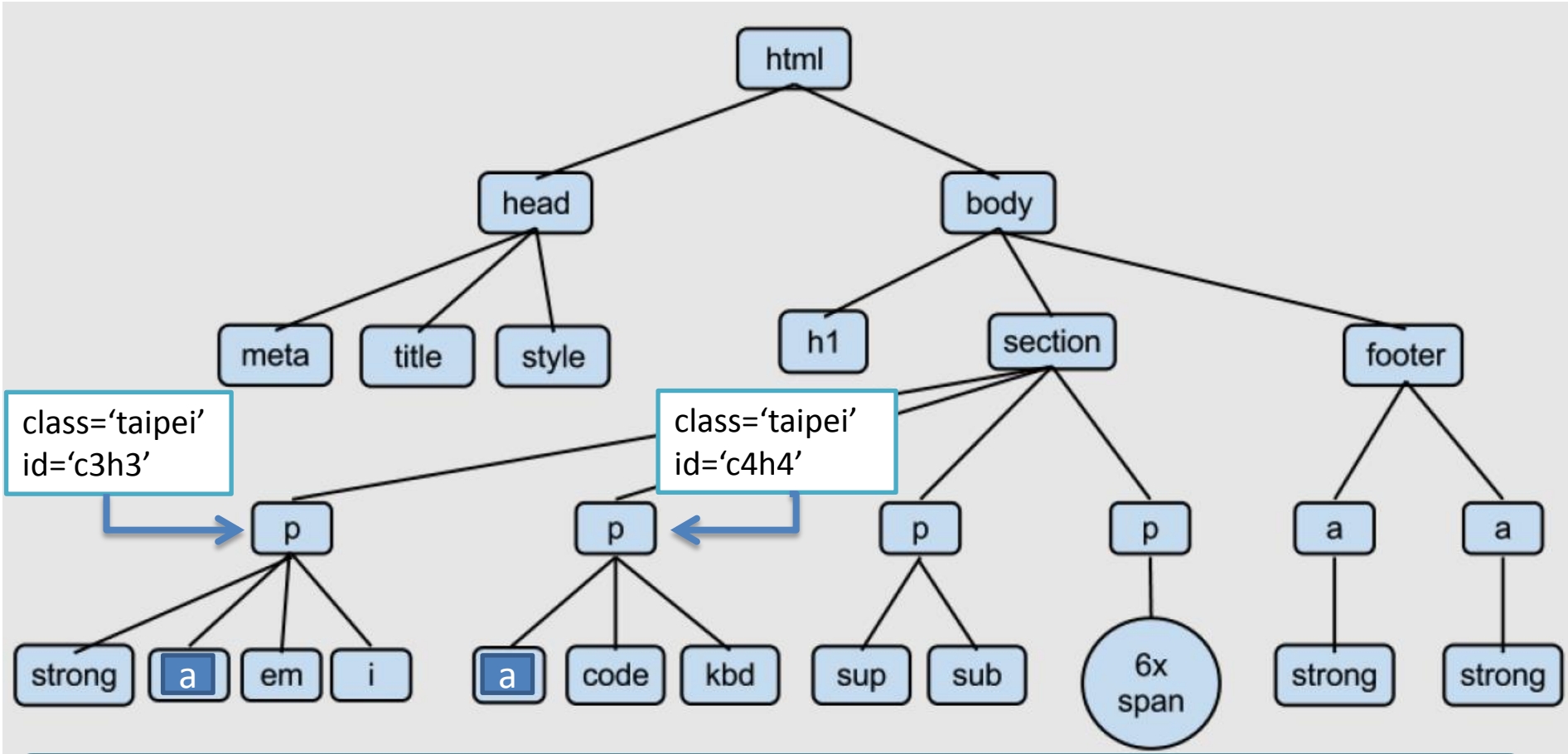
GET

- **GET**
  - 直接敲網址拿資料
  - 注意網址有時候會有參數
  - Function：httr::GET(URL)

Yahoo News

- Yahoo 運動新聞 為例, 瀏覽軌跡:
  - 1. 到https://tw.news.yahoo.com/sports/
  - 2. 點連結文章 => 看文章 (文章為所需資料)
- 爬蟲程式
  - 1. 抓新聞連結(資料頁的網址)
  - 2. 對每個連結去抓新聞內容

