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

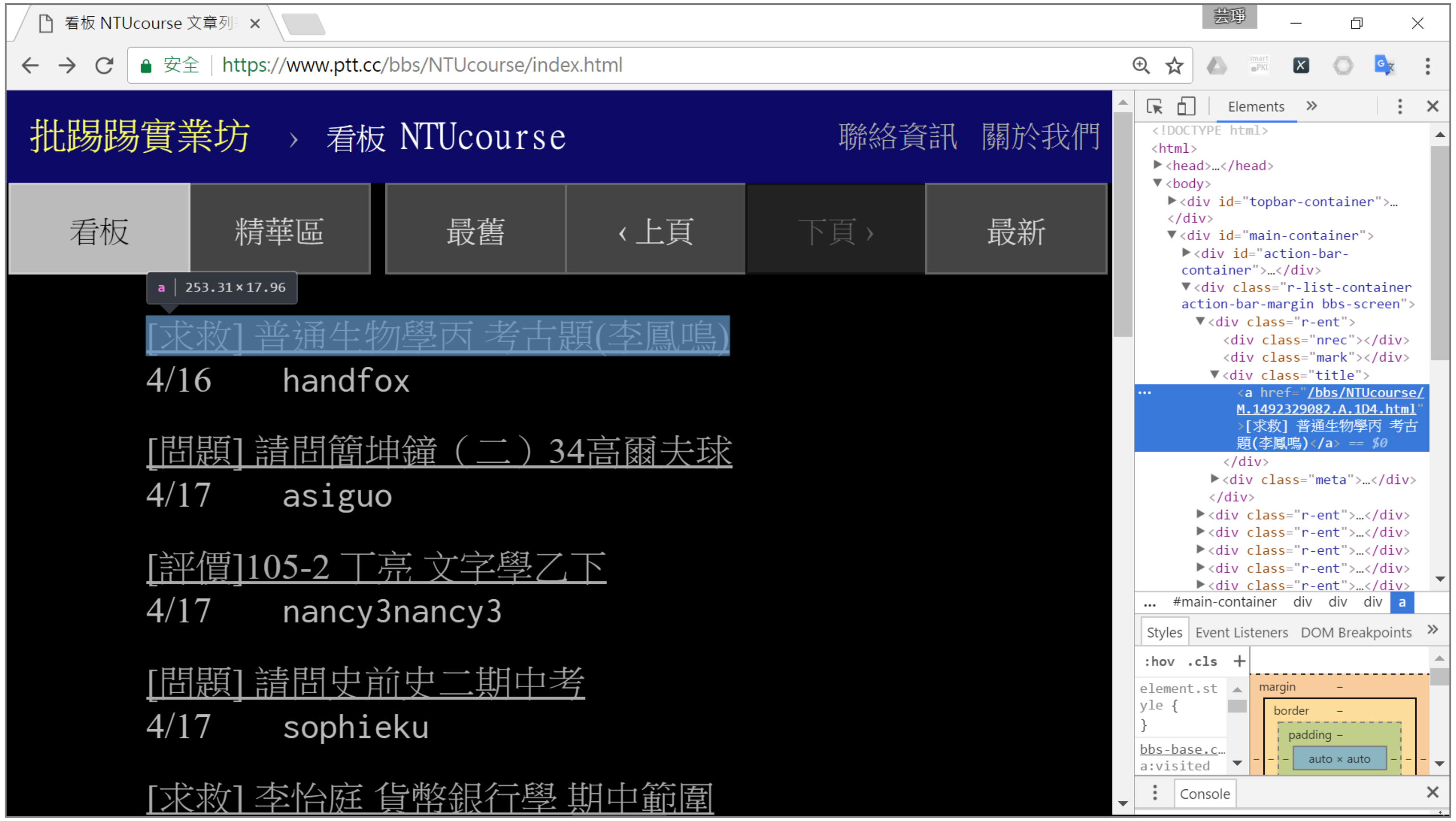
套件安裝

```
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



看板 NTUcourse 文章列

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

