



國立臺灣大學
National Taiwan University

中文文字探勘

國立臺灣大學共同教育中心

蔡芸琤

套件安裝

library(tmcn) : https://r-forge.r-project.org/R/?group_id=1571

library(NLP)

library(tm)

library(jiebaRD)

library(jiebaR)

library(RColorBrewer)

library(wordcloud)

library(rvest)

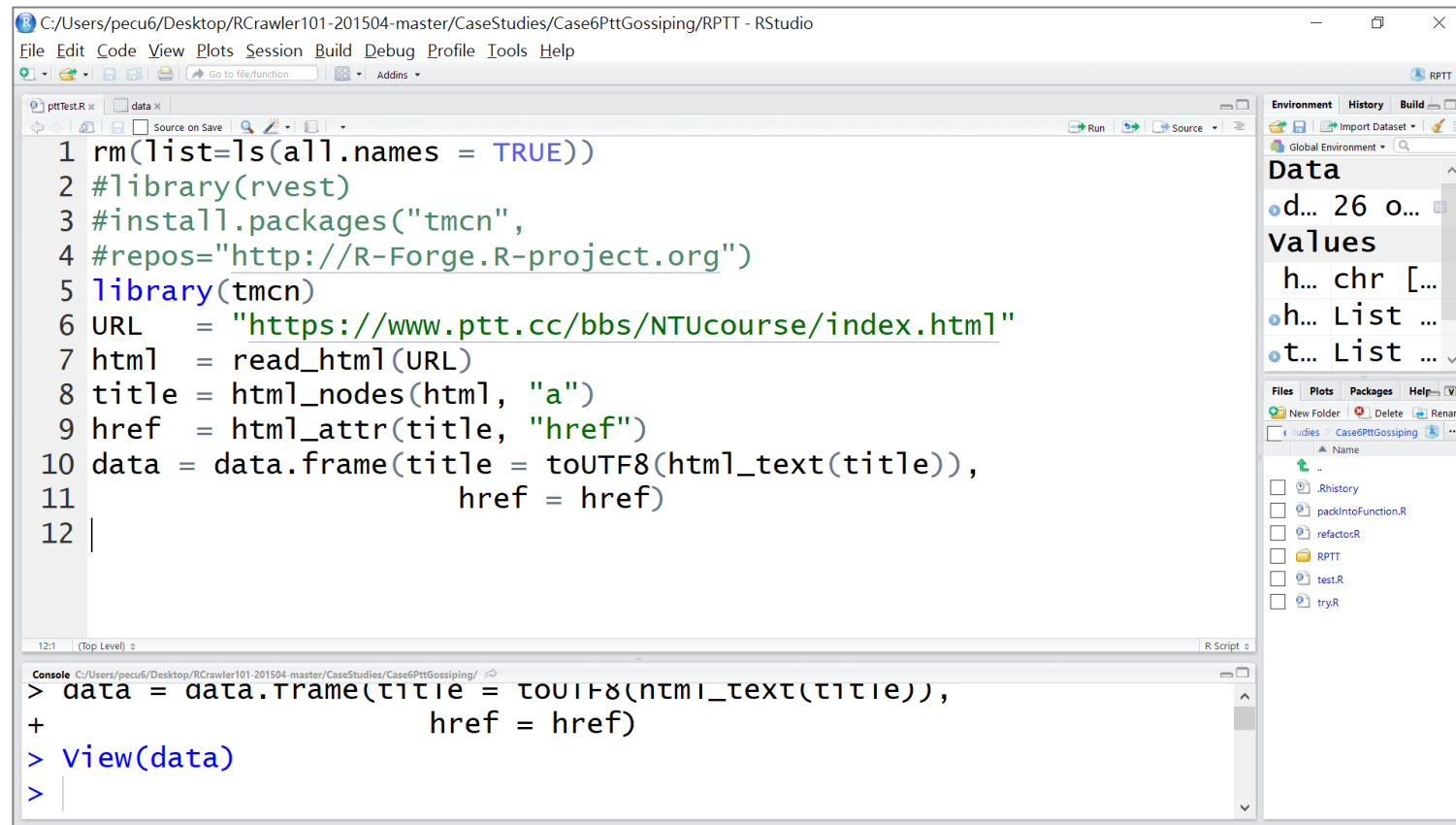
文本蒐集

1. 網路爬蟲，範例程式：
<https://github.com/pecu/RCrawler101-201504>
2. 本機端純文字檔

網路爬蟲文本蒐集



網路爬蟲文本蒐集



The screenshot shows an RStudio interface with the following components:

- Source Editor:** Contains R code for scraping a PTT forum page. The code uses the `tm` package to fetch and parse HTML data into a data frame.
- Environment:** Shows the current data environment with a list of objects including `data`, `html`, `href`, `title`, `URL`, and `html_nodes`.
- Files:** A file explorer on the right showing the project structure, including files like `.Rhistory`, `packIntoFunction.R`, `refactor.R`, `RPTT`, `test.R`, and `try.R`.
- Console:** Displays the execution of the `data = data.frame(title = toUTF8(html_text(title)), href = href)` command and the subsequent `view(data)` command.

```
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 data

Filter

	title	href
1	批發路實業坊	/
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] 普通生物學門 考古題(李鳳鳴)	/bbs/NTUcourse/M.1492329082.A.1D4.html
12	[問題] 請問麻坤錢 (二) 34高爾夫球	/bbs/NTUcourse/M.1492364474.A.141.html
13	[評價] 105-2 丁亮 文學學乙下	/bbs/NTUcourse/M.1492410691.A.528.html
14	[問題] 請問史前史二期中考	/bbs/NTUcourse/M.1492421244.A.118.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 國語二期中考題	/bbs/NTUcourse/M.1492439974.A.BC4.html
17	[問題] 請問吳洪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) 黃淑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 李碧山	/bbs/NTUcourse/M.1492585301.A.4F4.html
20	[問題] 一般醫學保健有那些名?	/bbs/NTUcourse/M.1492591816.A.65C.html
21	[求助] 歐陽志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

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

