



Spider and R

2013. 10. 16

Yen + 1 @ NCHU



About Me ..

- ☐ 顏嘉儀 Yen + 1
- ☐ 台大經濟所碩二
- ☐ yen.chiayi@gmail.com

- ☒ R user
- ☒ Python user

About R ...

...

都已經到最後一場了耶
還需要介紹 R 是什麼嗎XD

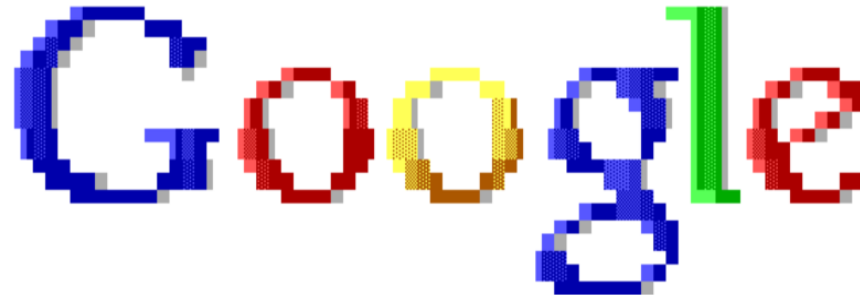
Pass ~~

About Spider...

Spider is everywhere !

Google, Amazon, 台灣公車通 , Whoscall,

NumerInfo (<-- 我放上來了 !) , Ebay(EC Lee) ,求職小幫手(Ronny Wang)



ANSI forever

Google Search

I'm Feeling Lucky

Are you ready to be a

SpideR man ?

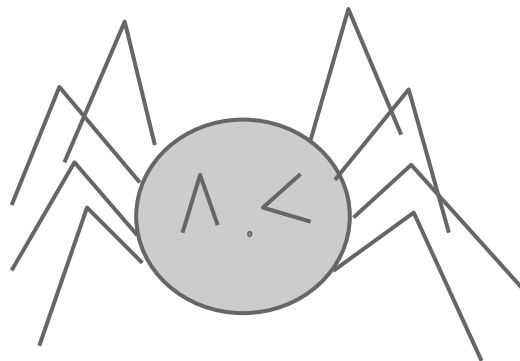


Wait...

**How can I
shot the web =
=?**

一隻最簡單的蜘蛛
要有身體跟八隻腳

?



一隻最簡單的爬蟲
包含兩個部份：

Connector

+

Parser



?

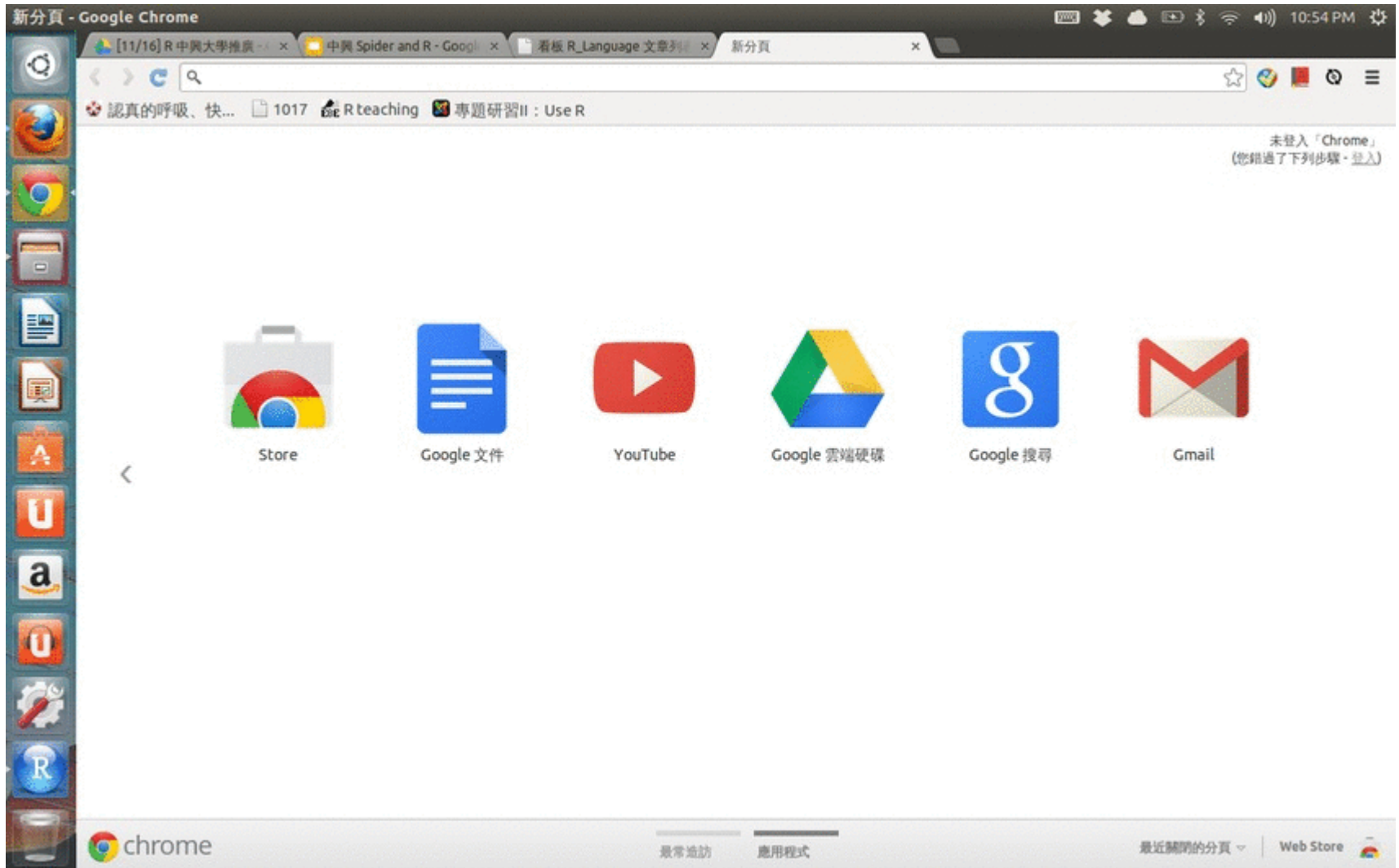
Connector & Parser ?

What is it ?? (可以吃嗎 ??)

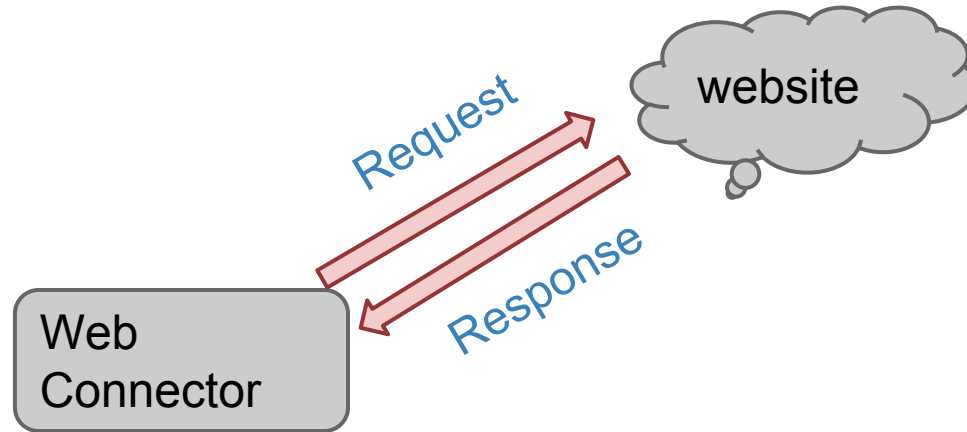
Think of the way you (human beings)
scrap some info from a webpage ...

