



國立臺灣大學

National Taiwan University

使用R語言進行資料分析 Using R for Data Analysis

國立臺灣大學共同教育中心

助理教授 蔡芸琇

Chapter 06

» 網路爬蟲

爬蟲基本概念介紹

- ▶ 網頁的內容是由 HTML+CSS+JavaScript 組合成。
- ▶ HTML+CSS+JavaScript 透過瀏覽器編譯後呈現給使用者。
- ▶ 爬蟲是自動抓網頁內容 HTML+CSS+JavaScript 的程式。
- ▶ 爬蟲透過網址 (URL : Uniform Resource Locator) 提取網頁內容。

URL 的格式

▶ URL 分成三部分：

<http://www.ntu.edu.tw/about/about.html>

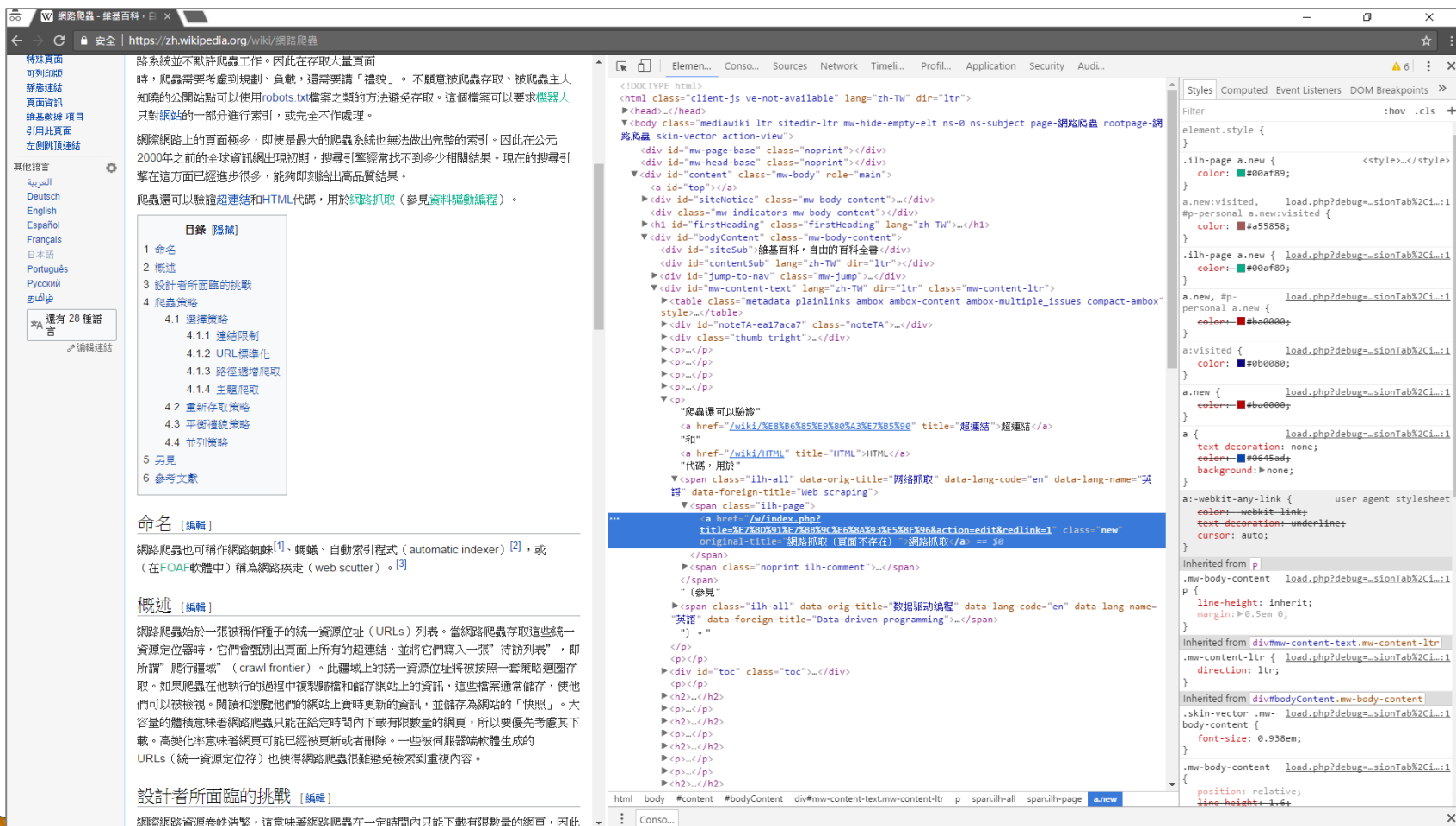
- 第一部分是協議 (http、https)。
- 第二部分是存有該資源的主機位址 (140.112.8.116、www.ntu.edu.tw)。
- 第三部分是主機上的具體目錄或文件名等 (about/about.html)。
- 第一部分和第二部分用「://」符號隔開。
- 第二部分和第三部分用「/」符號隔開。
- 第一部分和第二部分不可缺少，第三部分看實際狀況決定是否加上。

網頁基本架構介紹



<http://httpbin.org/>
<https://www.w3.org/Protocols/rfc2616/rfc2616.html>

檢視原始碼



[illegible]