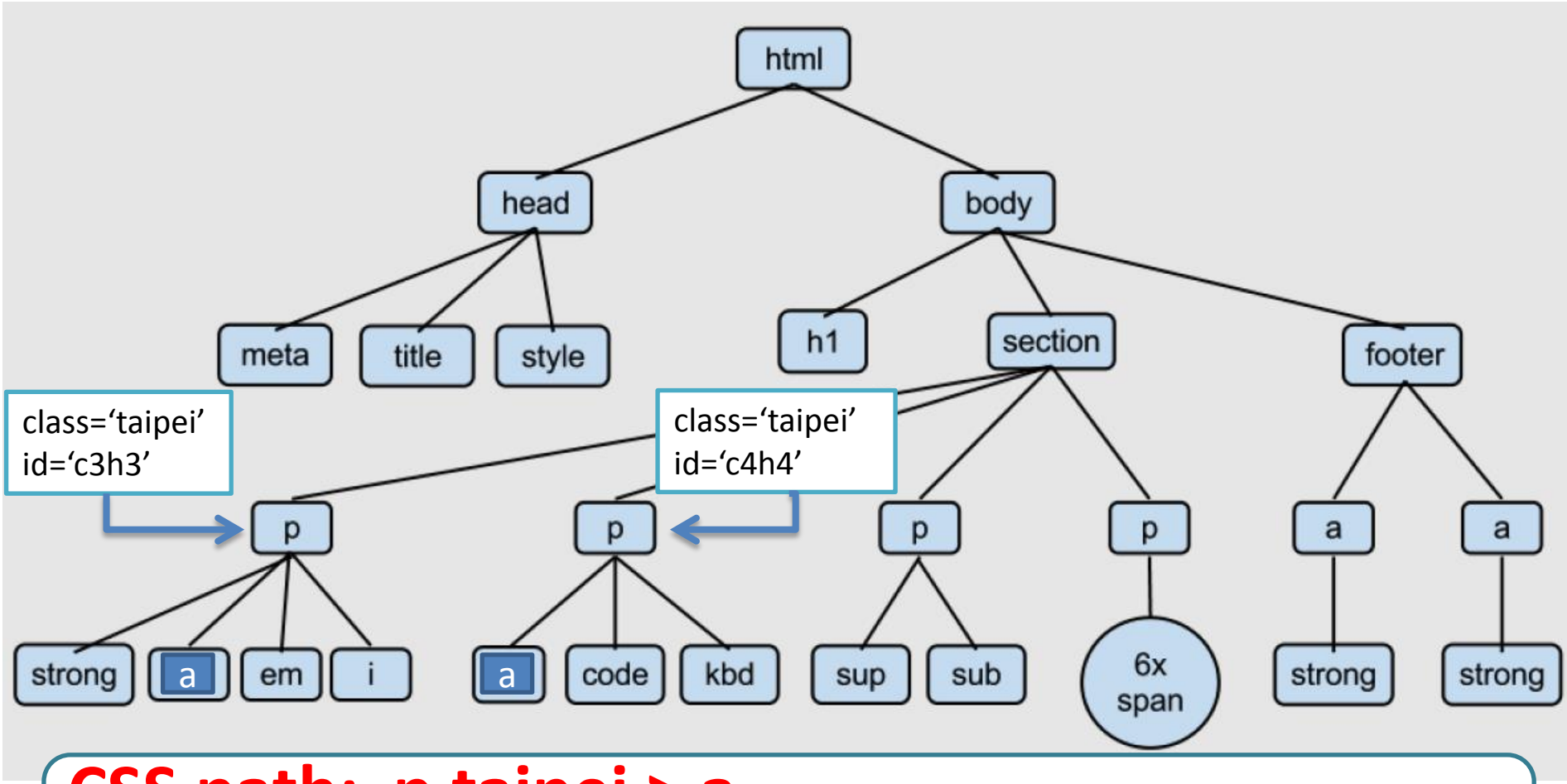
HTML



CSS path: html > body > section > p > a

Xpath : /html/body/section/p/a

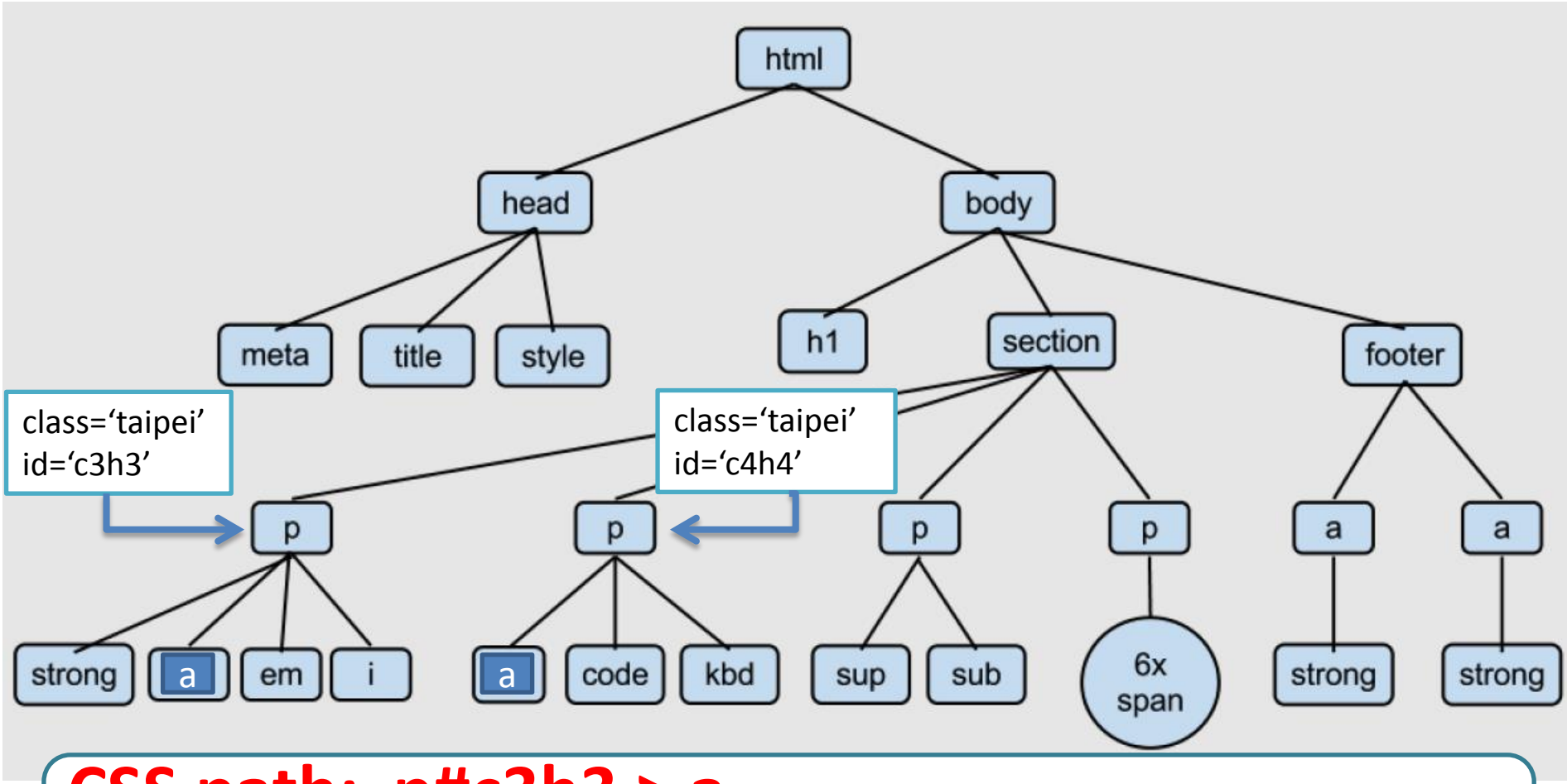
HTML



CSS path: p.taipei > a

Xpath : //p[@class='taipei']/a

HTML



CSS path: p#c3h3 > a

Xpath : //p[@id='c3h3']/a

中文網址


屬性(attribute)
<!-- -->
<div class="yom-mod yom-blist" id=
"mediablistmixedlpcatemp">
::before