```
> data = data.frame(title = toupper(substr(text(title), 1, 10)),  
+ href = href)  
> view(data)  
>
```

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

udies Case6PttGossiping

Name

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

網路爬蟲文本蒐集

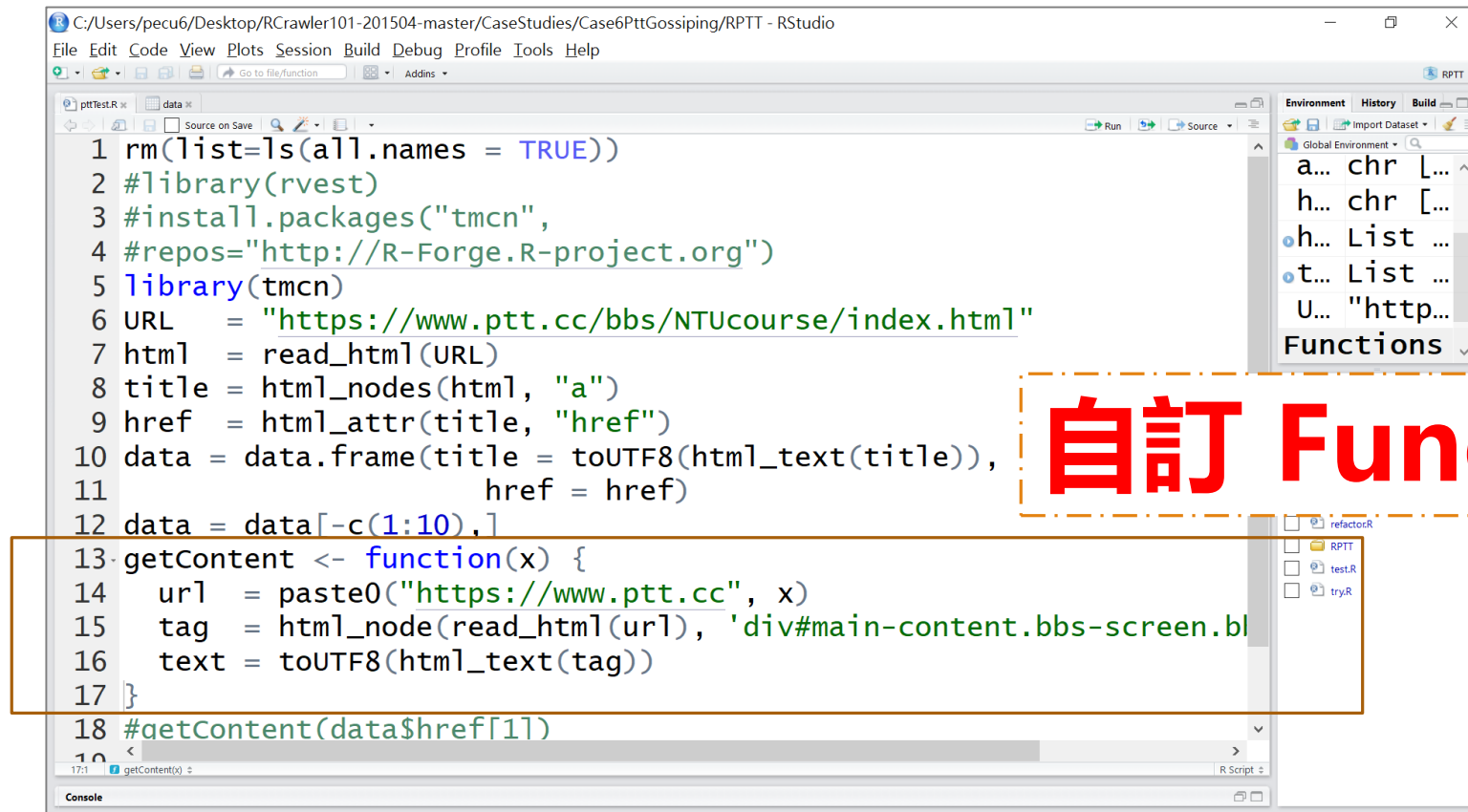
The screenshot displays a web browser window with the address bar showing `https://www.ptt.cc/bbs/NTUcourse/M.1492329082.A.1D4.html`. The page content is a forum post from the NTUcourse board. The post header includes the board name "批踢踢實業坊" and "NTUcourse", the author "handfox (handwolf)", the title "[求救] 普通生物學丙 考古題(李鳳鳴)", and the time "Sun Apr 16 15:51:19 2017". The post body contains the text: "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！". Below the post body, it shows the sender's IP "223.136.95.125" and the article URL. At the bottom, there are buttons for "返回看板", "分享", "Like 0", and a "G+1" button.

The browser's developer tools are open on the right side, showing the "Elements" panel. The selected element is a `div` with the class `article-metaling`. The DOM tree shows the following structure:

```
<div id="navigation-container">...</div>
<div id="main-container">
  <div id="main-content" class="bbs-screen bbs-content">
    <div class="article-metaling">...</div>
    <div class="article-metaling-right">...</div>
    <div class="article-metaling">...</div>
  </div>
  <div class="article-metaling">...</div>
</div>
```

The "Styles" panel shows the default styles for the `div` element, including `margin`, `border`, and `padding`.

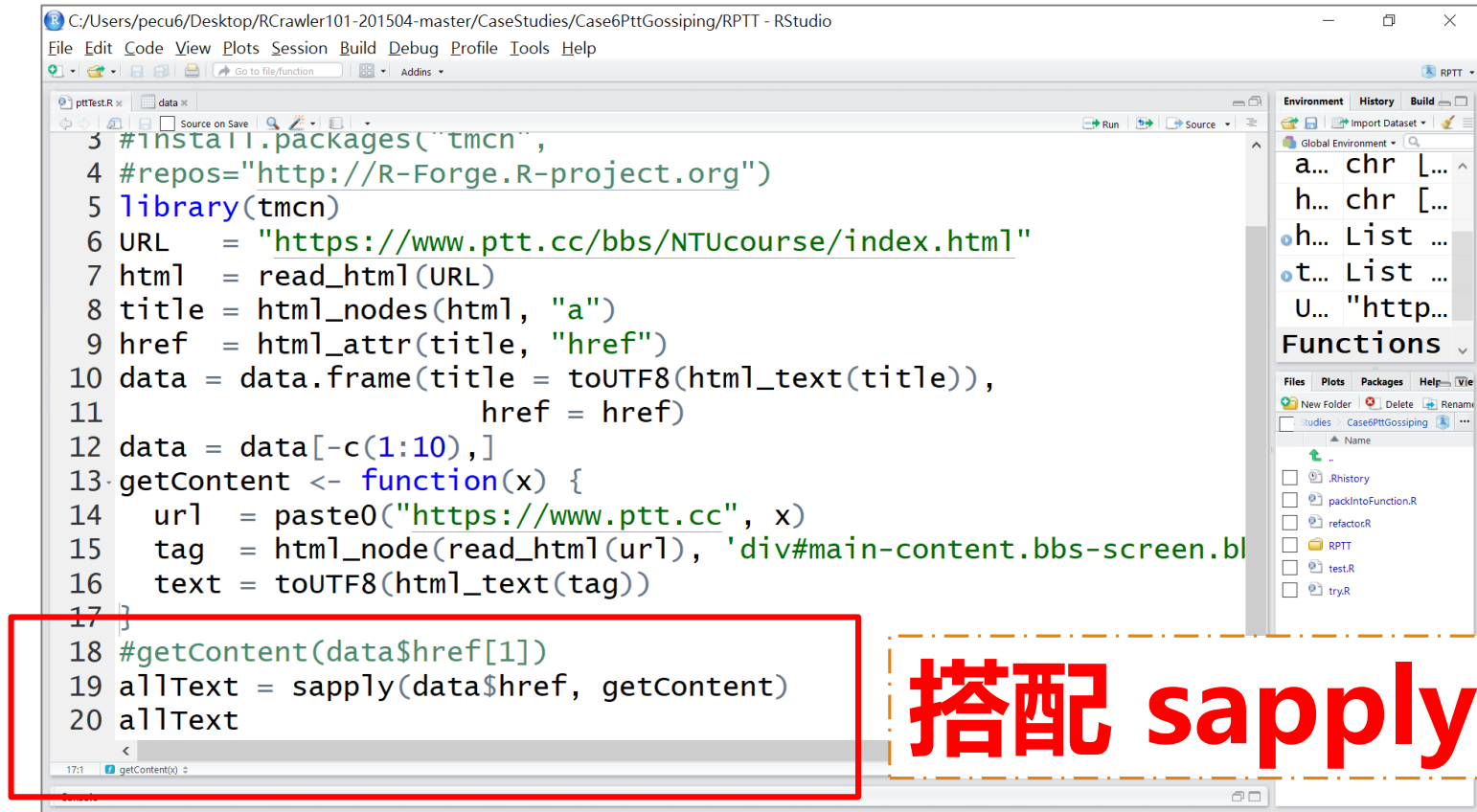
網路爬蟲文本蒐集