開發人員工具

- ▶ 在網頁任何位置按右鍵，選擇“檢查”，就可以看到相對應的原始碼。
- ▶ **Network** 頁面，可以看到網頁各項的執行細節。
- ▶ **Console** 頁面，可以檢查錯誤訊息。
- ▶ 直接對 **CSS** 樣式表更改參數，畫面就會直接可以預覽。
- ▶ 爬蟲觀察技巧：<http://tech-marsw.logdown.com/blog/2016/01/10/crawler-tips-mining-chrome>。

網頁架構及語法

- <標籤>內容</標籤>

```
<html>
  <head>
  </head>
  <body>
    <h1> 標題 </h1>
    <p> 段落 </p>
    <ol>
      <li> 項目內容 1 </li>
      <li> 項目內容 2 </li>
    </ol>
  </body>
</html>
```

網路爬蟲文本蒐集

看板 NTUcourse 文章列 x

安全 | <https://www.ptt.cc/bbs/NTUcourse/index.html>

批踢踢實業坊 > 看板 NTUcourse 聯絡資訊 關於我們

看板 精華區 最舊 < 上頁 下頁 > 最新

a | 253.31 x 17.96

[\[求救\] 普通生物學丙 考古題\(李鳳鳴\)](#)
4/16 handfox

[\[問題\] 請問簡坤鐘 \(二\) 34高爾夫球](#)
4/17 asiguo

[\[評價\]105-2 丁亮 文字學乙下](#)
4/17 nancy3nancy3

[\[問題\] 請問史前史二期中考](#)
4/17 sophieku

[\[求救\] 李怡庭 貨幣銀行學 期中範圍](#)

Elements

```
<!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
        action-bar-margin bbs-screen">
        <div class="r-ent">
          <div class="nrec"></div>
          <div class="mark"></div>
          <div class="title">
            <a href="/bbs/NTUcourse/
              M.1492329082.A.1D4.html"
              >[求救] 普通生物學丙 考古
                題(李鳳鳴)</a> == $0
          </div>
          <div class="meta">...</div>
        </div>
        <div class="r-ent">...</div>
        <div class="r-ent">...</div>
        <div class="r-ent">...</div>
        <div class="r-ent">...</div>
      </div>
    </div>
  </body>
</html>
```