Step 1: step into the right webpage

Step 1: step into the right webpage



The Structure of a Spider



Spiders just like counterpart of human beings.

Let's build our "Connector" in R.

“Connector” in R

```
if (!require(RCurl)) install.packages("RCurl")
```

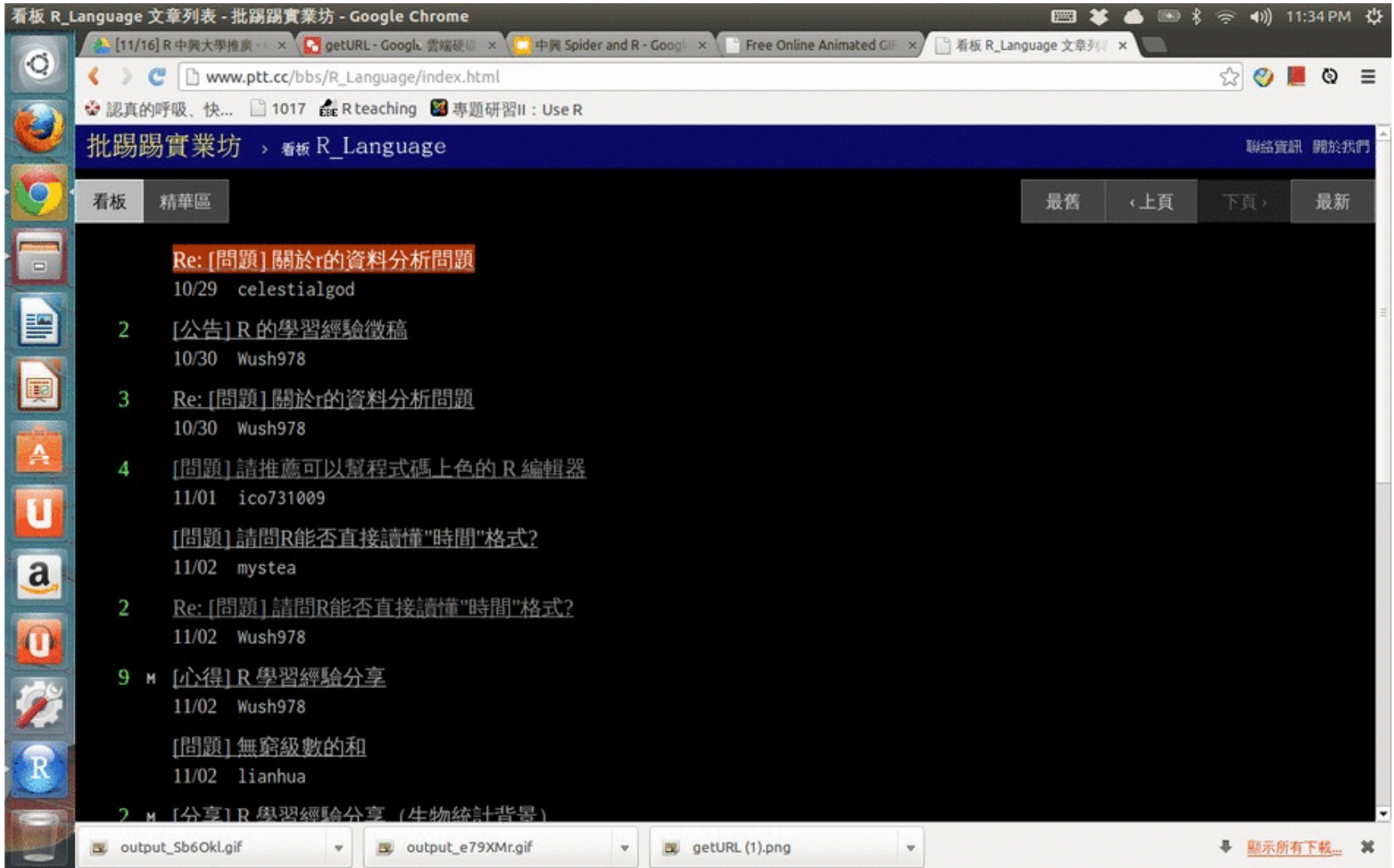
```
html = getURL(url = "http://www.ptt.cc/bbs/R_Language/index.html")
```

```
print(html)
```

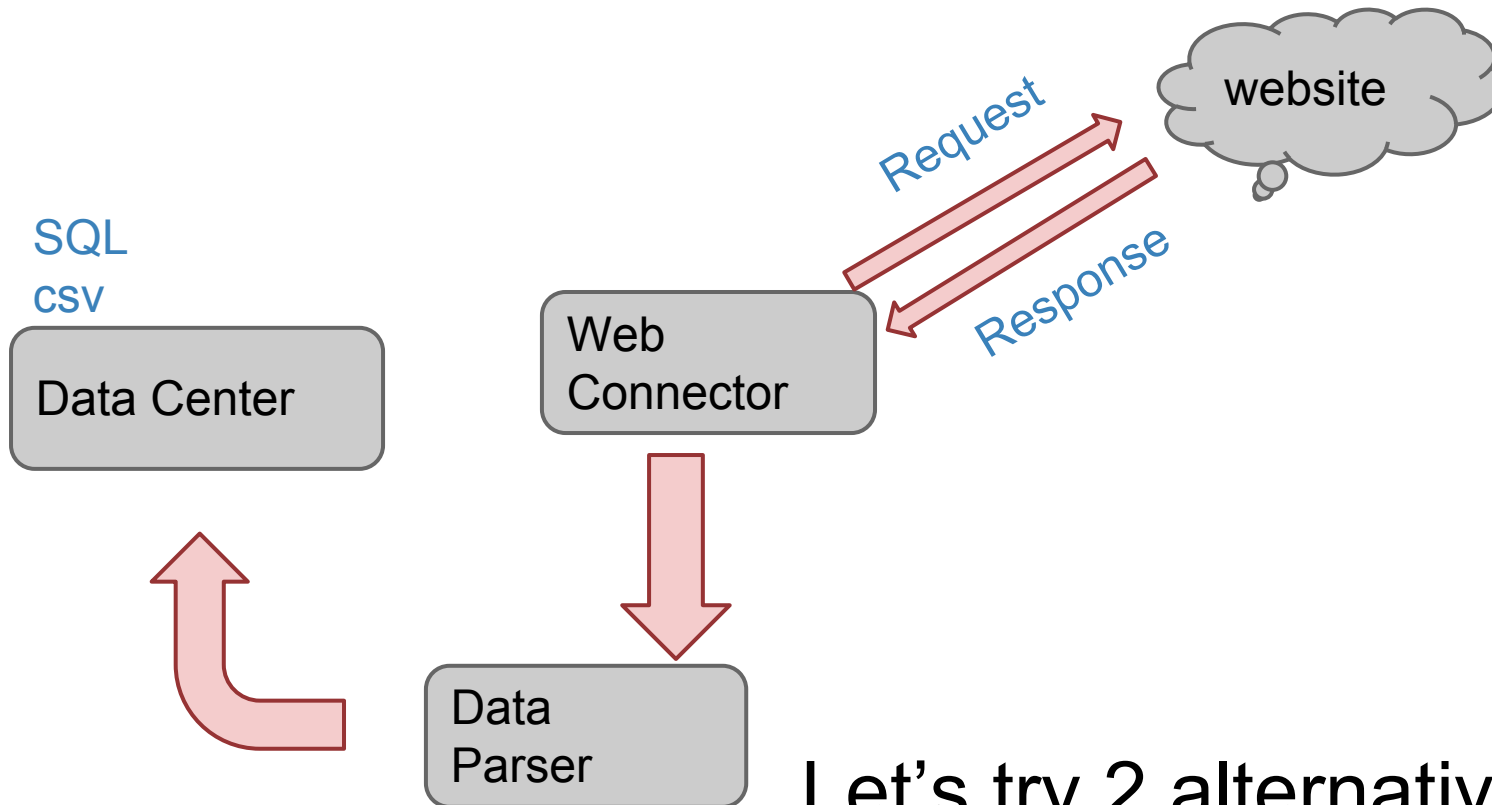
[<div class="r-list-container bbs-screen"><div class="r-ent"><div class="nrec"></div><div class="mark"></div><div class="title">Re: \[問題\] 關於r的資料分析問題</div><div class="meta"><div class="date">10/29</div><div class="author">celestalgod</div></div></div><div class="r-ent"><div class="nrec"></div><div class="mark"></div><div class="title">\[公告\] R 的學習經驗徵稿</div><div class="meta"><div class="date">10/30</div><div class="author">Wush978</div></div></div><div class="r-ent"><div class="nrec"></div><div class="mark"></div><div class="title">Re: \[問題\] 關於r的資料分析問題</div><div class="meta"><div class="date">10/30</div><div class="author">Wush978</div></div></div><div class="r-ent"><div class="nrec"></div><div class="mark"></div><div class="title">\[問題\] 請推薦可以幫程式碼上色的 R 編輯器</div><div class="meta"><div class="date">11/01</div><div class="author">ico731009</div></div></div><div class="r-ent"><div class="nrec"></div><div class="mark"></div><div class="title">\[問題\] 請問能否直接讀懂#34;時間#34;格式?</div><div class="meta"><div class="date">11/02</div><div class="author">mystee</div></div></div><div class="r-ent"><div class="nrec"></div><div class="mark"></div><div class="title">Re: \[問題\] 請問R能否直接讀懂#34;時間#34;格式?</div><div class="meta"><div class="date">11/02</div><div class="author">Wush978</div></div></div><div class="r-ent"><div class="nrec"></div><div class="mark"></div><div class="title">\[心得\] R 學習經驗分享</div><div class="meta"><div class="date">11/02</div><div class="author">Wush978</div></div></div><div class="r-ent"><div class="nrec"></div><div class="mark"></div><div class="title">\[問題\] 無窮級數的和</div><div class="meta"><div class="date">11/02</div><div class="author">lanhua</div></div></div><div class="r-ent"><div class="nrec"></div><div class="mark"></div><div class="title">\[分享\] R 學習經驗分享\(生物統計背景\)</div><div class="meta"><div class="date">11/04</div><div class="author">andrew43</div></div></div><div class="r-ent"><div class="nrec"></div><div class="mark"></div><div class="title">\[問題\] 尋找列聯表專用的重疊性function</div><div class="meta"><div class="date">11/04</div><div class="author">osuper</div></div></div></div>](#)