::before
<ul class="tpl-title yom-list list-style-disc" id=
"yui_3_9_1_1_1437205250113_1112">
::before
<li class="list-story first" id=
"yui_3_9_1_1_1437205250113_1111">
<div class="txt" id=
"yui_3_9_1_1_1437205250113_1110">
<a href="/mlb-rod開轟相挺-田中將大優質先發奪
勝-045801263--mlb.html" class="title " data-
ylk="pkg:7f84f790-ed0b-3bfc-9858-20ef2d770bb3;
ver:e81a9540-2d09-11e5-ae7f-a42629afe5ae;lt:i;
pos:1;" data-rapid_p="1" id=
"yui_3_9_1_1_1437205250113_1109">A-Rod開轟相挺
田中將大優質先發奪勝
... #yui_3_9_1_1_1437205250113_1110 a#yui_3_9_1_1_1437205250113_1109.title
Styles Event Listeners DOM Breakpoints Properties

網址中需要中文

- 網址不能有中文需要轉成一段**密碼**
- 使用URLencode 進行加密，產生網址
- 需使用相關函數:
- **tmcn::toUTF8 =>** 轉UTF8編碼
- **URLencode =>** 轉網頁用編碼
  - %2F 為 ‘/’ # charToRaw("/")
  - %3A 為 ‘:’ # charToRaw(":")
- **URLdecode** => 解網頁用編碼

GET

- 觀察網址:
- URL?var1=val1&var2=val2
  => URL(var1=val1, var2=val2)
  Ex: substr?start=1&length=3
  => substr(start=1, length=3)
- 命名有規則 => 觀察設計者的網址命名規範去得到新網址
  Ex: https://tw.stock.yahoo.com/d/s/major_2451.html

創造網址方式

- sprintf(replace_string, str1, str2, …)

  => 把replace_string 的第1個 %s 用str1取代

  => 把replace_string 的第2個 %s 用str2取代

  ⇒…依此類推

  EX: URL?var1=val1&var2=val2

  ⇒ sprintf('URL?var1=%s&var2=%s', val1, val2)

- Q: sprintf('http://http://agilearning.io/%s1?var1=%s&var2=%s', val1, val2) # 會有錯誤 因為前面有%

創造網址方式(續)

- paste(characterVector, collapse = ‘’)

  ⇒characterVector 用collapse 合併貼成一個字串

  ⇒Ex:

  paste(c(‘URL?var1=‘ , val1 , ‘&var2=‘, val2),

  collapse = ‘’)

- paste0(string1,string2, string3,.., stringN)

  ⇒string1到stringN合併貼成一個字串

  ⇒Ex:

  paste0(‘URL?var1=‘ , val1 , ‘&var2=‘, val2)

爬蟲設計概念

- 點網頁
  - 1. 用CSS selector or Xpath selector
    抓到**不完整的部分**網址
  - 2. 利用拿不完整的部分網址
    去**創造完整**網址
- 看到資料頁 => 抓資料
- 思考關鍵點:
  如何得到所需資料全部網址及相關參數

POST

- **POST**
  - 除了提交網址外，還要提交參數才能拿資料
  - Function：
- 寫法1:

```
httr::POST(URL, body= list(para1=paraVal1,
                           para2=paraVal2))
```