Styles Event Listeners DOM Breakpoints

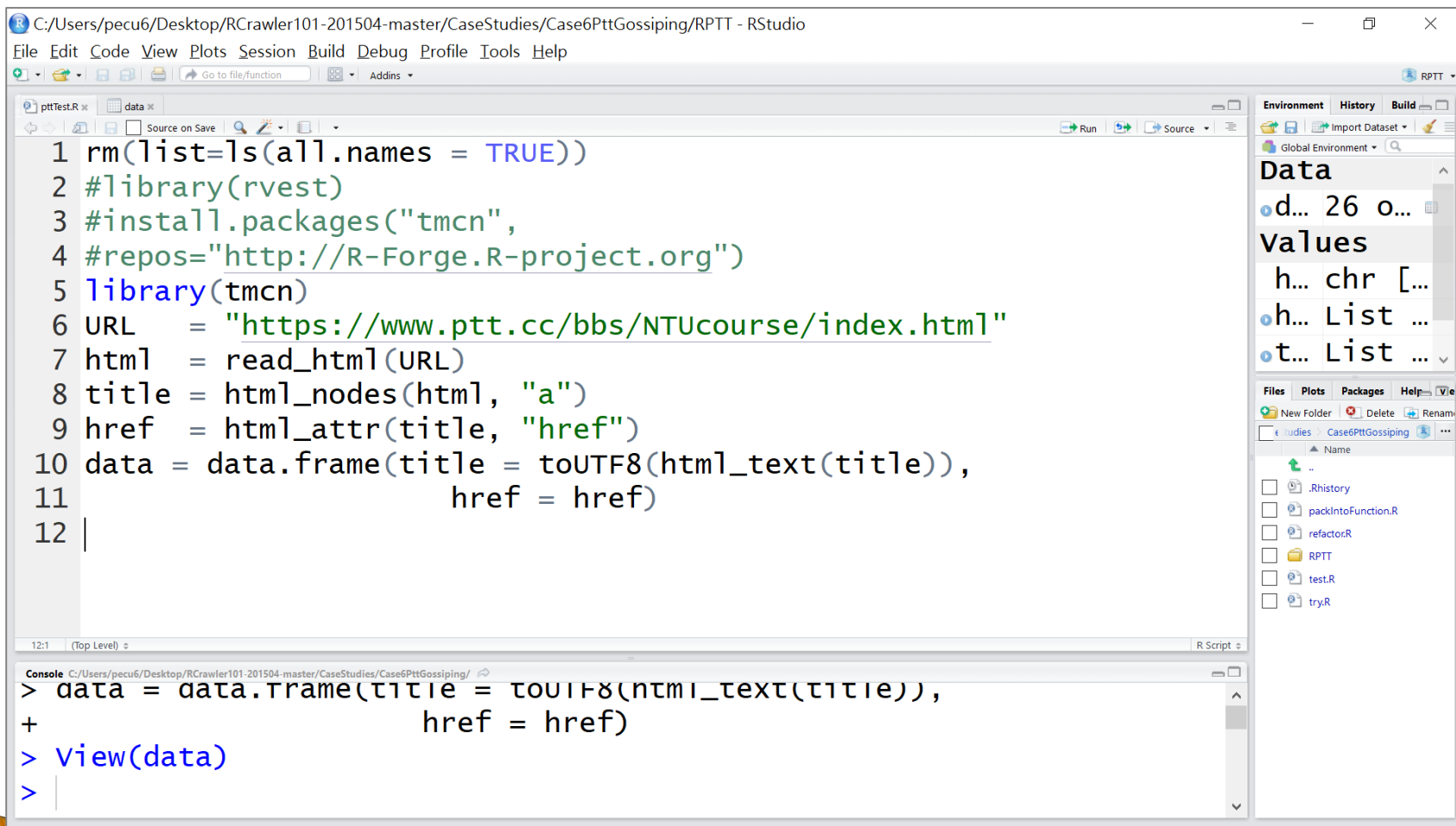
:hov .cls +

element.st
yle {
}
bbs-base.c.c
a:visited

margin -
border -
padding -
auto x auto

Console

網路爬蟲文本蒐集



The screenshot shows an RStudio interface with the following components:

- Menu Bar:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Includes icons for file operations, running code, and adding packages.
- Source Editor:** Contains R code for scraping a PTT forum page. The code uses the `tm` and `rvest` packages to fetch and parse HTML data.
- Environment Panel:** Shows the 'Global Environment' with a 'Data' section containing a list of values, including character strings and lists.
- Files Panel:** Displays the project file structure, including files like `.Rhistory`, `packIntoFunction.R`, `refactor.R`, `RPTT`, `test.R`, and `try.R`.
- Console:** Shows the execution of the code, with the final output being a data frame with columns for title and href.

```
1 rm(list=ls(all.names = TRUE))
2 #library(rvest)
3 #install.packages("tmcn",
4 #repos="http://R-Forge.R-project.org")
5 library(tmcn)
6 URL = "https://www.ptt.cc/bbs/NTUcourse/index.html"
7 html = read_html(URL)
8 title = html_nodes(html, "a")
9 href = html_attr(title, "href")
10 data = data.frame(title = toUTF8(html_text(title)),
11 href = href)
12 |
```

```
> data = data.frame(title = toUTF8(html_text(title)),
+ href = href)
> view(data)
> |
```

網路爬蟲文本蒐集

C:/Users/pecu6/Desktop/RCrawler101-201504-master/CaseStudies/Case6PttGossiping/RPTT - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

pttTest.R x data x

Filter

	title	href
1	批踢踢實業坊	/
2	看板 NTUcourse	/bbs/NTUcourse/index.html
3	關於我們	/about.html
4	聯絡資訊	/contact.html
5	看板	/bbs/NTUcourse/index.html
6	精華區	/man/NTUcourse/index.html
7	最舊	/bbs/NTUcourse/index1.html
8	<U+2039> 上頁	/bbs/NTUcourse/index1191.html
9	下頁 <U+203A>	/NA
10	最新	/bbs/NTUcourse/index.html
11	[求助] 普通生物學丙 考古題(李鳳鳴)	/bbs/NTUcourse/M.1492329082.A.1D4.html
12	[問題] 請問關坤鐘 (二) 34高爾夫球	/bbs/NTUcourse/M.1492364474.A.141.html
13	[評價] 105-2 丁亮 文字學乙下	/bbs/NTUcourse/M.1492410691.A.52B.html
14	[問題] 請問史前史二期中考	/bbs/NTUcourse/M.1492421244.A.11B.html
15	[求助] 李怡庭 貨幣銀行學 期中範圍	/bbs/NTUcourse/M.1492423076.A.E3E.html
16	[問題] 張亞中教授 國關二期中考題	/bbs/NTUcourse/M.1492439974.A.BC4.html
17	[問題] 請問吳英傑老師英美法名著這周要上課嗎?	/bbs/NTUcourse/M.1492525157.A.1D5.html
18	[求助] 4/14 (五) 黃昭元老師國際公法錄音檔	/bbs/NTUcourse/M.1492537737.A.304.html
19	[問題] 海峽兩岸關係史二 李君山	/bbs/NTUcourse/M.1492585301.A.4F4.html
20	[問題] 一般醫學保健有幾名嗎?	/bbs/NTUcourse/M.1492591816.A.65C.html
21	[求助] 歐陽彥正現代科學與心靈科學期中考~	/bbs/NTUcourse/M.1492610720.A.1A3.html
22	[求助] 四人幫經濟學上下解答	/bbs/NTUcourse/M.1492693159.A.BA5.html

Showing 1 to 23 of 26 entries

Environment History Build

Global Environment

Data

d... 26 o...
values
h... chr [...
h... List ...
t... List ...