```
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[-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])
```

自訂 Function

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

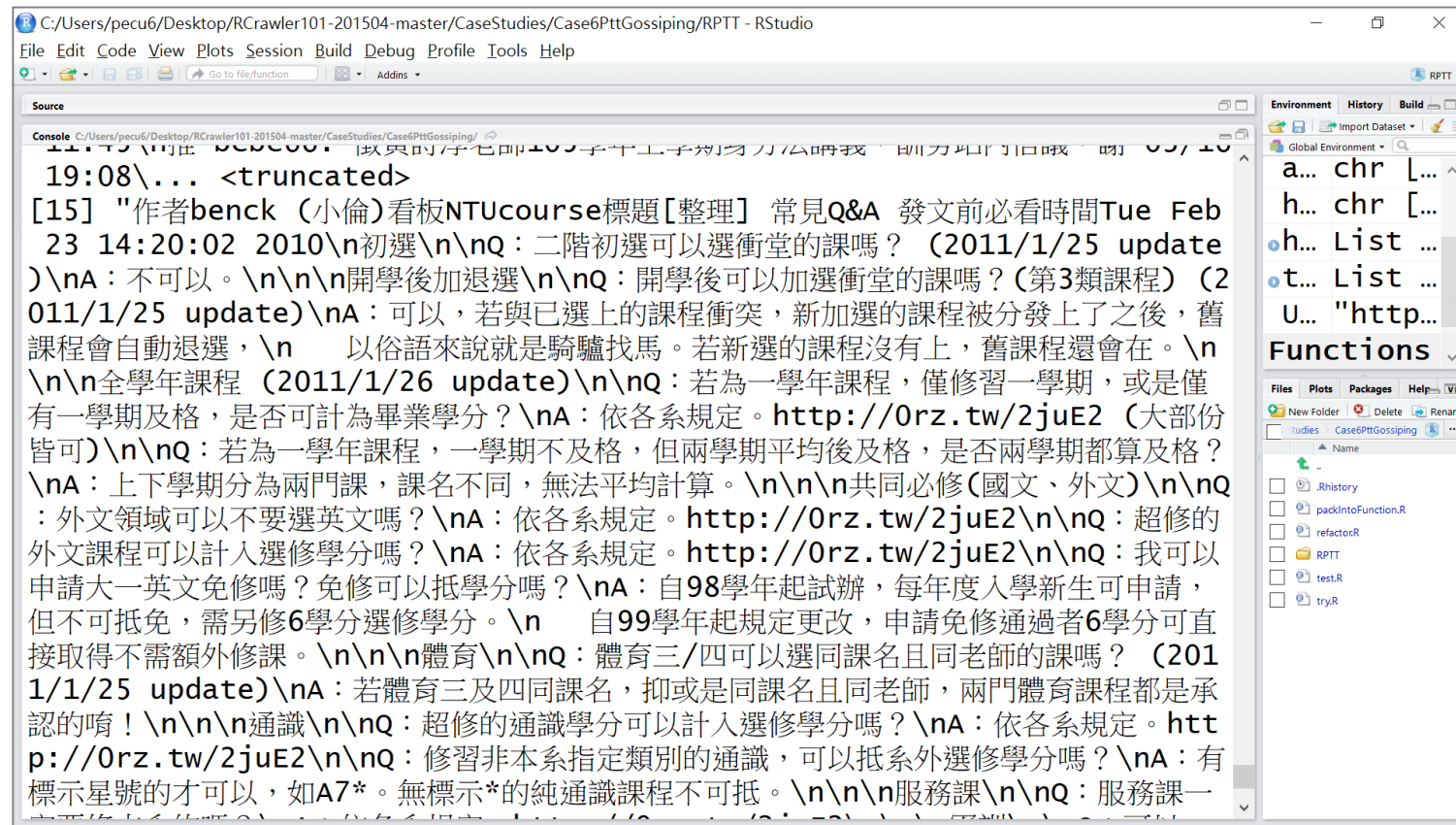
pttTest.R x data x
Source on Save
Run Source

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

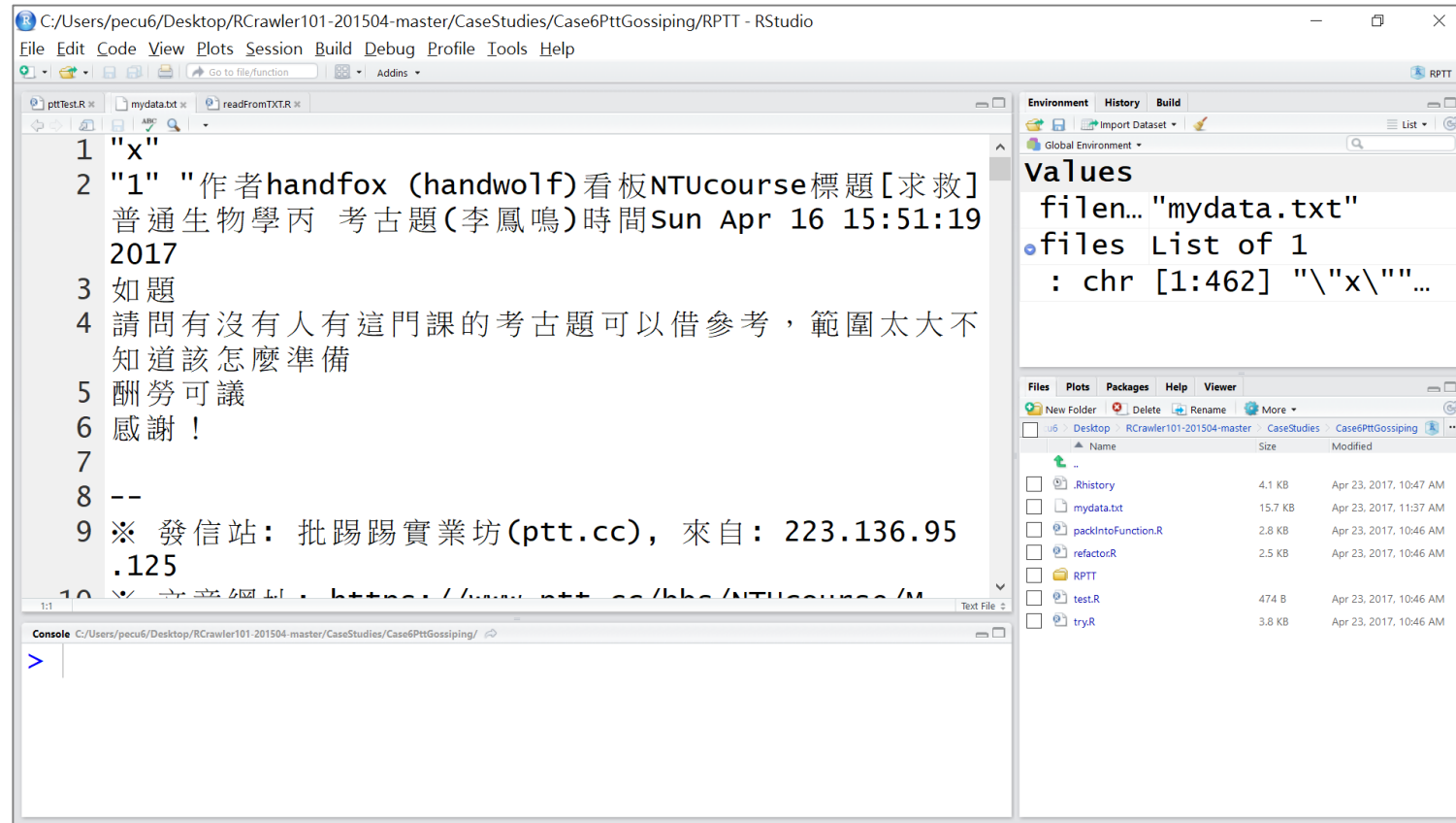
17:1 | getConten(x) |
```