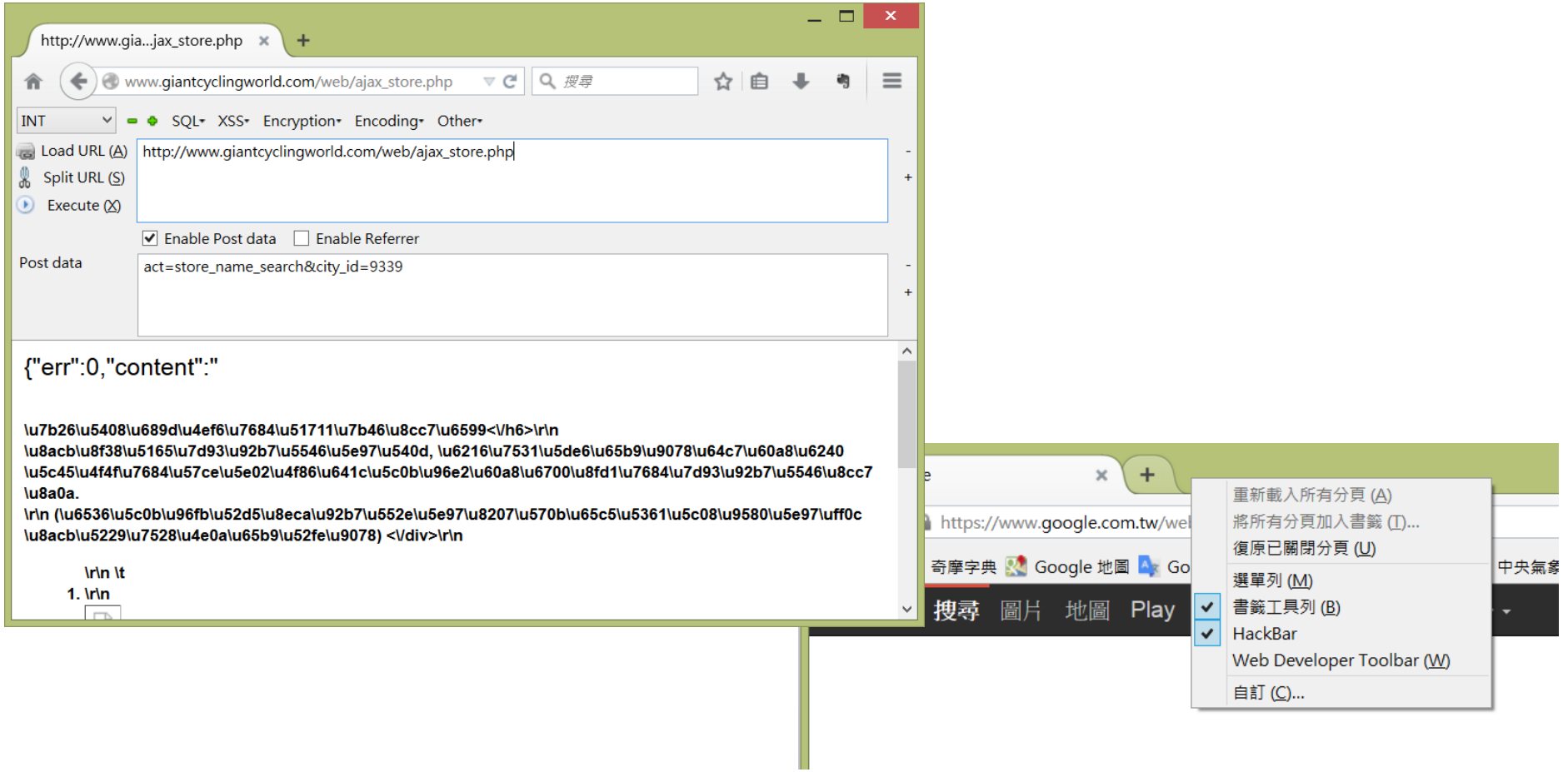
- 註: 中文請注意編碼
  (特別是作業系統為windows)

POST

- 寫法2:
  - httr::POST(URL, body= ‘para1=**paraVal1_encode**&para2=**paraVal2_encode**’)
- 寫法3(有時候需加參數encode):
  - httr::POST(URL, body= para1=paraVal1_encode&para2=paraVal2_encode, encode=‘form’)