Step 2: scrap the info you want

Step 2: scrap the info you want



The Structure of a Spider



(1) RegEx

(2) XPath

Let's try 2 alternatives
to build our "Parser" in R

“Parser” in R (regular expression)

```
div = "<div class=\"meta\">
      <div class=\"date\">10/29</div>
      <div class=\"author\">celestialgod</div>
    </div> "
date = gsub("[^((0?[1-9]|1[012])[ /](0?[1-9]|[12][0-9]|3[01]))]", "", div)
```

```
[1] "1029"
```

What is “RegEx” ?

example: (0 ? [1-9] | 1[012]) [/]
(0 ? [1-9] | [12][0-9] | 3[01])

10/29

see details: <http://atedev.wordpress.com/2007/11/23/%E6%AD%A3%E8%A6%8F%E8%A1%A8%E7%A4%BA%E5%BC%8F-regular-expression/>

“Parser” in R (Xpath)

批踢踢實業坊 > 看板 R_Language

看板 精華區

Re: [問題] 關於r的資料分析問題
10/29 celestialgod


2 [公告] R 的學習經驗徵稿
10/30 Wush978

3 Re: [問題] 關於r的資料分析問題
10/30 Wush978

Elements Resources Network Sources Timeline Profiles Audits Console

```
<!DOCTYPE html>
<!-- saved from url=(0043)http://www.ptt.cc/bbs/R_Language/index.html -->
<html>
  <head>...</head>
  <body>
    <div id="topbar-container">...</div>
    <div id="main-container">
      <div id="action-bar-container">...</div>
      <div class="r-list-container bbs-screen">
        <div class="r-ent">
          <div class="nrec"></div>
          <div class="mark"></div>
          <div class="title">
            <a href="http://www.ptt.cc/bbs/R_Language/M.1383062358.A.0A3.html">Re: [問題] 關於r的資料分析問題</a>
          </div>
          <div class="meta">
            <div class="date">10/29</div>
            <div class="author">celestialgod</div>
          </div>
          <div class="r-ent">...</div>
          <div class="r-ent">...</div>
        </div>
      </div>
    </body>
  </html>
```

```
html
/body
/div[@id='main-container']
/div[@class='r-list-container bbs-screen']
/div[@class='title']
/a
```



What is “Xpath” ?

like “address” in HTML doc

Google Developer Tool (ctrl + shift + I)

批踢踢實業坊 > 看板 R_Language 聯絡資訊 關於我們

看板 精華區

最舊 < 上頁 下頁 > 最新

1 [Re: \[問題\] 關於r的資料分析問題](#)
10/29 celestialgod

2 [\[公告\] R 的學習經驗徵稿](#)
10/30 Wush978

3 [Re: \[問題\] 關於r的資料分析問題](#)
10/30 Wush978

www.ptt.cc/bbs/R_Language/M.1383062358.A.0A3.html

Elements Resources Network Sources Timeline Profiles Audits Console

```
<!DOCTYPE html>
<html>
  <head>...</head>
  <body>
    <div id="topbar-container">...</div>
    <div id="main-container">
      <div id="action-bar-container">...</div>
      <div class="r-list-container bbs-screen">
        <div class="r-ent">
          <div class="nrec"></div>
          <div class="mark"></div>
          <div class="title">
            <a href="/bbs/R_Language/M.1383062358.A.0A3.html">Re: [問題] 關於r的資料分析問題</a>
          </div>
          <div class="meta">...</div>
        </div>
        <div class="r-ent">...</div>
        <div class="r-ent">...</div>
        <div class="r-ent">...</div>
      </div>
    </div>
  </body>
</html>
```

Computed Style ☐ Show inherited

Styles

element.style {

Matched CSS Rules

media="screen" [index.html](#)
a:visited { [bbs.css:117](#)
color: #888;

media="screen" [index.html](#)
a:hover { [bbs.css:113](#)
color: #333;
background-color: #ccc;

media="screen" [index.html](#)
a:link { [bbs.css:110](#)
color: #aaa;

html body div#main-container div.r-list-container.bbs-screen div.r-ent div.title a

Chrome Plugin: Xpath Helper

Query:

```
//div[@class='title']/a
```

Results (18):

Re: [問題] 關於r的資料分析問題

[公告] R 的學習經驗徵稿

Re: [問題] 關於r的資料分析問題

[問題] 請推薦可以幫程式碼上色的 R 編輯器

Re: [問題] 關於r的資料分析問題

10/29 celestialgod

2

[公告] R 的學習經驗徵稿

10/30 Wush978

3

Re: [問題] 關於r的資料分析問題

10/30 Wush978

4

[問題] 請推薦可以幫程式碼上色的 R 編輯器

11/01 ico731009

[問題] 請問R能否直接讀懂"時間"格式?

11/02 mystea

2

Re: [問題] 請問R能否直接讀懂"時間"格式?

11/02 Wush978

9 M

[心得] R 學習經驗分享

11/02 Wush978

[問題] 無窮級數的和

11/02 lianhua

The shorter the Xpath,
the more Robust the result.