搭配 sapply

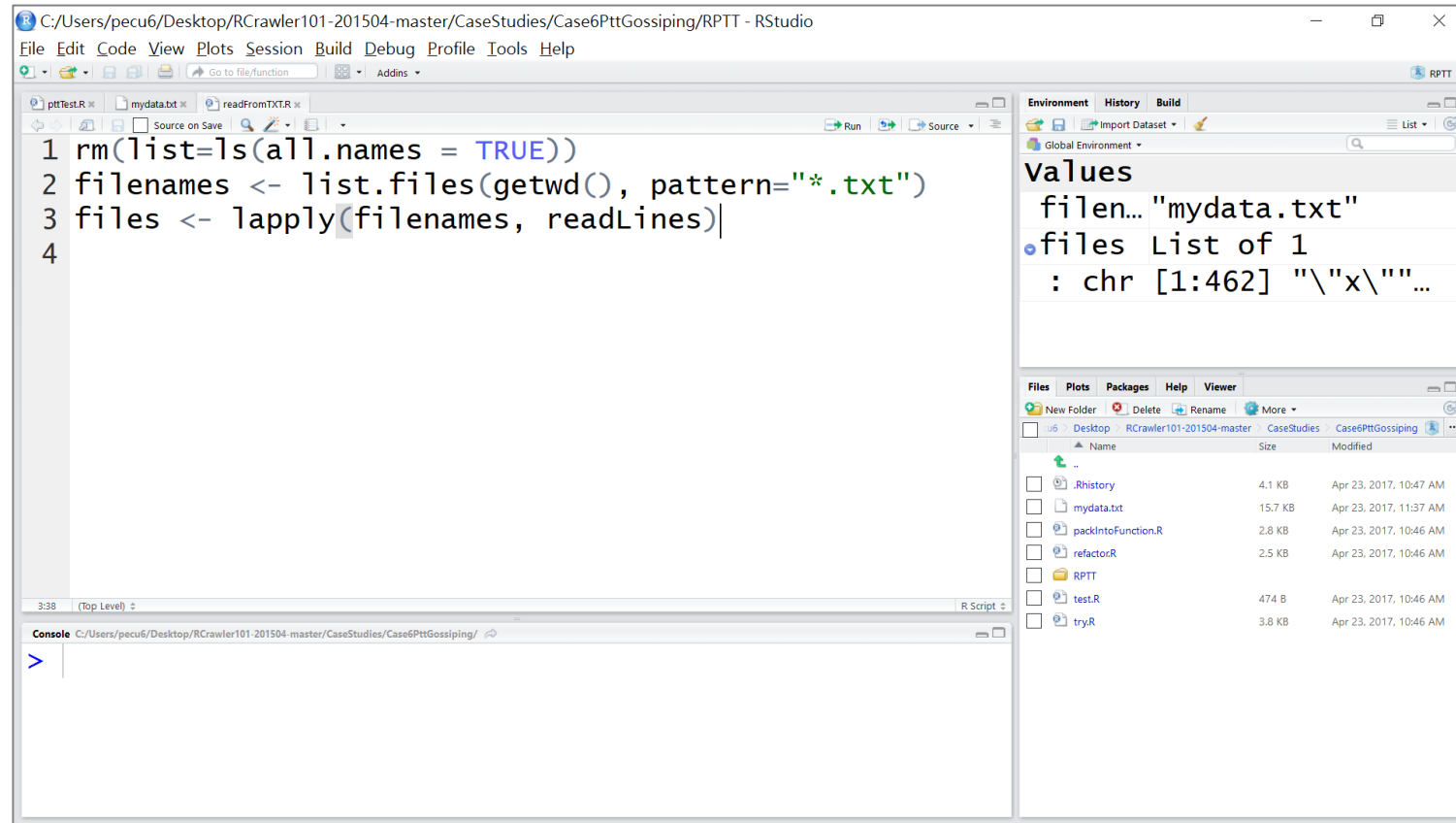
網路爬蟲文本蒐集



本機端純文字檔本蒐集



本機端純文字檔本蒐集



課堂練習

連續產生 10 頁以上的純文字檔

1. 人工複製貼上？
2. 網路爬蟲？
3. 現成的資料庫？

課堂練習

The screenshot displays the RStudio interface. The source editor on the left contains the following R code:

```
1 source('pttTestFunction.R')
2 id = c(1:10)
3 URL = paste0("https://www.ptt.cc/bbs/NTUcourse/index",
4 filename = paste0(id, ".txt")
5 pttTestFunction(URL[1], filename[1])
6 mapply(pttTestFunction,
7       URL = URL, filename = filename)
8
```

The Environment pane on the right shows the following values:

Variable	Type	Value
file...	chr [1:10]	"1.t...
file...	chr [1:10]	"mydata.txt"
files	List of 1	
id	int [1:10]	1 2 ...
URL	chr [1:10]	"htt...

The Files pane at the bottom right shows a directory listing of generated files:

Name	Size	Modified
main.R	247 B	Apr 23, 2017, 12:03 PM
pttTest.R	688 B	Apr 23, 2017, 11:37 AM
pttTestFunction.R	740 B	Apr 23, 2017, 11:48 AM
readFromTXT.R	121 B	Apr 23, 2017, 11:42 AM
RPTT	3.5 KB	Apr 23, 2017, 10:46 AM
1.txt	19.1 KB	Apr 23, 2017, 12:02 PM
2.txt	21.8 KB	Apr 23, 2017, 12:02 PM
3.txt	22.7 KB	Apr 23, 2017, 12:02 PM
4.txt	24.8 KB	Apr 23, 2017, 12:02 PM
5.txt	22.3 KB	Apr 23, 2017, 12:02 PM
6.txt	18.1 KB	Apr 23, 2017, 12:02 PM
7.txt	17 KB	Apr 23, 2017, 12:02 PM
8.txt	26.3 KB	Apr 23, 2017, 12:02 PM
9.txt	20.8 KB	Apr 23, 2017, 12:02 PM
10.txt	26.1 KB	Apr 23, 2017, 12:02 PM