Firefox::HackBar

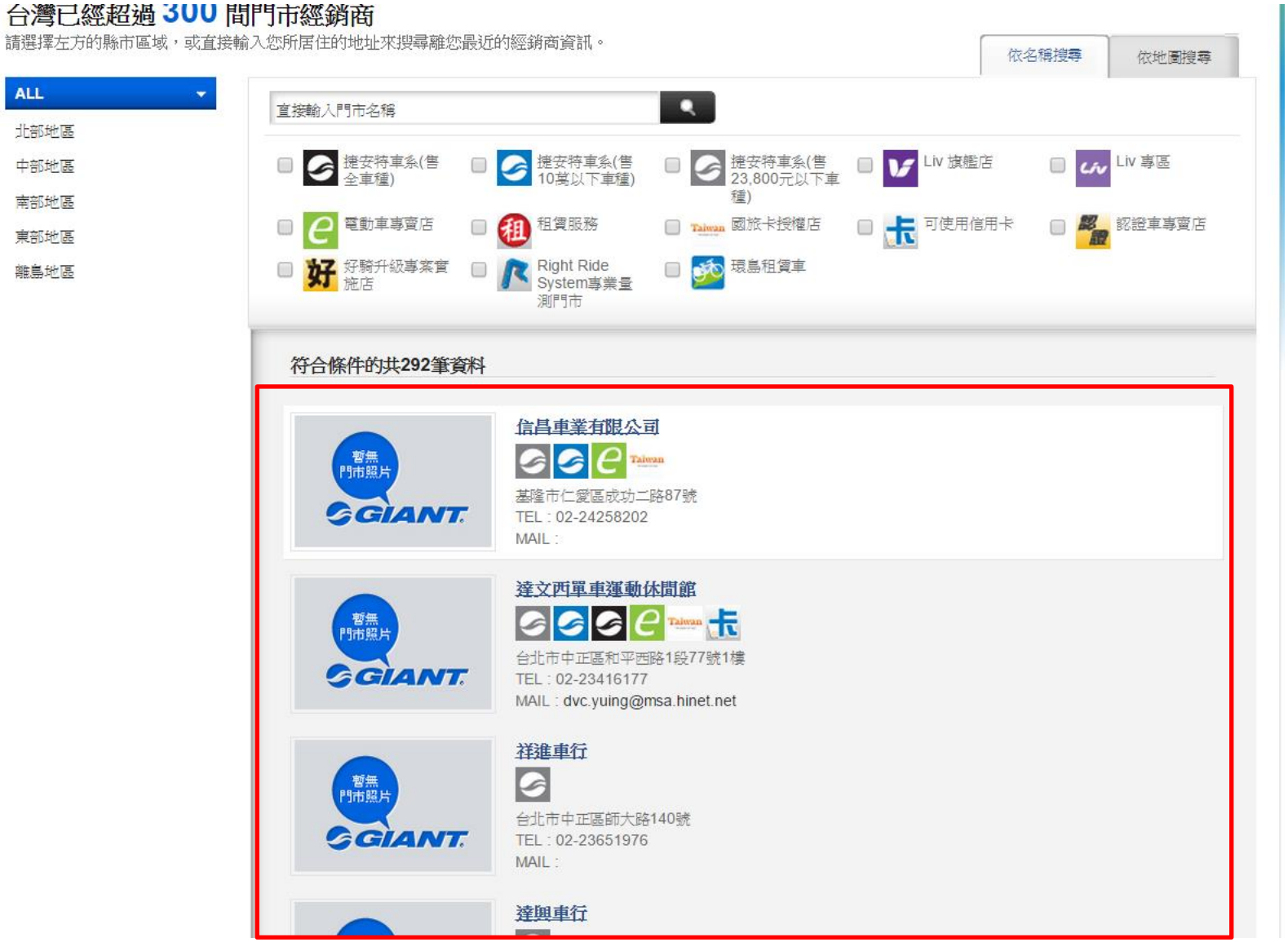
- 開關HackBar快捷鍵: F9



上市股票ID

公司代號	公司名稱	住址
1101	台灣水泥股份有限公司	台北市中山北路2段113號
1102	亞洲水泥股份有限公司	台北市大安區敦化南路2段207號30、31樓
1103	嘉新水泥股份有限公司	台北市中山北路2段96號
1104	環球水泥股份有限公司	台北市南京東路二段125號10樓

捷安特


台灣已經超過 300 間門市經銷商
請選擇左方的縣市區域，或直接輸入您所居住的地址來搜尋離您最近的經銷商資訊。
依名稱搜尋
依地圖搜尋
ALL
北部地區
中部地區
南部地區
東部地區
離島地區
直接輸入門市名稱
捷安特車系(售全車種)
捷安特車系(售10萬以下車種)
捷安特車系(售23,800元以下車種)
Liv 旗艦店
Liv 專區
電動車專賣店
租賃服務
國旅卡授權店
可使用信用卡
認證車專賣店
好騎升級專案實施店
Right Ride System專業量測門市
環島租賃車
符合條件的共292筆資料
暫無門市照片
GIANT
信昌車業有限公司
基隆市仁愛區成功二路87號
TEL：02-24258202
MAIL：
達文西單車運動休閒館
台北市中正區和平西路1段77號1樓
TEL：02-23416177
MAIL：dvc.yuing@msa.hinet.net
祥進車行
台北市中正區師大路140號
TEL：02-23651976
MAIL：
達興車行

腦中想要結果(捷安特)