//div[@class='title'] /a

“Parser” in R (Xpath)

```
if (!require(RCurl)) install.packages("RCurl")
if (!require(XML)) install.packages("XML")
html = getURL("http://www.ptt.cc/bbs/R_Language/index.html")
```

```
xml = htmlParse(html)
Xpath = "//div[@class='title']/a/text()"
titles = xml [Xpath]
```

```
[1] "Re: [問題] 關於r的資料分析問題"      "[公告] R 的學習經驗徵稿"
[3] "Re: [問題] 關於r的資料分析問題"      "[問題] 請推薦可以幫程式碼上色的 R 編輯器"
[5] "[問題] 請問R能否直接讀懂\"時間\"格式?"  "Re: [問題] 請問R能否直接讀懂\"時間\"格式?"
[7] "[心得] R 學習經驗分享"              "[問題] 無窮級數的和"
[9] "[分享] R 學習經驗分享(生物統計背景)" "[問題] 尋找列聯表專用的的重抽樣 function"
[11] "Re: [問題] 尋找列聯表專用的的重抽樣 function" "[問題] 初學R的問題"
[13] "Re: [問題] 初學R的問題"              "Re: [問題] 初學R的問題"
[15] "[問題] error in xts"                  "[問題] 想請問計算極值問題"
[17] "[問題] 新手用R, 有關 Tinn-R與Rcmdr"
```

RegEx v.s. XPath Parser

Regular Expression

Every characters are
treated as the same
(difficult to maintain)

Alternatives: XPath

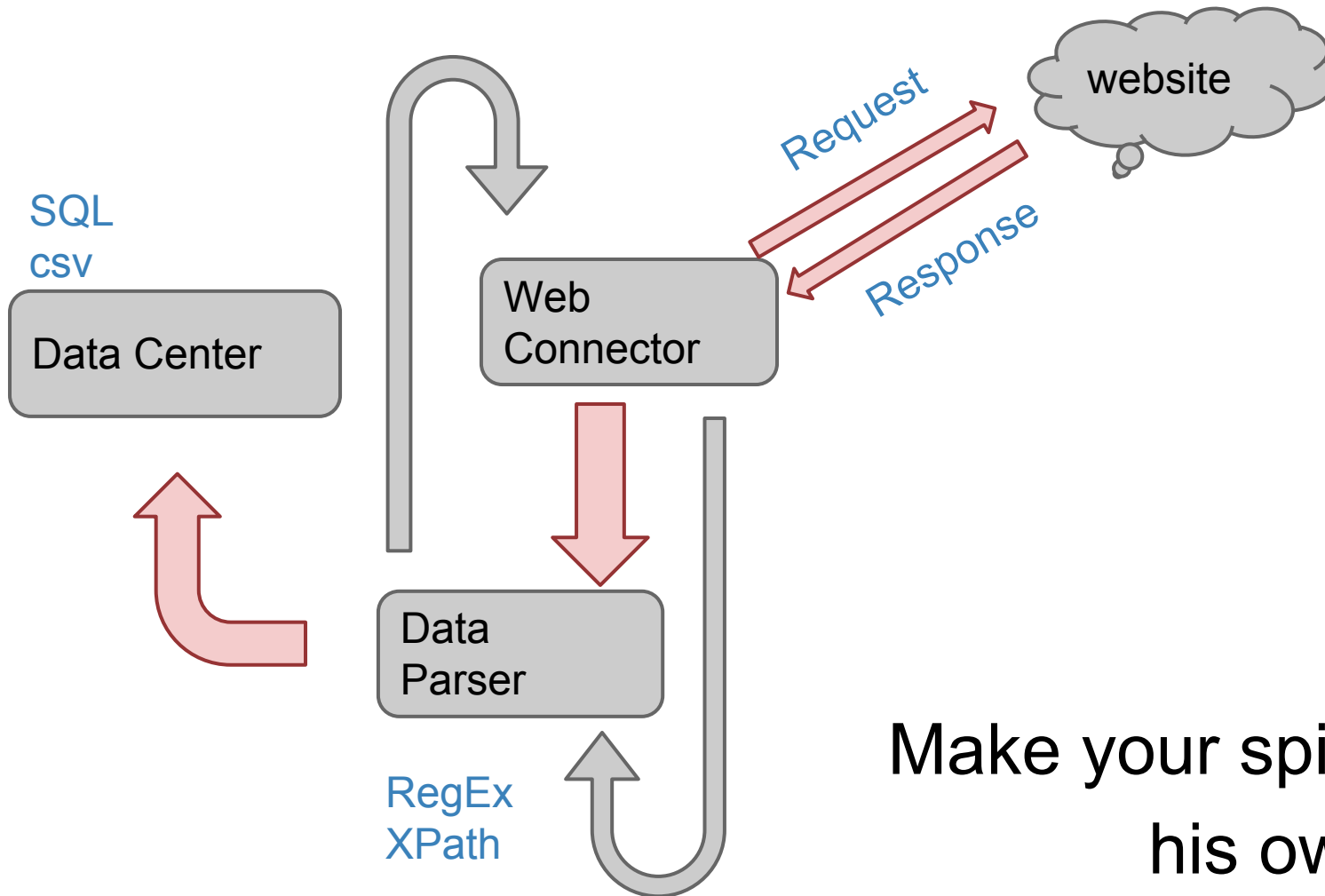
html doc can be a
structured data
(rely on clean html)

```
1643 <th>買進<br>期貨商代號</th>
1644 <th>買進期貨商名稱</th>
1645 <th>賣出<br>期貨商代號</th>
1646 <th>賣出期貨商名稱</th>
1647 </tr>
1648
1649 <tr>
1650 <td align="center">7900</td>
1651 <td align="center">313</td>
1652 <td align="center">3</td>
1653 <td align="center">S888</td>
1654 <td align="left">國泰證券</td>
1655 <td align="center">F004</td>
1656 <td align="left">凱基期貨</td>
1657 </tr>
1658
1659 <tr>
1660 <td align="center">7900</td>
1661 <td align="center">316</td>
1662 <td align="center">3</td>
1663 <td align="center">F034</td>
1664 <td align="left">漢帝華期貨</td>
1665 <td align="center">F004</td>
1666 <td align="left">凱基期貨</td>
1667 </tr>
1668
1669 <tr>
1670 <td align="center">7950</td>
1671 <td align="center">269</td>
1672 <td align="center">1</td>
1673 <td align="center">S980</td>
1674 <td align="left">元大寶來證券</td>
1675 <td align="center">F039</td>
1676 <td align="left">大昌期貨</td>
1677 </tr>
1678
1679 <tr>
1680 <td align="center">8000</td>
1681 <td align="center">223</td>
1682 <td align="center">1</td>
```