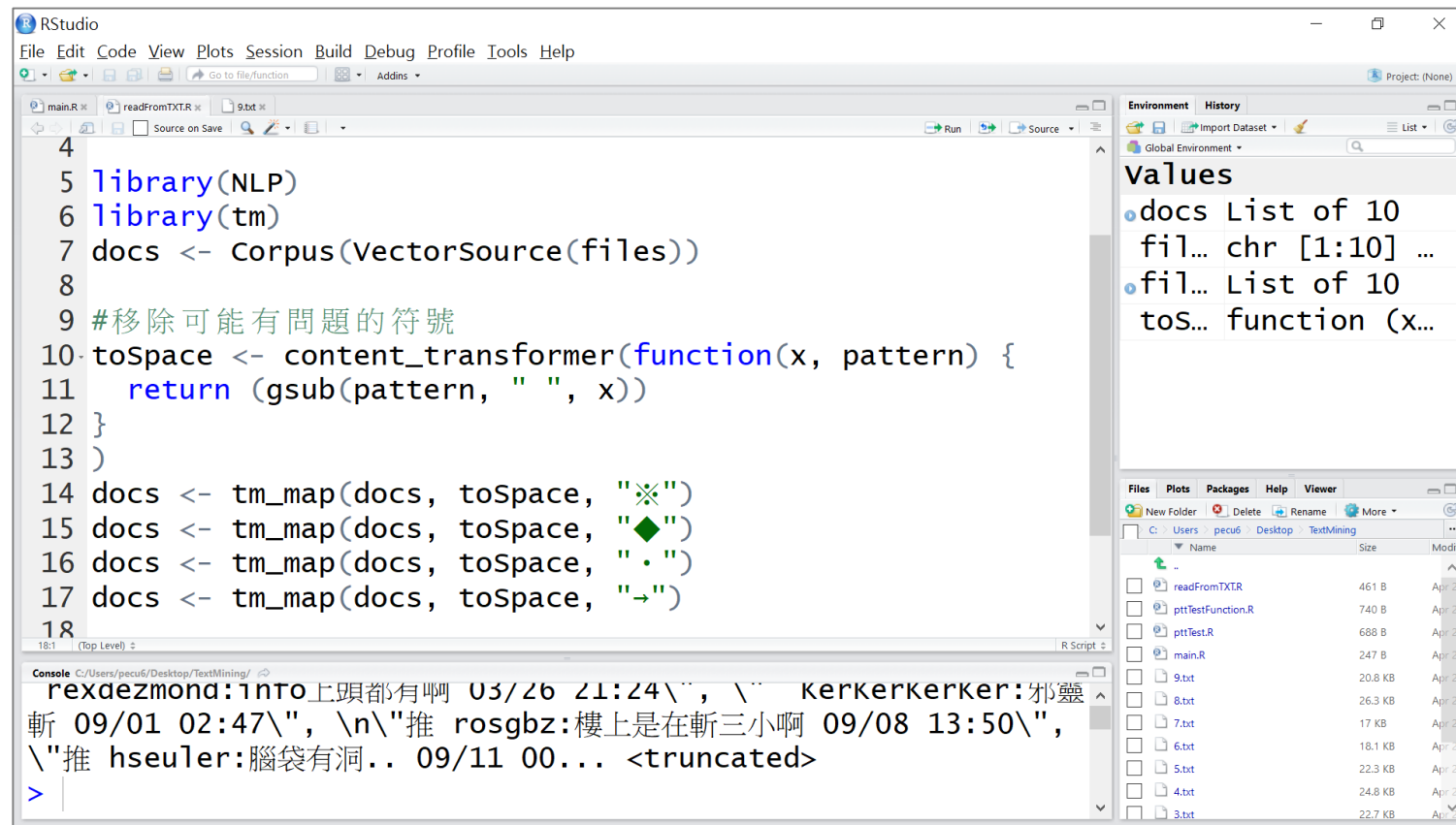
自訂 Function
搭配 mapply

自動生成檔案

文本清理

- 參考資料：https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf
- 大小寫轉換
- 標點符號、數字移除
- URLs 移除
- 表情符號、停用詞移除

文本清理



The screenshot shows the RStudio interface with a script editor, environment pane, and console.

```
4  
5 library(NLP)  
6 library(tm)  
7 docs <- Corpus(VectorSource(files))  
8  
9 #移除可能有問題的符號  
10 toSpace <- content_transformer(function(x, pattern) {  
11   return (gsub(pattern, " ", x))  
12 }  
13 )  
14 docs <- tm_map(docs, toSpace, "※")  
15 docs <- tm_map(docs, toSpace, "◆")  
16 docs <- tm_map(docs, toSpace, ".")  
17 docs <- tm_map(docs, toSpace, "→")  
18
```

Environment pane (Values):

- docs List of 10
fil... chr [1:10] ...
- fil... List of 10
toS... function (x...

Files pane (C:\Users\pecu6\Desktop\TextMining):

Name	Size	Modified
readFromTXT.R	461 B	Apr 24
pttTestFunction.R	740 B	Apr 23
pttTest.R	688 B	Apr 23
main.R	247 B	Apr 23
9.txt	20.8 KB	Apr 23
8.txt	26.3 KB	Apr 23
7.txt	17 KB	Apr 23
6.txt	18.1 KB	Apr 23
5.txt	22.3 KB	Apr 23
4.txt	24.8 KB	Apr 23
3.txt	22.7 KB	Apr 23

Console (C:\Users\pecu6\Desktop\TextMining/):