Files Plots Packages Help View

New Folder Delete Rename

studies Case6PttGossiping

Name

- ..
- .Rhistory
- packIntoFunction.R
- refactorR
- RPTT
- test.R
- try.R

Console

```
> data = data.frame(title = toupper(html_text(title)),  
+ href = href)  
> view(data)  
>
```

網路爬蟲文本蒐集

[求救] 普通生物學丙 考古

安全 | <https://www.ptt.cc/bbs/NTUcourse/M.1492329082.A.1D4.html>

批踢踢實業坊

看板 NTUcourse

聯絡資訊 關於我們

作者 handfox (handwolf)

標題 [求救] 普通生物學丙 考古題(李鳳鳴)

時間 Sun Apr 16 15:51:19 2017

如題

請問有沒有人有這門課的考古題可以借參考，範圍太大不知道該怎麼準備

酬勞可議

感謝！

※ 發信站：批踢踢實業坊(ptt.cc)，來自：223.136.95.125

※ 文章網址：
<https://www.ptt.cc/bbs/NTUcourse/M.1492329082.A.1D4.html>

返回看板

分享

Like 0

G+1

0

Elements

</div>

<div id="navigation-container">...

</div>

<div id="main-container">

<div id="main-content" class="bbs-screen bbs-content">

<div class="article-metaline">...</div>

<div class="article-metaline-right">...</div>

<div class="article-metaline">...</div>

<div class="article-metaline"> == \$0

時間

Sun Apr 16 15:51:19 2017

</div>

如題

請問有沒有人有這門課的考古題可以借參考，範圍太大不知道該怎麼準備

酬勞可議

感謝！

--

#main-content

div.article-metaline

Styles

Event Listeners

DOM Breakpoints

:hov .cls +

element.style {

yle {

}

bbs-base.c...