What if

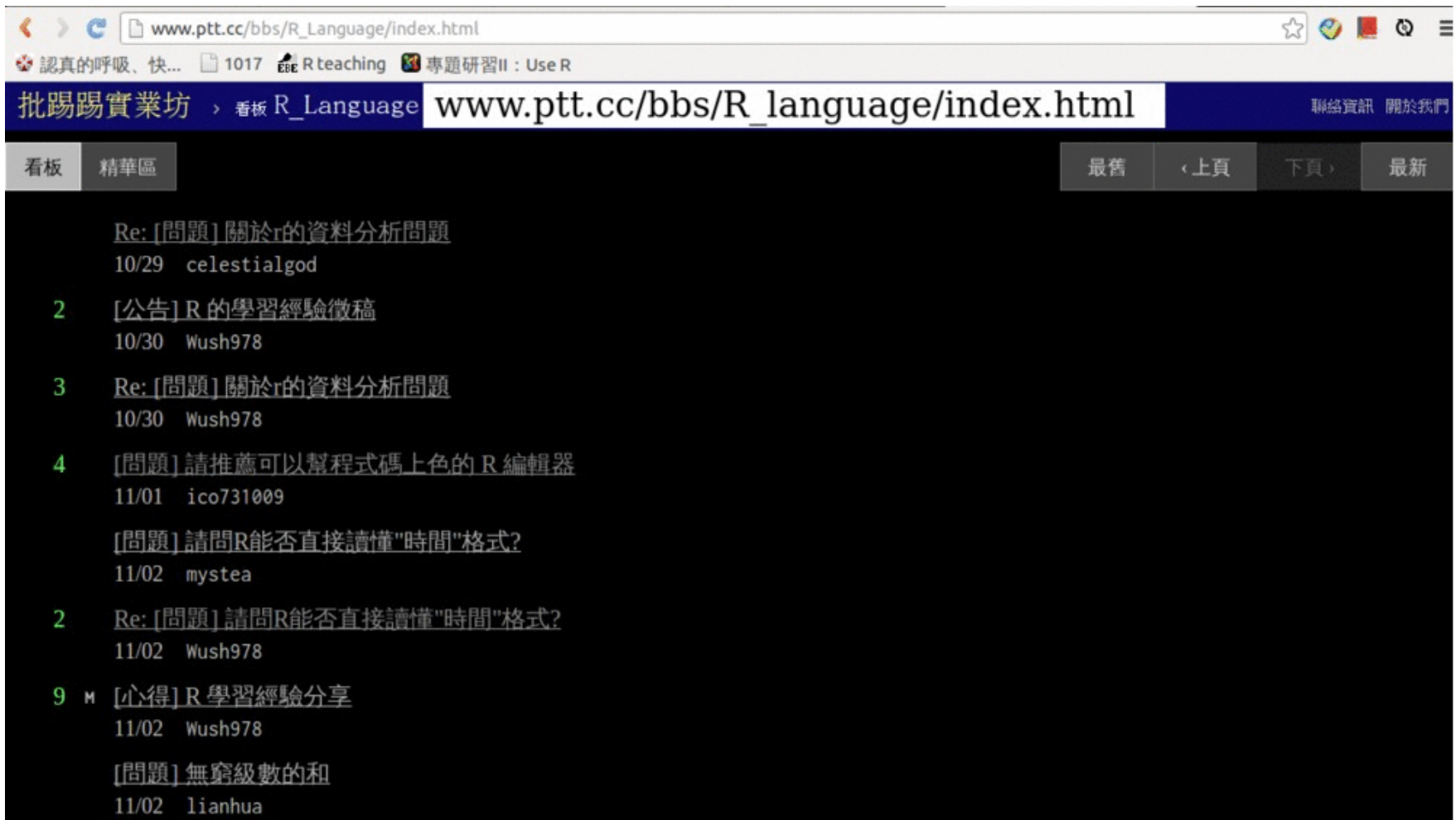
I want to crawl into the next page..

The Structure of a Spider



Make your spider run
his own way!

Observe the “RULE” of target URL !



The screenshot shows a web browser window displaying the Ptt forum page for the R_Language board. The address bar shows the URL www.ptt.cc/bbs/R_Language/index.html. The page header includes the Ptt logo, the board name "批踢踢實業坊", and the board name "R_Language". The main content area lists several forum posts with their titles, dates, and authors. The posts are numbered 2, 3, 4, 2, and 9. The titles of the posts are: "Re: [問題] 關於r的資料分析問題", "[公告] R 的學習經驗徵稿", "Re: [問題] 關於r的資料分析問題", "[問題] 請推薦可以幫程式碼上色的 R 編輯器", "[問題] 請問R能否直接讀懂"時間"格式?", "Re: [問題] 請問R能否直接讀懂"時間"格式?", "[心得] R 學習經驗分享", and "[問題] 無窮級數的和". The authors of the posts are celestialgod, Wush978, ico731009, mystea, Wush978, and lianhua.

認真的呼吸、快... 1017 R teaching 專題研習II : Use R

批踢踢實業坊 > 看板 R_Language www.ptt.cc/bbs/R_language/index.html 聯絡資訊 關於我們

看板 精華區 最舊 < 上頁 下頁 > 最新

Re: [\[問題\] 關於r的資料分析問題](#)
10/29 celestialgod

2 [\[公告\] R 的學習經驗徵稿](#)
10/30 Wush978

3 [Re: \[問題\] 關於r的資料分析問題](#)
10/30 Wush978

4 [\[問題\] 請推薦可以幫程式碼上色的 R 編輯器](#)
11/01 ico731009

[\[問題\] 請問R能否直接讀懂"時間"格式?](#)
11/02 mystea

2 [Re: \[問題\] 請問R能否直接讀懂"時間"格式?](#)
11/02 Wush978

9 [\[心得\] R 學習經驗分享](#)
11/02 Wush978

[\[問題\] 無窮級數的和](#)
11/02 lianhua

Ex: Scrap across pages

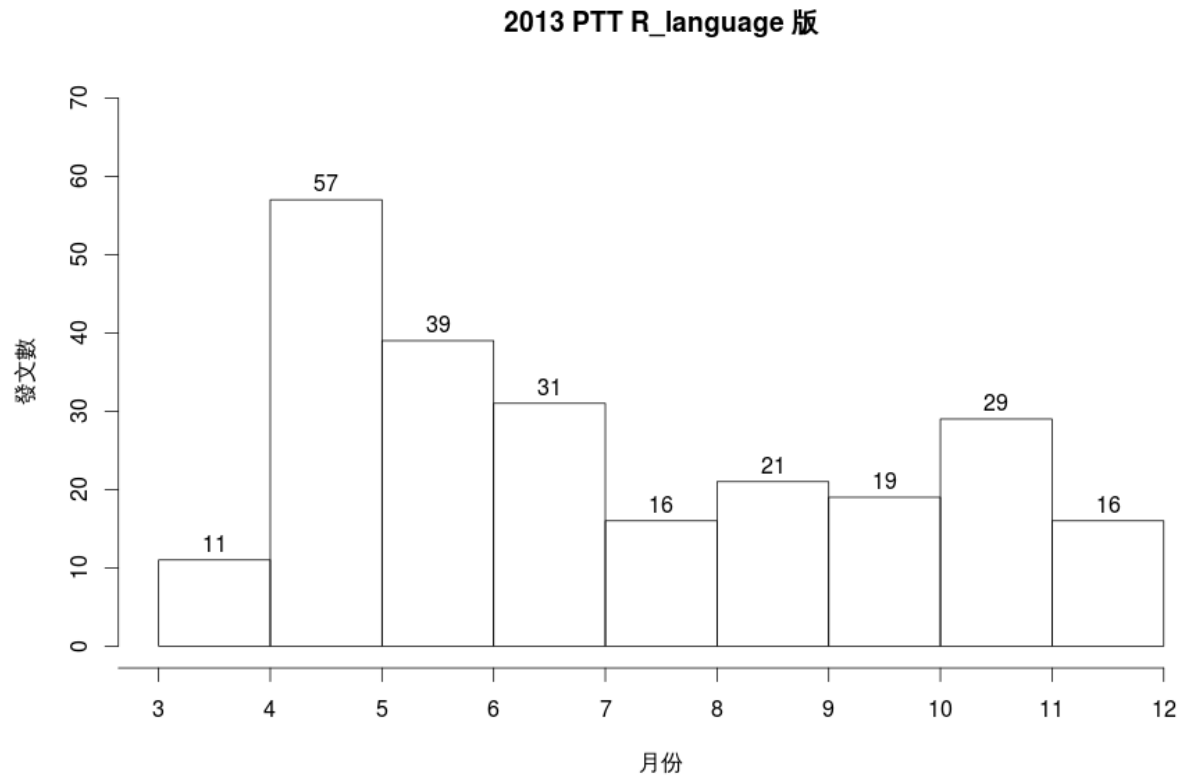
```
# install RCurl, XML
extract = function(xml, Xpath){
  value = unlist(xml[Xpath])
  if (length(value) == 0) return(NA)
  if (class(value) != "character" ) return(sapply(xml[Xpath], xmlValue))
  return(value) }

getPageData = function(url){
  html = getURL(url)
  xml = htmlParse(html)
  page_data= data.frame()
  page_data = data.frame( title = extract(xml, Xpath = "//div[@class='title']/a/text()",
    author= extract(xml, Xpath = "//div[@class='author']/text()",
    href = paste("http://www.ptt.cc",extract(xml, Xpath = "//div[@class='title']/a/@href"),
    sep=""),
    date = extract(xml, Xpath = "//div[@class='date']"))
  page_data$push = sapply(1:nrow(page_data),
    function(x) extract(xml, Xpath = sprintf("//div[@class='r-ent'][%d]/span[@class='hl f2']",x)))
  return(page_data)
}

data = data.frame()
for (i in 1:12){
  url = sprintf("http://www.ptt.cc/bbs/R_Language/index%s.html", i )
  data = rbind(data, getPageData(url))
}
```