Filter

store	addr	tel_no	email
信昌車業有限公司	基隆市仁愛區成功二路87號	0224258202	
達文西單車運動休閒館	台北市中正區和平西路1段77號1樓	0223416177	dvc.yuing@msa.hinet.net
祥進車行	台北市中正區師大路140號	0223651976	
達興車行	台北市大同區太原路142之11號	0225554803	
民昇自行車行	台北市松山區民生東路五段35號1樓	0227669501	
長春車行	台北市文山區興隆路一段279號1樓	0229319569	al43@ms9.hinet.net
順天自行車行	台北市文山區木柵路三段33號一樓	0229388187	A22053168@yahoo.com.tw
世明車行	台北市內湖區內湖路一段597號一樓	0227986196	shemin@umail.hinet.net
雄輪車行	06台北市大安區和平東路三段308巷30號1樓	0227327328	
德昌裝璜車業行	台北市信義區莊敬路458號1樓	0227208854	
榮泰車行	台北市大同區重慶北路3段295巷12號	0225925889	
捷輪自行車行	台北市萬華區康定路141號1樓	0223612863	
[illegible]	[illegible]	[illegible]	[illegible]

1 to 13 of 292 entries

Cookie

- **Cookie**:
  - 需帶cookie的才能拿資料，需在函數中多一個參數
  - config=set_cookies(cookie1=ckVal1, cookie2=ckVal2)

EX1:

```
 res =
GET("https://www.ptt.cc/bbs/Gossiping/index.html",
                        config=set_cookies("over18"="1"))
```