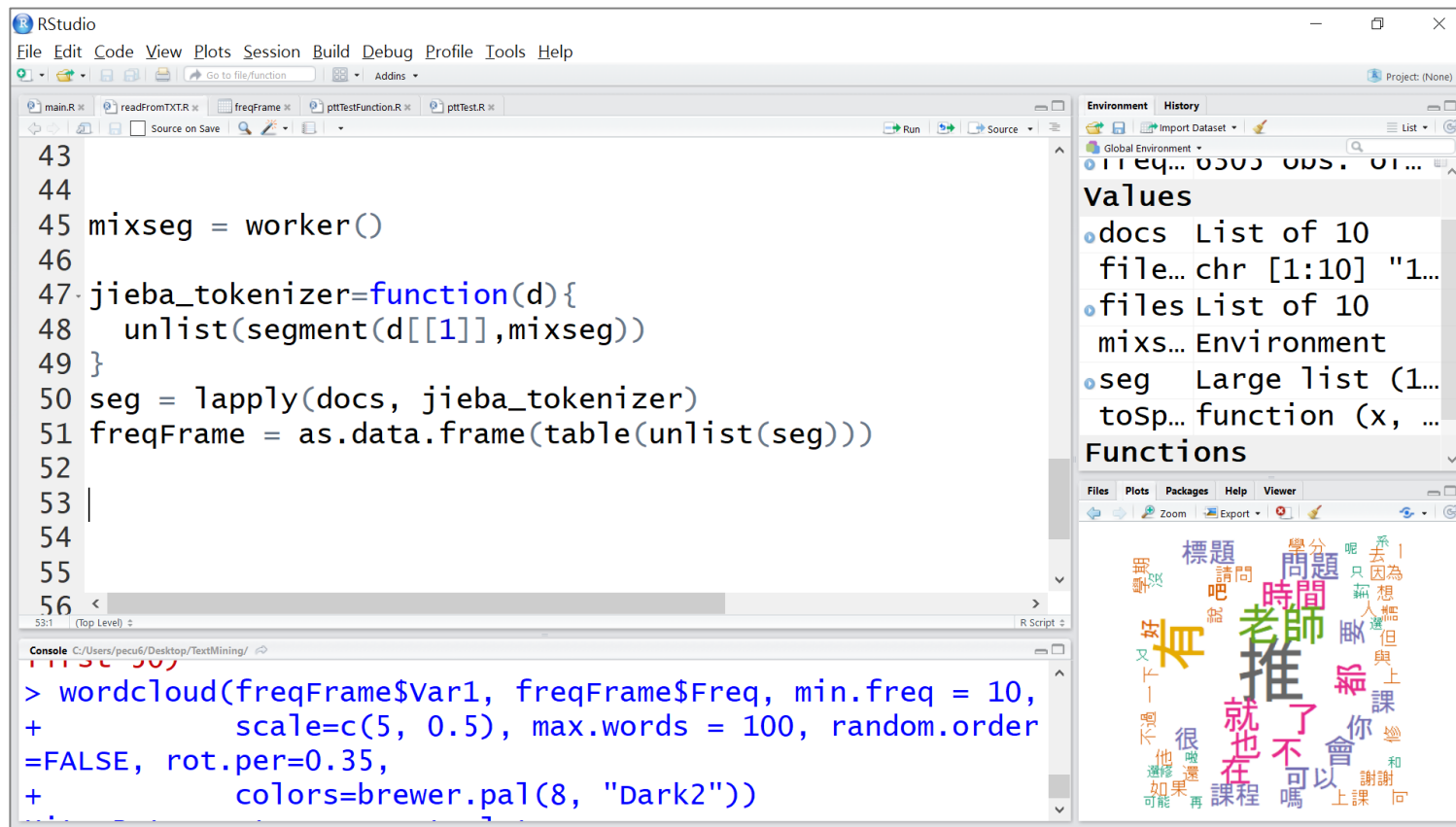
```
rexdezmond: into上頭都有啊 03/26 21:24\", \" kerkkerkerker: 邪惡  
斬 09/01 02:47\", \"推 rosgbz: 樓上是在斬三小啊 09/08 13:50\",  
\"推 hseuler: 腦袋有洞.. 09/11 00... <truncated>  
>
```