	title	author	href	date	push
1	Fw: [公告] R_Language 成立試閱	cleanwind	http://www.ptt.cc/bbs/R_Language/M.1364407184.A.FD6.html	3/28	1
2	[公告] 初版版務相關事項	Wush978	http://www.ptt.cc/bbs/R_Language/M.1364448745.A.011.html	3/28	8
3	[情報] Moving to R 3.0.0 on Ubuntu	Wush978	http://www.ptt.cc/bbs/R_Language/M.1364486158.A.1B2.html	3/28	1
4	[問題] 如何建構這個問題的程式	MIZUYAMA	http://www.ptt.cc/bbs/R_Language/M.1364517882.A.408.html	3/29	5
5	[分享] 簡易的建立及使用arima model	Wush978	http://www.ptt.cc/bbs/R_Language/M.1364561519.A.384.html	3/29	NA
6	Fw: [筆記] R的設定-平行運算 (CPU)	gsuper	http://www.ptt.cc/bbs/R_Language/M.1364646766.A.115.html	3/30	1
7	Fw: [筆記] R的字串處理	gsuper	http://www.ptt.cc/bbs/R_Language/M.1364646786.A.452.html	3/30	2
8	Fw: [程式] 寫出含文字的函數?	gsuper	http://www.ptt.cc/bbs/R_Language/M.1364646926.A.0C2.html	3/30	NA
9	Fw: [程式] a[order(a[,1])], 排序與延伸選取範圍	gsuper	http://www.ptt.cc/bbs/R_Language/M.1364646940.A.882.html	3/30	NA
10	[分享] Lots of data != Big Data	Wush978	http://www.ptt.cc/bbs/R_Language/M.1364720206.A.CE5.html	3/31	NA
11	[情報] 20130401 MLDM Monday 會前通知	Cayley	http://www.ptt.cc/bbs/R_Language/M.1364741273.A.DE5.html	3/31	2
12	Fw: [程式] R跑出來的結果如何存檔?	andrew43	http://www.ptt.cc/bbs/R_Language/M.1364776498.A.F27.html	4/01	NA
13	Fw: [問題] 請問R語言	andrew43	http://www.ptt.cc/bbs/R_Language/M.1364776508.A.EC3.html	4/01	NA
14	Fw: [問題] 迴歸模型類別變數之綜合檢定	andrew43	http://www.ptt.cc/bbs/R_Language/M.1364776522.A.6CF.html	4/01	NA
15	Fw: [程式] 以 R 運算常見的檢定	andrew43	http://www.ptt.cc/bbs/R_Language/M.1364776535.A.CBB.html	4/01	NA
16	Fw: [問題] 請問R的中文書	andrew43	http://www.ptt.cc/bbs/R_Language/M.1364776569.A.C62.html	4/01	2
17	Fw: [情報] R 演習室開張了	andrew43	http://www.ptt.cc/bbs/R_Language/M.1364776576.A.925.html	4/01	6
18	[分享] 十個給新手的R tips	Wush978	http://www.ptt.cc/bbs/R_Language/M.1364915485.A.3DC.html	4/02	1
19	Re: [問題] 請問R的中文書	Wush978	http://www.ptt.cc/bbs/R_Language/M.1364920563.A.CA9.html	4/03	NA
20	Re: [問題] 請問R的中文書	andrew43	http://www.ptt.cc/bbs/R_Language/M.1364926499.A.A5B.html	4/03	1
21	Re: [問題] 請問R的中文書	Yukirin	http://www.ptt.cc/bbs/R_Language/M.1364963864.A.2AC.html	4/03	5
22	Re: [問題] 請問R的中文書	Cayley	http://www.ptt.cc/bbs/R_Language/M.1364998897.A.8AE.html	4/03	NA
23	[問題] R 關於例外處理	ckkt	http://www.ptt.cc/bbs/R_Language/M.1365084660.A.A57.html	4/04	1
24	Re: [問題] R 關於例外處理	andrew43	http://www.ptt.cc/bbs/R_Language/M.1365085670.A.F01.html	4/04	1
25	[心得] R 的例外處理	Wush978	http://www.ptt.cc/bbs/R_Language/M.1365090307.A.8E8.html	4/04	1
26	[情報] 20130408 MLDM Monday 會前通知	Cayley	http://www.ptt.cc/bbs/R_Language/M.1365091090.A.98D.html	4/04	NA
27	[問題] 兩組資料合併	MIZUYAMA	http://www.ptt.cc/bbs/R_Language/M.1365220507.A.F33.html	4/06	2
28	[問題] 生成隨機數據矩陣	goddirk	http://www.ptt.cc/bbs/R_Language/M.1365348858.A.77F.html	4/07	NA
29	Re: [問題] 生成隨機數據矩陣	Yukirin	http://www.ptt.cc/bbs/R_Language/M.1365350876.A.531.html	4/08	NA

It shows...



author	count
Wush978	70
celestialgod	18
andrew43	16
EXILESPACER	12
DrRd	7
gsuper	6
MIZUYAMA	6
youngce	5
cog5566	5
Cayley	4

Summary

To write a basic spider...