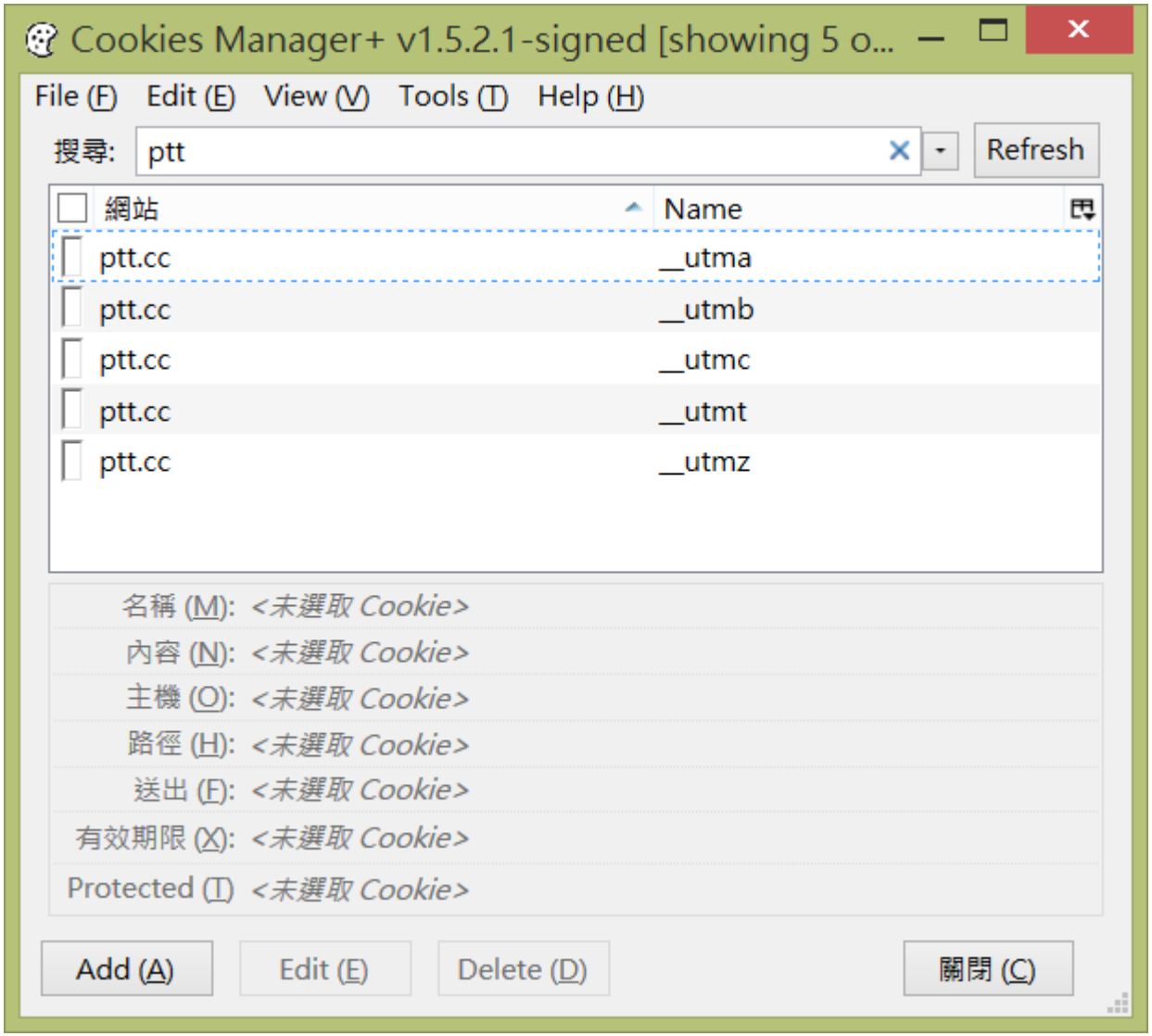
PTT Gossiping

- PTT Movie為例, 瀏覽軌跡:
  - 1. https://www.ptt.cc/bbs/Gossiping/index.html
  - 2. 點連結文章 => 看文章 (文章為所需資料)
- 爬蟲程式
  - 0. https://www.ptt.cc/bbs/Gossiping/index.html
    抓頁數(最後一頁是第幾頁)
  - 1. 抓資料頁的網址
  - 2. 拿資料頁的網址抓內容

腦中想要結果(Gossiping)

	postId	postUrl	pushTag	userId	text	time
1	M.1431338763.A.1BF	https://www.ptt.cc/bbs/Gossiping/M.1431338763.A.1BF....	→	conca	: 是的 台女最美	05/11 18:06
2	M.1431338763.A.1BF	https://www.ptt.cc/bbs/Gossiping/M.1431338763.A.1BF....	→	MeiLu5566	: 土耳其帥勾我可以<3	05/11 18:06
3	M.1431338763.A.1BF	https://www.ptt.cc/bbs/Gossiping/M.1431338763.A.1BF....	→	james732	: 台灣去土耳其的機票要怎麼買？	05/11 18:07
4	M.1431338763.A.1BF	https://www.ptt.cc/bbs/Gossiping/M.1431338763.A.1BF....	推	Rocks5566	: 土耳其火槍兵還可以讓你看可愛的小兔兔	05/11 18:07
5	M.1431338763.A.1BF	https://www.ptt.cc/bbs/Gossiping/M.1431338763.A.1BF....	→	conca	: 土耳其算亞洲？	05/11 18:08
6	M.1431338763.A.1BF	https://www.ptt.cc/bbs/Gossiping/M.1431338763.A.1BF....	推	Jason0813	: 土耳其人可是 小隻馬 稱號的發明人 公認真正大師	05/11 18:09
7	M.1431338763.A.1BF	https://www.ptt.cc/bbs/Gossiping/M.1431338763.A.1BF....	推	BestGarenTw	: 倫家女生耶	05/11 18:11
8	M.1431338763.A.1BF	https://www.ptt.cc/bbs/Gossiping/M.1431338763.A.1BF....	→	wanghong	: 伊朗也很美很帥~~	05/11 18:11
9	M.1431338763.A.1BF	https://www.ptt.cc/bbs/Gossiping/M.1431338763.A.1BF....	噓	Pony5566	: remove kebab	05/11 18:28
10	M.1431338763.A.1BF	https://www.ptt.cc/bbs/Gossiping/M.1431338763.A.1BF....	→	lobotime	: 西亞中亞都白種人底的電爆東亞好嗎	05/11 18:32
11	M.1431338763.A.1BF	https://www.ptt.cc/bbs/Gossiping/M.1431338763.A.1BF....	推	bluenicole	: 我覺得女生的話印度的比較美耶	05/11 21:02
12	M.1431338763.A.1BF	https://www.ptt.cc/bbs/Gossiping/M.1431338763.A.1BF....	→	Leeng	: 東俄算亞洲的話都俄最強 高加索>>>>其他種族不可避	05/12 08:49

Firefox::Cookies manager



Cookie

- **Cookie**:
  - 需帶cookie的才能拿資料，需在函數中多一個參數
  - config=set_cookies(cookie1=ckVal1, cookie2=ckVal2)
- EX:

```
res = POST("http://lvr.land.moi.gov.tw/N11/login.action",
        body=list(command='login', rand_code='1015',
                  in_type='land'),
            config=set_cookies('JSESSIONID'=
                  "0607E9012D564796528C206E2CA09074.jvm2",
                   'slb_cookie'= "33728704.20480.0000")
        )
```