斷詞處理

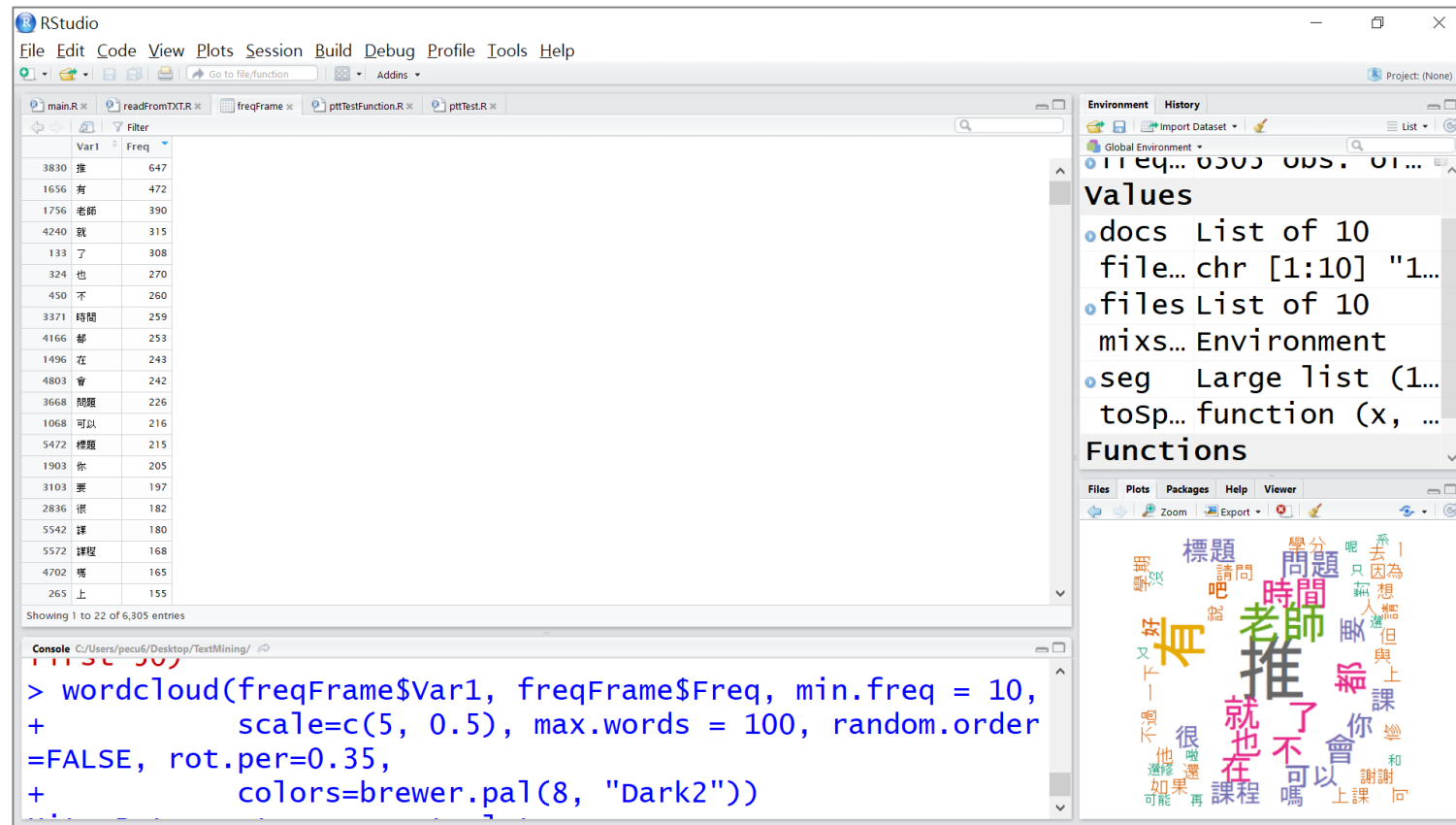
產生正確中文詞頻矩陣

1. 使用 `cutter=worker()` 產生切詞器。
2. 使用 `new_user_word` 將新詞彙加入詞庫。
3. 使用 `cutter=worker("tag")` 可切割出詞彙與提供詞彙的詞性。

詞頻矩陣



詞頻矩陣

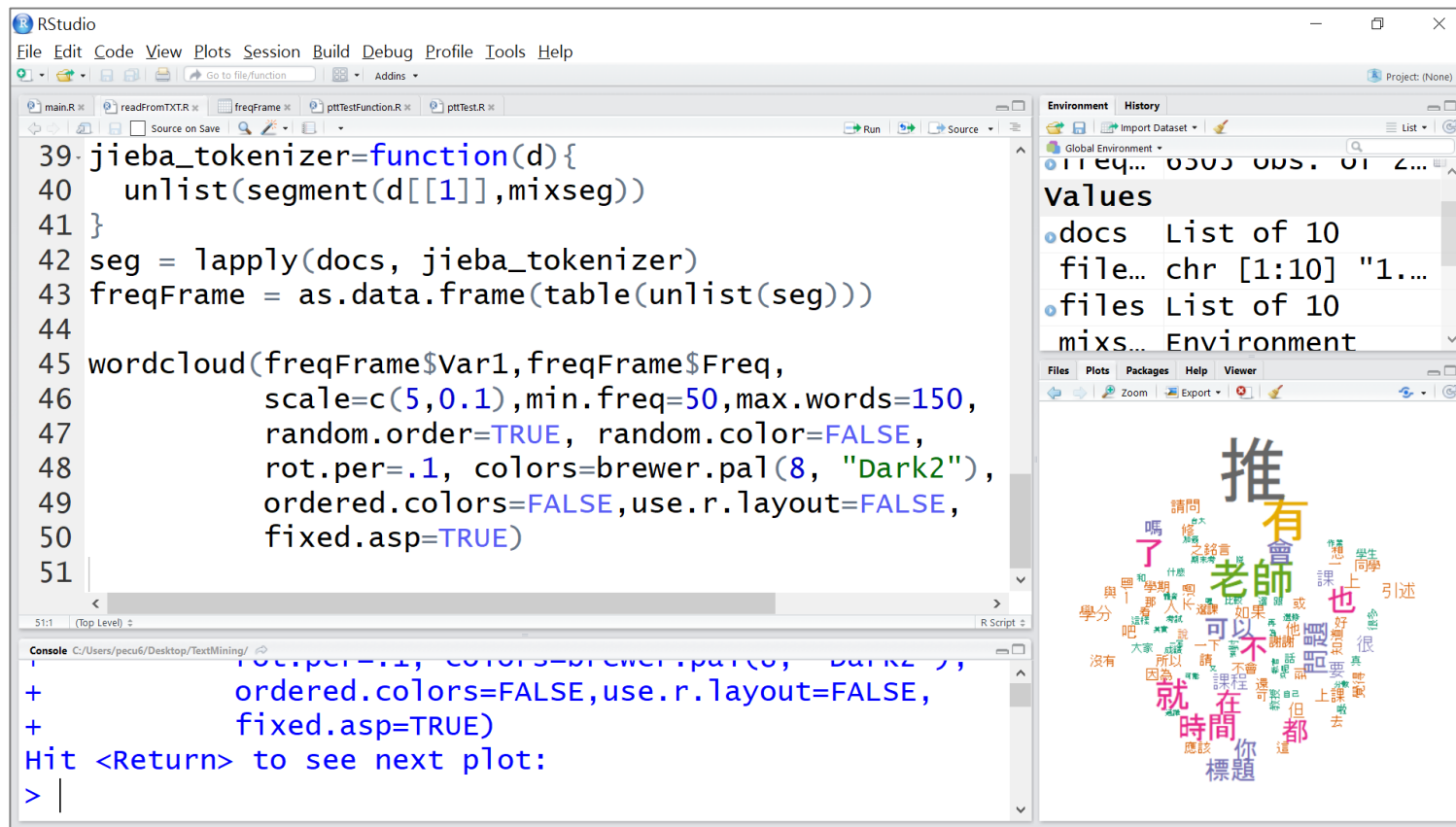


文字雲

<https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf>

```
wordcloud(words,freq,scale=c(4,.5),min.freq=3,max.words=Inf,  
  random.order=TRUE, random.color=FALSE, rot.per=.1,  
  colors="black",ordered.colors=FALSE,use.r.layout=FALSE,  
  fixed.asp=TRUE, ...)
```

文字雲



HW5 請完成一個文字雲

文字探勘目標

<https://buzzorange.com/techorange/2017/04/13/data-to-pxmart-hsu/>

1. 文字雲
2. 詞語詞間的關係
3. 文本關聯