.article-

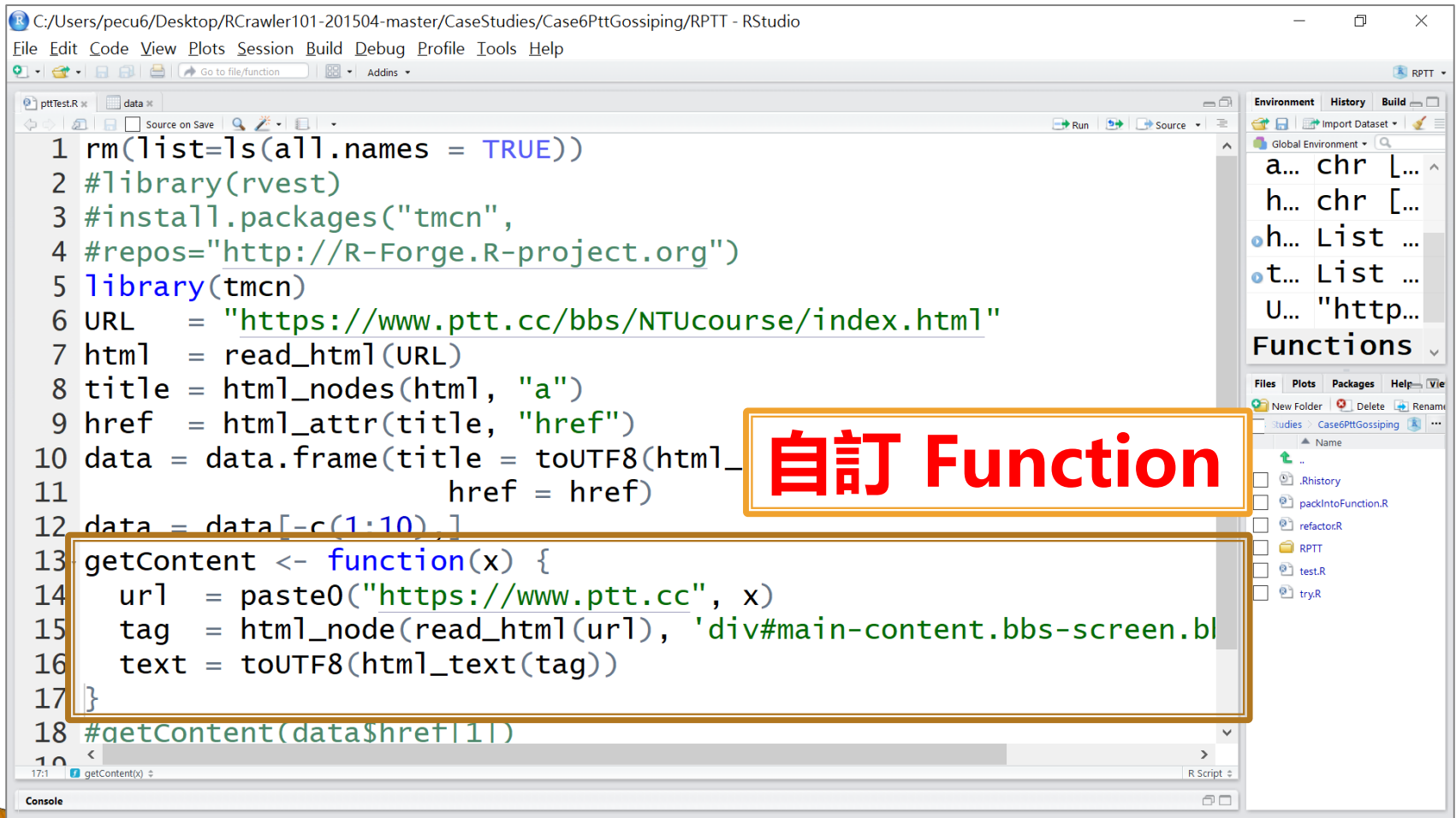
margin -

border -

padding -

480.327 x 16

網路爬蟲文本蒐集

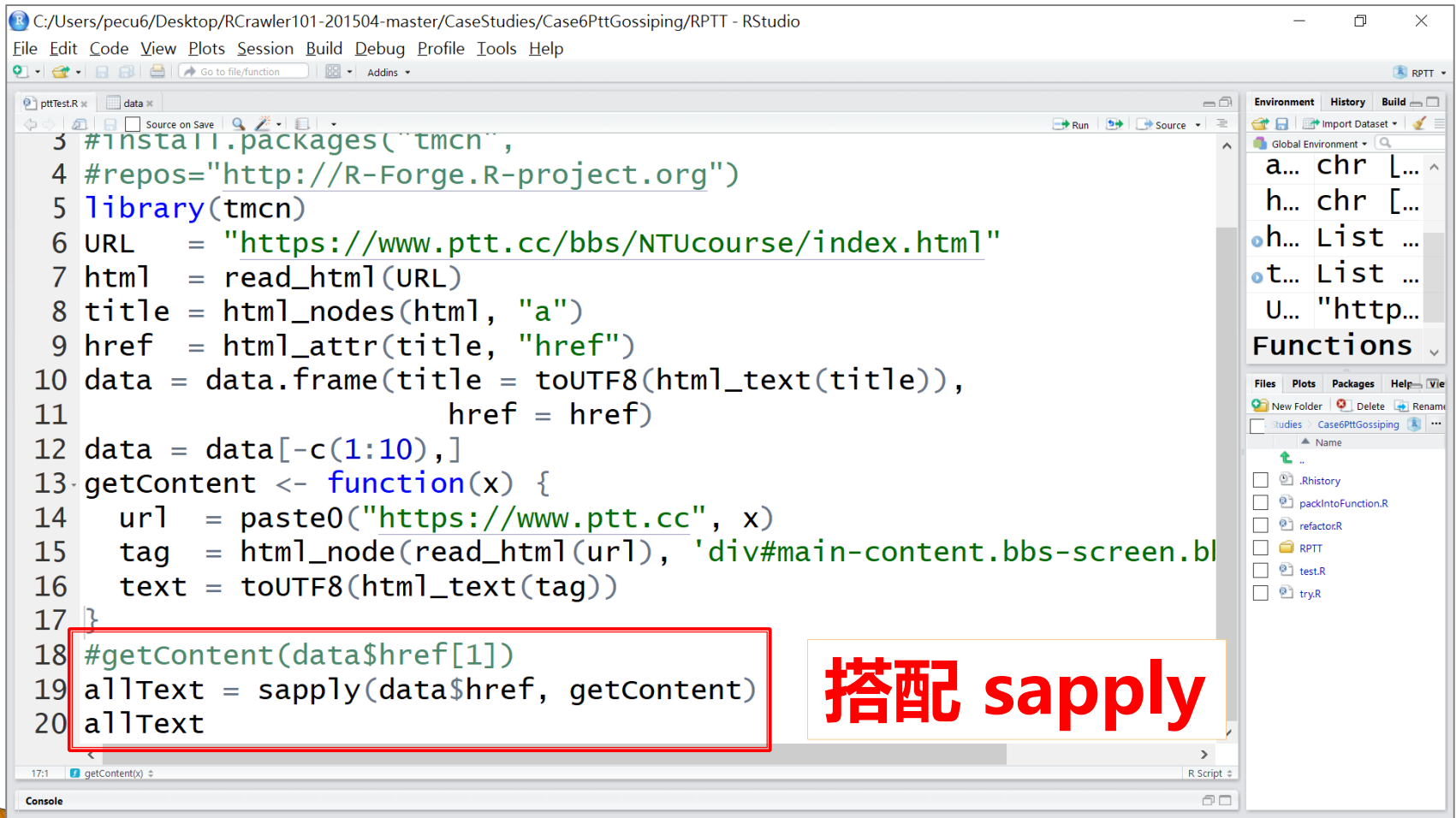


```
C:/Users/pecu6/Desktop/RCrawler101-201504-master/CaseStudies/Case6PttGossiping/RPTT - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
pttTest.R x data x
Source on Save
1 rm(list=ls(all.names = TRUE))
2 #library(rvest)
3 #install.packages("tmcn",
4 #repos="http://R-Forge.R-project.org")
5 library(tmcn)
6 URL = "https://www.ptt.cc/bbs/NTUcourse/index.html"
7 html = read_html(URL)
8 title = html_nodes(html, "a")
9 href = html_attr(title, "href")
10 data = data.frame(title = toUTF8(html_
11 href = href)
12 data = data[-c(1:10), ]
13 getContent <- function(x) {
14 url = paste0("https://www.ptt.cc", x)
15 tag = html_node(read_html(url), 'div#main-content.bbs-screen.bl
16 text = toUTF8(html_text(tag))
17 }
18 #getContent(data$href[1])
19 <
17:1 | getConten(x) |
Console
```