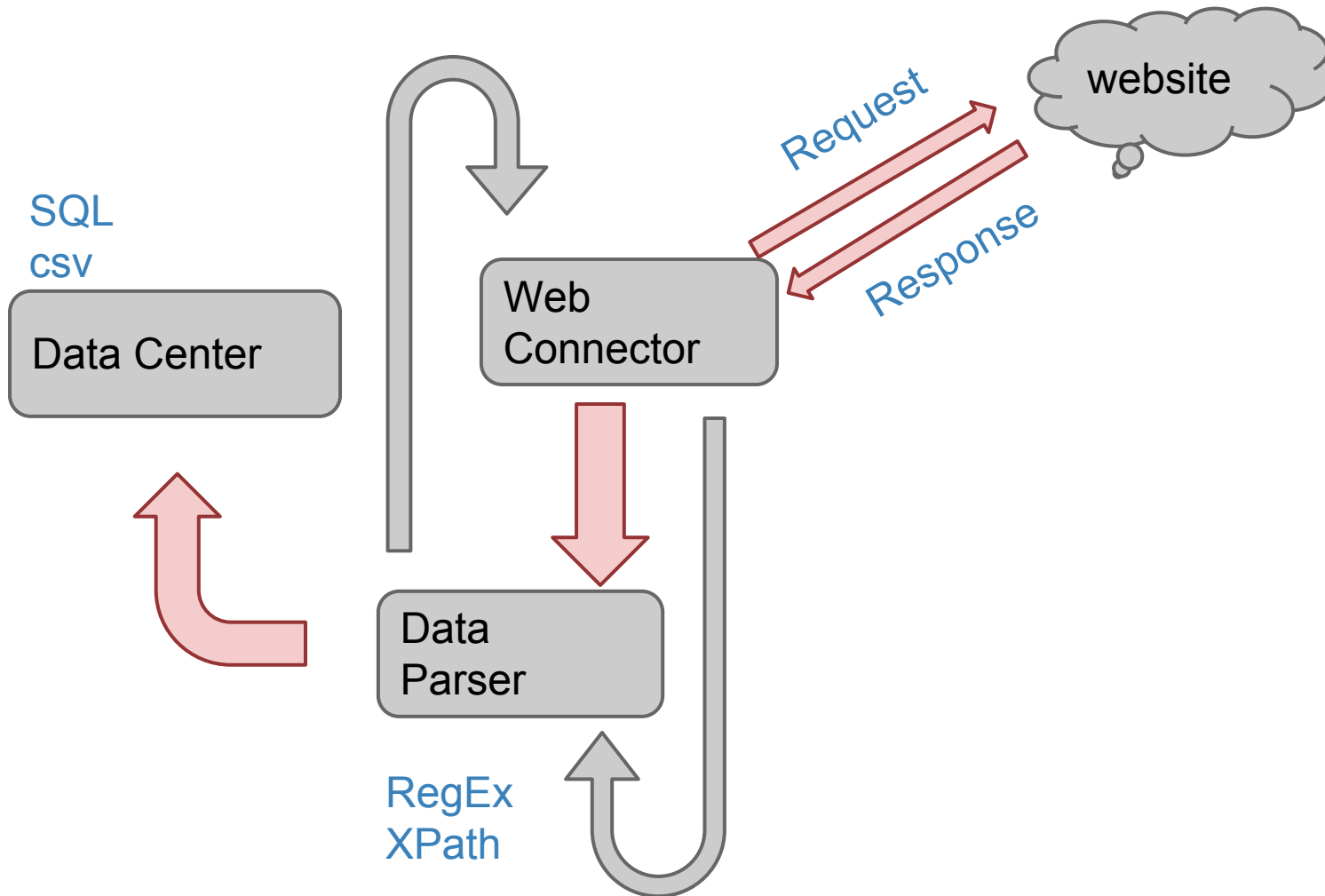
Step 1: step into the right webpage (connector)

Step 2: scrap the info you want (parser)

more detail...

1. Observe the URL and decide what to request
2. Make a request
3. Parse the response
4. Store results

The Structure of a Spider



Some Advanced Suggestions...

What's the reality?

Sometimes, you will encounter
a more complicated case...

- ❑ Fail to connect
- ❑ Fail to parse
- ❑ Try Catch
- ❑ Legal Issue

Fail to connect : Cookies / Session

你有通關密碼嗎...



Fail to connect : Cookies / Session

Elements	Resources	Network	Sources	Timeline	Profiles	Audits	Console
▼ Frames							
(index.html)							
▶ Web SQL							
IndexedDB							
Local Storage							
Session Storage							
▼ Cookies							
www.ptt.cc							
▶ Application Cache							

Name	Value	Domain	Path	Expires / ...	Size	HTTP	Secure
__utma	156441338.1678767527.1384401096.1384532139.1384...	.ptt.cc	/	Sun, 15 ...	62		
__utmb	156441338.1.10.1384540620	.ptt.cc	/	Fri, 15 N...	31		
__utmc	156441338	.ptt.cc	/	Session	15		
__utmz	156441338.1384532139.12.3.utmcsr=google utmccn=(or...	.ptt.cc	/	Sat, 17 ...	101		

RCurl:: getCurlHandle

(pseudo code)

```
cookie = 'cookiefile.txt'
```

```
handler = getCurlHandle(cookiefile=cookie, cookiejar=cookie)
```

```
data = getURL(url , curl=handler )
```

Fail to connect: next url ?

www.taifex.com.tw/chinese/3/fcm_opt_rep.asp

認真的呼吸、快... 1017 EBE R teaching 專題研習II : Use R

English / 網站地圖 / 回首頁 / 電子報訂閱

臺灣期貨交易所
TAIWAN FUTURES EXCHANGE

搜尋

公司簡介 商品 交易資訊 交易制度 結算業務 法令規章 統計資料 期貨業專區 交易人服務與保護 出版與研究 最新消息 外資專區 相關網站

榮耀傳承
期交所專注期貨領域 保障交易權益

首頁 > 交易資訊 > 盤後資訊 > 期貨商買賣日報表 > 選擇權

交易資訊

- 盤後資訊
 - 期貨
 - 選擇權
 - 鉅額交易
- 期貨商買賣日報表
 - 期貨
 - 選擇權
 - 鉅額交易
- 三大法人
- 大額交易人未沖銷部位結構
- 每日外幣參考匯率查詢
- 交易歷史資料申請
- 資料下載專區

選擇權

期貨商買賣日報表 (資料日期:2013/11/15)

契約: 請選擇

到期月份(選別): 請選擇

買/賣權: 請選擇

送出查詢

•本查詢功能僅提供期貨市場當日交易資料
•資料產製時間: 約每交易日下午4時

友答列印

query string:

http://www.taifex.com.tw/chinese/3/fcm_opt_rep.asp?

commodity_id=TXO & commodity_idt=TXO &

settlemon=201310W1 & pccode=P &

curpage=1 & pagenum=1

counter_flag:	1
curpage:	1
qtype:	2
commodity_id:	TXO
commodity_id2:	
commodity_name:	臺指選擇權(TXO)
downloadflag:	
qflag:	
commodity_idt:	TXO
commodity_id2t:	
settlemon:	201310W1
pccode:	P

www.taifex.com.tw/chinese/3/fcm_opt_rep.asp

交易資訊

- 盤後資訊
- 期貨
- 選擇權
- 鉅額交易
- 期貨商買賣日報表
- 期貨
- 選擇權
- 鉅額交易
- 三大法人
- 大額交易者未沖銷部位結構
- 每日外幣參考匯率查詢
- 交易歷史資料申請
- 資料下載專區
- 資訊廠商
- 台灣期貨交易所行情資訊網站

選擇權

期貨商買賣日報表 (資料日期: 2013/09/27)

契約: 臺指選擇權(TXO)

到期月份(週別): 201310

買/賣權: 賣權(Put)

送出查詢

•本查詢功能僅提供期貨市場當日交易資料
•資料產製時間: 約每交易日下午4時

交易日期	20130927	契約代號	TXO	契約名稱	臺指選擇權	到期月份(週別)	201310	買/賣權	賣權	成交口數	42,295
履約價格	成交價格	全市場成交量	買進期貨商代號	買進期貨商名稱	賣出期貨商代號	賣出期貨商名稱					
6600	0.1	20	B224	中國信託商業銀行	B224	中國信託商業銀行					
6600	0.3	1	F002	永豐期貨	B224	中國信託商業銀行					

Resources Network Sources Timeline Profiles Audits Console

Name Path

- fcmm_opt_rep.asp/chinese/3
- rich_calendar.css/chinese/js/rich_calendar
- rich_calendar.js/chinese/js/rich_calendar
- rc_lang_en.js/chinese/js/rich_calendar
- rich_calendar_customize/chinese/js
- global.css/chinese/css
- jquery-1.4.2.min.js/chinese/js

Headers Preview Response Cookies Timing

User-Agent: Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.22 (KHTML, like Gecko) Chrome/25.0.1364.172 Safari/537.22

Form Data

counter_flag: 1
curpage: 1
qtype: 2
commodity_id: TXO
commodity_id2:
commodity_name: 臺指選擇權(TXO)
downloadflag:
qflag:
commodity_idt: TXO
commodity_id2t:
settlemon: 201310
pccode: P

Response Headers

Cache-control: private
Content-Length: 100976

Fail to parse : Encoding

I HATE 'ENCODING'!!!