自訂 Function

Environment History Build
Global Environment
a... chr [...]
h... chr [...]
h... List ...
t... List ...
U... "http...
Functions
Files Plots Packages Help (View)
New Folder Delete Rename
studies Case6PttGossiping
Name
..
.Rhistory
packIntoFunction.R
refactorR
RPTT
test.R
try.R

網路爬蟲文本蒐集



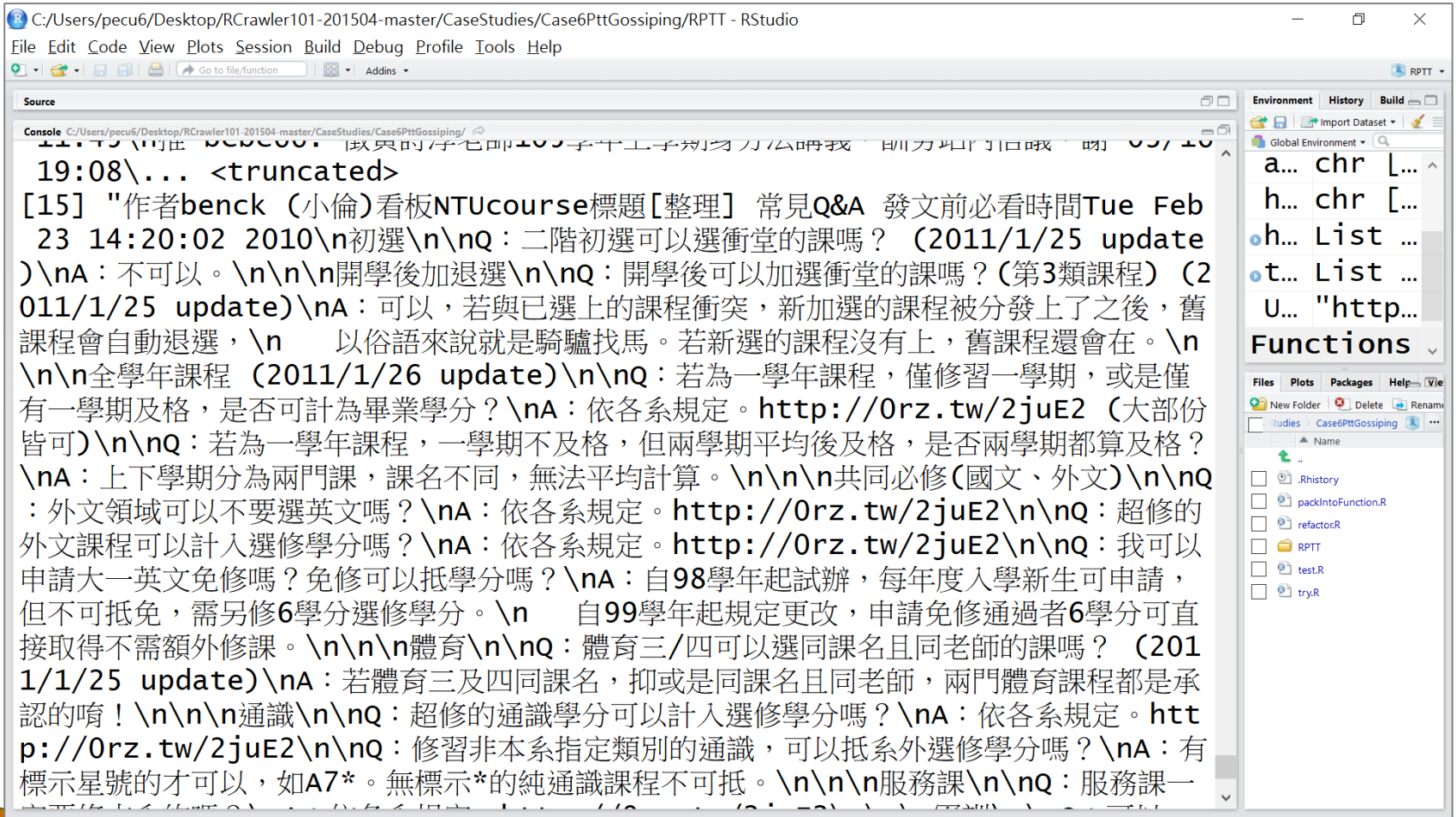
```
C:/Users/pecu6/Desktop/RCrawler101-201504-master/CaseStudies/Case6PttGossiping/RPTT - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
pttTest.R x data x
Source on Save Run Source
3 #install.packages("tmcn",
4 #repos="http://R-Forge.R-project.org")
5 library(tmcn)
6 URL = "https://www.ptt.cc/bbs/NTUcourse/index.html"
7 html = read_html(URL)
8 title = html_nodes(html, "a")
9 href = html_attr(title, "href")
10 data = data.frame(title = toUTF8(html_text(title)),
11 href = href)
12 data = data[-c(1:10),]
13 getContent <- function(x) {
14 url = paste0("https://www.ptt.cc", x)
15 tag = html_node(read_html(url), 'div#main-content.bbs-screen.bl
16 text = toUTF8(html_text(tag))
17 }
18 #getContent(data$href[1])
19 allText = sapply(data$href, getContent)
20 allText
```

搭配 sapply

Environment History Build
Global Environment
a... chr [...]
h... chr [...]
h... List ...
t... List ...
U... "http...
Functions
Files Plots Packages Help (View)
New Folder Delete Rename
studies Case6PttGossiping
Name
..
[] .Rhistory
[] packIntoFunction.R
[] refactorR
[] RPTT
[] test.R
[] try.R

17:1 [] getContent(x) R Script Console

網路爬蟲文本蒐集



正規表示式 Regular Expression

- ▶ 正規表示式: 以單個字符串描述一系列符合某句法規則的字符串。
- ▶ 正規表示式列表:
<https://atedev.wordpress.com/2007/11/23/正規表示式-regular-expression/>
- ▶ 使用目的: 有效清洗資料內容，批次移除不需要的字串與特殊字元。
- ▶ 用正規表示式取出所有開頭大寫的英文單字

- [A-Z] 大寫字母之字串
- \w+ 一或多個數字、字母、底線
- 所以得到 [A-Z]\w+ 為所有開頭大寫的英文單字

正規表示式 Regular Expression

RegExr v2.1

by gskinner RegExr v1 GitHub Tutorial

Library

- Help
- Reference
- Cheatsheet
- Examples
- Community
- Favourites

RegExr is an online tool to learn, build, & test Regular Expressions (RegEx / RegExp).

- Results update in **real-time** as you type.
- Roll over** a match or expression for details.
- Save & share** expressions with others.
- Use **Tools** to explore your results.
- Browse the **Library** for help & examples.
- Undo & Redo** with Ctrl-Z / Y.
- Search for & rate **Community** patterns.

Expression

`[A-Z]\w+/$`

21 matches

Text

Welcome to RegExr v2.1 by gskinner.com, proudly hosted by Media Temple!

Edit the Expression & Text to see matches. Roll over matches or the expression for details. Undo mistakes with ctrl-z. Save Favorites & Share expressions with friends or the Community. Explore your results with Tools. A full Reference & Help is available in the Library, or watch the video Tutorial.

Sample text for testing:

abcdefghijklmnopqrstuvwxyz ABCDEFGHIJKLMNOPQRSTUVWXYZ

0123456789 _+-. ,!@#\$%^&*() ;\|/ <> " ' " "

12345 -98.7 3.141 .6180 9,000 +42

555.123.4567 +1-(800)-555-2468

foo@demo.net bar.ba@test.co.uk

www.demo.com http://foo.co.uk/

http://regexr.com/foo.html?q=bar

https://mediatemple.net

<http://regexr.com/>

爬蟲程式的資料清洗

- XML的節點內容包含特殊符號

```
> title
```

[1] "\n\t\t\t\n\t\t\t\t[討論] 你的名字去哪看? \n\t\t\t\n\t\t\t\t"

[2] "\n\t\t\t\n\t\t\t\t[新聞] 三星影業將製作《牧羊少年奇幻之旅》電影\n\t"

[3] "\n\t\t\t\n\t\t\t\t[問片] 追龍捲風\n\t\t\t\t\n\t\t\t\t"

[4] "\n\t\t\t\n\t\t\t[新聞] 李安談首度指導李淳拍戲 對兒子的評價是.\n"

[5] "\n\t\t\t\n\t\t\t\tRe: [討論] 你的名字去哪看?\n\t\t\t\t\n\t\t\t\t"

- 用正規表示式取出上方出現的所有特殊字元

\ 跳脫字元

\[n-t] \ 配上n或t

所以得到 $\backslash[n-t]$ 為上方出現的所有特殊字元

爬蟲程式的資料清洗

RegExr v2.1

by gskinner RegExr v1 GitHub Tutorial

Library

Help

Reference

Cheatsheet

Examples

Community

Favourites

RegExr is an online tool to learn, build, & test Regular Expressions (RegEx / RegExp).

Results update in **real-time** as you type.

Roll over a match or expression for details.

Save & share expressions with others.

Use **Tools** to explore your results.

Browse the **Library** for help & examples.

Undo & Redo with Ctrl-Z / Y.

Search for & rate **Community** patterns.