CHARACTER ENCODING / 字元編碼 / 字符編碼

我愛你

UTF-8 (檔首無 BOM): 我愛你

Big5: ?[]?霏?

GB2312: 鋁憂副浣



Try Catch

優雅的出錯～

(pseudo code)

```
tryCatch( getPageData(url),  
          error = function(e) e,  
          finally=print("Hello")) # 優雅的跳出 ^.<
```

Legal Issue

Hi,

May I come in ?

Legal Issue

robots.txt



www.amazon.com/robots.txt

```
User-agent: *
Disallow: /exec/obidos/account-access-login
Disallow: /exec/obidos/change-style
Disallow: /exec/obidos/flex-sign-in
Disallow: /exec/obidos/handle-buy-box
Disallow: /exec/obidos/tg/cm/member/
Disallow: /gp/cart
Disallow: /gp/flex
Disallow: /gp/product/e-mail-friend
Disallow: /gp/product/product-availability
Disallow: /gp/product/rate-this-item
Disallow: /gp/sign-in
Disallow: /gp/reader
Disallow: /gp/sitbv3/reader
Disallow: /gp/richpub/syltguides/create
Disallow: /gp/gfix
Disallow: /gp/associations/wizard.html
Disallow: /gp/dmusic/order
Disallow: /gp/legacy-handle-buy-box.html
Disallow: /gp/aws/ssop
Disallow: /gp/yourstore
Disallow: /gp/gift-central/organizer/add-wishlist
Disallow: /gp/vote
Disallow: /gp/voting/
```

```
Disallow: /gp/music/wma-pop-up
Disallow: /gp/customer-images
Disallow: /gp/richpub/listmania/createpipeline
Disallow: /gp/content-form
Disallow: /gp/pdp/invitation/invite
Disallow: /gp/customer-reviews/common/du
Disallow: /gp/customer-reviews/write-a-review.html
Disallow: /gp/associations/wizard.html
Disallow: /gp/music/clipserve
Disallow: /gp/offer-listing
Disallow: /gp/customer-media/upload
Disallow: /gp/history
Disallow: /gp/item-dispatch
Disallow: /gp/dmusic/order/handle-buy-box.html
Disallow: /gp/recsradio
Disallow: /gp/slredirect
Disallow: /dp/shipping/
Disallow: /dp/twister-update/
Disallow: /dp/manual-submit/
```

www.ntu.edu.tw/robots.txt



```
User-Agent: Googlebot
Disallow: /president/
```

Legal Issue (policy)



維基百科

自由的百科全書

首頁

分類索引

特色內容

新聞動態

最近更新

隨機條目

▼ 幫助

幫助

社群入口

方針與指引

互助客棧

詢問處

字詞轉換

IRC即時聊天

聯繫我們

關於維基百科

資助維基百科

► 工具箱

▼ 其他語言

Alemannisch

العربية

條目

討論

台灣正體

漢

漢

閱讀

編輯

檢視歷史

搜尋

搜尋維基百科[alt-shift]

歡迎報名參與Wiki協作聚-台中場教學講座活動，體驗改變維基百科的樂趣。

[關閉]

神鬼奇航：鬼盜船魔咒

[編輯]

維基百科，自由的百科全書

(重定向自神鬼奇航)

《**神鬼奇航：鬼盜船魔咒**》（**英語：****Pirates of the Caribbean: The Curse of the Black Pearl**）是一部2003年的歷險奇幻電影，根據**迪士尼主題公園**的同名景點製作，由**戈爾·維賓斯基**執導，**傑瑞·布魯克海默**擔任製片人。主要講述了一位名叫威爾·特納（Will Turner，**奧蘭多·布魯姆**）的年輕鐵匠和**海盜傑克·斯派羅**船長（**強尼·戴普**飾）一起前去營救遭到綁架的伊莉莎白·斯旺（Elizabeth Swann，**凱拉·奈特莉**飾），綁架她的是由哈克特·巴博沙（Hector Barbossa）船長帶領的「黑珍珠號」海盜船成員，而斯派羅之前正是該船的船長。

傑·沃爾伯特於2001年根據主題公園景點開發了一個劇本，斯圖爾特·貝亞蒂耶於2002年初予以改寫。製片人傑瑞·布魯克海默加入了這個項目，請來特里·魯西奧和泰德·艾略特對劇本進行進一步改寫，在故事線索中增加了超自然**詛咒**的內容。影片於2002年10月至2003年3月在**聖文森及格瑞那丁**等地進行拍攝，許多外景都是在**洛杉磯**周邊地區建成。電影的全球首映式於2003年6月28日在**美國加利福尼亞州阿納海姆**的**迪士尼樂園度假區**舉行。由於商業上獲得了很大的成功，電影發展成為**神鬼奇航**系列電影的開山之作，首部續集《**神鬼奇航2：加勒比海盜**》於2006年上映，第二部續集《**神鬼奇航3：世界的盡頭**》則於2007年上映。2011年，電影的第3部續集《**神鬼奇航4：驚濤怪浪**》也予以上映。

影片獲得了影評人的好評，商業上更有著出人意料的成功，全球票房超過6.54億美元。強尼·戴普扮演的**傑克·斯派羅**獲得了普遍讚譽，為他贏得了**美國演員工會獎最佳男主角獎**，並提名**奧斯卡最佳男主角獎**、**英國電影學院獎最佳男主角獎**以及**金球獎最佳音樂及喜劇類電影男主角**。此外，《神鬼奇航：鬼盜船魔咒》還獲得了**奧斯卡金像獎**的其他4項提名，並在**英國電影學院獎**上有所斬獲。

神鬼奇航：鬼盜船魔咒

Pirates of the Caribbean: The Curse of the Black Pearl



本頁面最後修訂於**2013年11月2日（星期六）15:23**。

本站的全部文字在**創用CC 姓名標示-相同方式分享 3.0 協議**之條款下提供，附加條款亦可能應用。（請參閱**使用條款**）
Wikipedia®和**維基百科**標誌是**維基媒體基金會**的註冊商標；維基™是維基媒體基金會的商標。
維基媒體基金會是在**美國佛羅里達州**登記的**501(c)(3)免稅**、**非營利**、**慈善機構**。

[隱私政策](#) [關於維基百科](#) [免責聲明](#) [開發人員](#) [手機版](#)

Legal Issue

Solution (skip ?)

Act like a human being

1. pause [R] Sys.sleep(SleepSeconds)
2. off-hour ? rush-hour

Happy time is always flying.
It's time to make an end.

In Brief

R might not be so serious.

It can be very close to your daily life.

For example, SpideR:

分析網站資料, 網拍監控

Let's start your first **spider**,
and have fun with **R** :)

Thank you :)

Extended Reading

You might be also interested in

c3h3

https://www.youtube.com/watch?v=P3Xm_JFmh04

https://www.youtube.com/watch?v=Sr6JLjgX_30

https://www.youtube.com/watch?v=MI_Qlc-mjy0

Ronny Wang

<https://www.youtube.com/watch?v=qmtgeaajcew>

EC Lee

<https://www.youtube.com/watch?v=ixEz4GpTP5g>

[v=ixEz4GpTP5g&list=PLM7HGQkDNOHsaFMMbABbMCdMy7oVdVxxg](https://www.youtube.com/watch?v=ixEz4GpTP5g&list=PLM7HGQkDNOHsaFMMbABbMCdMy7oVdVxxg)

1 encoding