Expression

share save flags

/\\[n-t]/g

204 matches

Text

\\n\\t\\t\\t\\n\\t\\t\\t\\tFw: [心得] 第53屆金馬獎入圍名單\\n\\t\\t\\t\\n\\t\\t\\t

\\n\\t\\t\\t\\n\\t\\t\\t\\t[討論] 李安的半場無戰事有多高畫質?\\n\\t\\t\\t\\n\\t\\t\\t

\\n\\t\\t\\t\\n\\t\\t\\t\\tRe: [心得] 第53屆金馬獎入圍名單\\n\\t\\t\\t\\n\\t\\t\\t

\\n\\t\\t\\t\\n\\t\\t\\t\\t[負音] 怪奇孤兒院\\n\\t\\t\\t\\n\\t\\t\\t

\\n\\t\\t\\t\\n\\t\\t\\t\\t[讀益] 怪奇孤兒院一間\\n\\t\\t\\t\\n\\t\\t\\t

\\n\\t\\t\\t\\n\\t\\t\\t\\t[新聞] 【金馬入圍名單】國片《一路順風》8面威\\n\\t\\t\\t\\n\\t\\t\\t

\\n\\t\\t\\t\\n\\t\\t\\t\\t[好音] 怪奇孤兒院 \\n\\t\\t\\t\\n\\t\\t\\t

\\n\\t\\t\\t\\n\\t\\t\\t\\t[Live] 探訪The Visit HBO首播21:00\\n\\t\\t\\t\\n\\t\\t\\t

\\n\\t\\t\\t\\n\\t\\t\\t\\t[問片] 懸疑鬥智的劇情片?\\n\\t\\t\\t\\n\\t\\t\\t

\\n\\t\\t\\t\\n\\t\\t\\t\\t[音音] 暫時停止呼吸 看氣氛不要注意邏輯 \\n\\t\\t\\t\\n\\t\\t\\t

\\n\\t\\t\\t\\n\\t\\t\\t\\t[有音] 厲陰宅最後如何降伏邪靈? \\n\\t\\t\\t\\n\\t\\t\\t

\\n\\t\\t\\t\\n\\t\\t\\t\\t[問片] 有關娃娃的鬼片\\n\\t\\t\\t\\n\\t\\t\\t

Tools

Replace

List

Details

Explain

+

爬蟲程式的資料清洗

- 全部特定字元取代: `gsub()`

```
title <- gsub("\\[n-t]", "", title)
```

或

```
title <- gsub("\n", "", title)
```

```
title <- gsub("\t", "", title)
```

- 清洗完畢的資料

```
> title
```

```
[1] "[討論] 你的名字去哪看?"
```

```
[2] "[新聞] 三星影業將製作《牧羊少年奇幻之旅》電影"
```

```
[3] "[問片] 追龍捲風"
```

```
[4] "[新聞] 李安談首度指導李淳拍戲 對兒子的評價是."
```

```
[5] "Re: [討論] 你的名字去哪看?"
```

將整理好的資料輸出成 csv 檔

The screenshot shows the RStudio interface with a data table in the Environment pane and R code in the Console pane. The data table has 11 rows and 5 columns: title, author, path, date, and response. The R code in the Console pane shows the steps to export the data to a CSV file.

	title	author	path	date	response
1	請問 曾祥志 先生 曾祥志 先生 (有雷)	W000000000000	/bbs/movie/M.1475332674.A.F3C.html	10/01	2
2	請問 曾祥志 先生 曾祥志 先生 (有雷)	beatify22	/bbs/movie/M.1475332674.A.C9C.html	10/01	22
3	請問 曾祥志 先生 曾祥志 先生 (有雷)	lowes	/bbs/movie/M.1475332674.A.S6G.html	10/01	3
4	請問 曾祥志 先生 曾祥志 先生 (有雷)	suffice	/bbs/movie/M.1475332674.A.S5G.html	10/01	4
5	請問 曾祥志 先生 曾祥志 先生 (有雷)	incomspicuous	/bbs/movie/M.1475332674.A.AF6.html	10/01	2
6	請問 曾祥志 先生 曾祥志 先生 (有雷)	hayato24	/bbs/movie/M.1475332674.A.34A.html	10/01	1
7	請問 曾祥志 先生 曾祥志 先生 (有雷)	yanyun85106	/bbs/movie/M.1357887191.A.800.html	4/19	22
8	請問 曾祥志 先生 曾祥志 先生 (有雷)	encl129	/bbs/movie/M.1468848054.A.77E.html	7/18	59
9	請問 曾祥志 先生 曾祥志 先生 (有雷)	indusosp	/bbs/movie/M.1475122628.A.DC9.html	9/29	95
10	請問 曾祥志 先生 曾祥志 先生 (有雷)	CatchPlay	/bbs/movie/M.1475231628.A.386.html	9/30	84
11	請問 曾祥志 先生 曾祥志 先生 (有雷)	moneyan	/bbs/movie/M.1475251601.A.41C.html	10/01	8

```
1. 1wan.950;LC_MONETARY=Chinese (Traditional)_Taiwan.950;LC_NUMERIC=C;LC_TIME=Chinese (Traditional)_Taiwan.950
2. > urlPath <- "https://www.ptt.cc/bbs/movie/index.html"
3. > temp <- getURL(urlPath, encoding = "big5")
4. > xmldoc <- htmlParse(temp)
5. > title <- xpathSApply(xmldoc, "//div[@class='title']", xmlValue)
6. > title <- gsub("\n", "", title)
7. > title <- gsub("\t", "", title)
8. > author <- xpathSApply(xmldoc, "//div[@class='author']", xmlValue)
9. > path <- xpathSApply(xmldoc, "//div[@class='title']/a[@href]", xmlValue)
10. > date <- xpathSApply(xmldoc, "//div[@class='date']", xmlValue)
11. > response <- xpathSApply(xmldoc, "//div[@class='nrec']", xmlValue)
12. > alldata <- data.frame(title, author, path, date, response)
13. > View(alldata)
```

The Environment pane shows the 'alldata' object with 11 observations and 5 variables: author, date, path, response, title, and urlPath. The Console pane shows the R code used to create the 'alldata' object and view it.

文本蒐集

- ▶ 網路爬蟲，範例程式：
<https://github.com/pecu/RCrawler101-201504>