[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

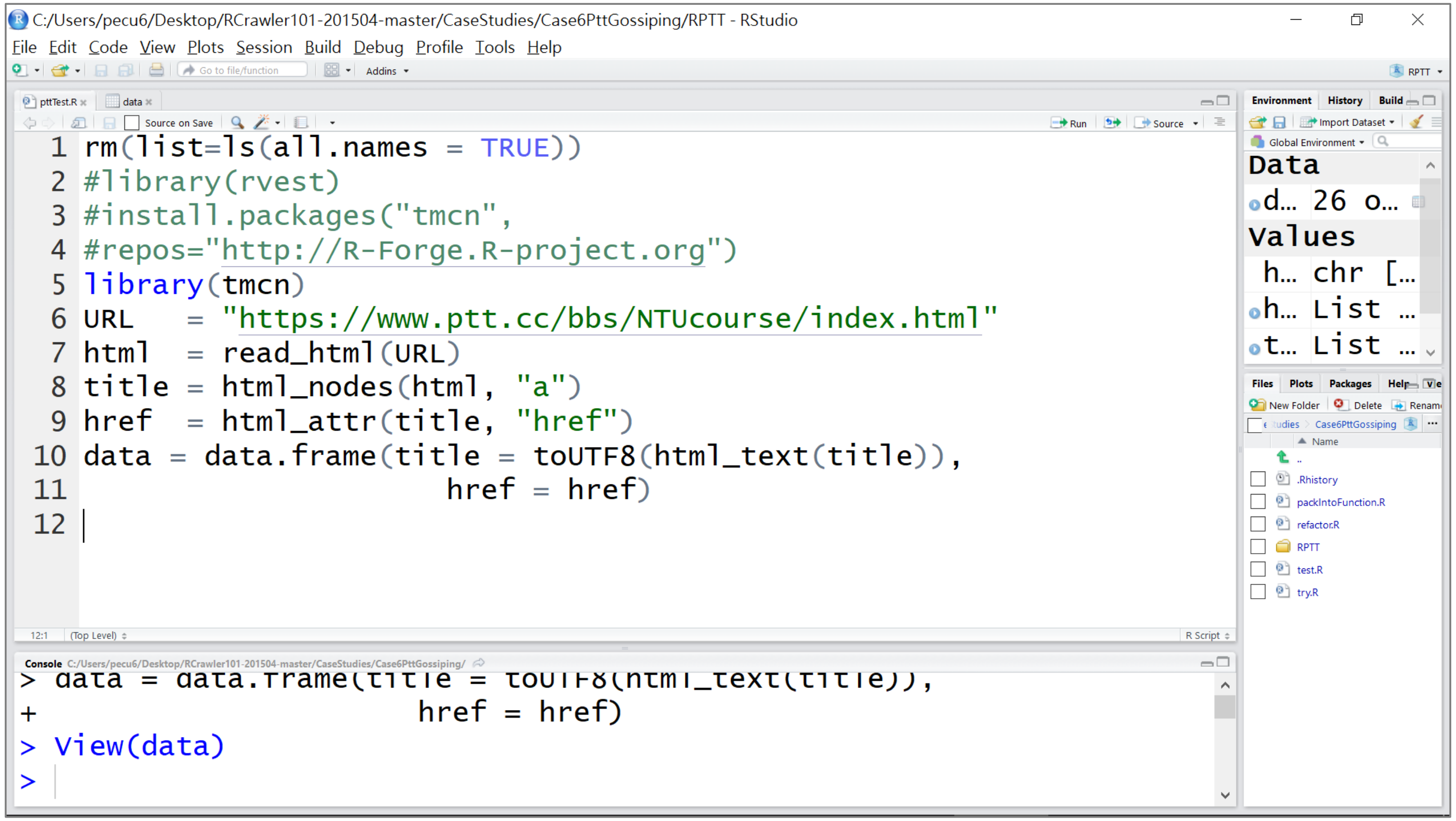
[評價]105-2 丁亮 文字學乙下
4/17 nancy3nancy3

[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

```
<!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 class="r-ent">…</div>
```

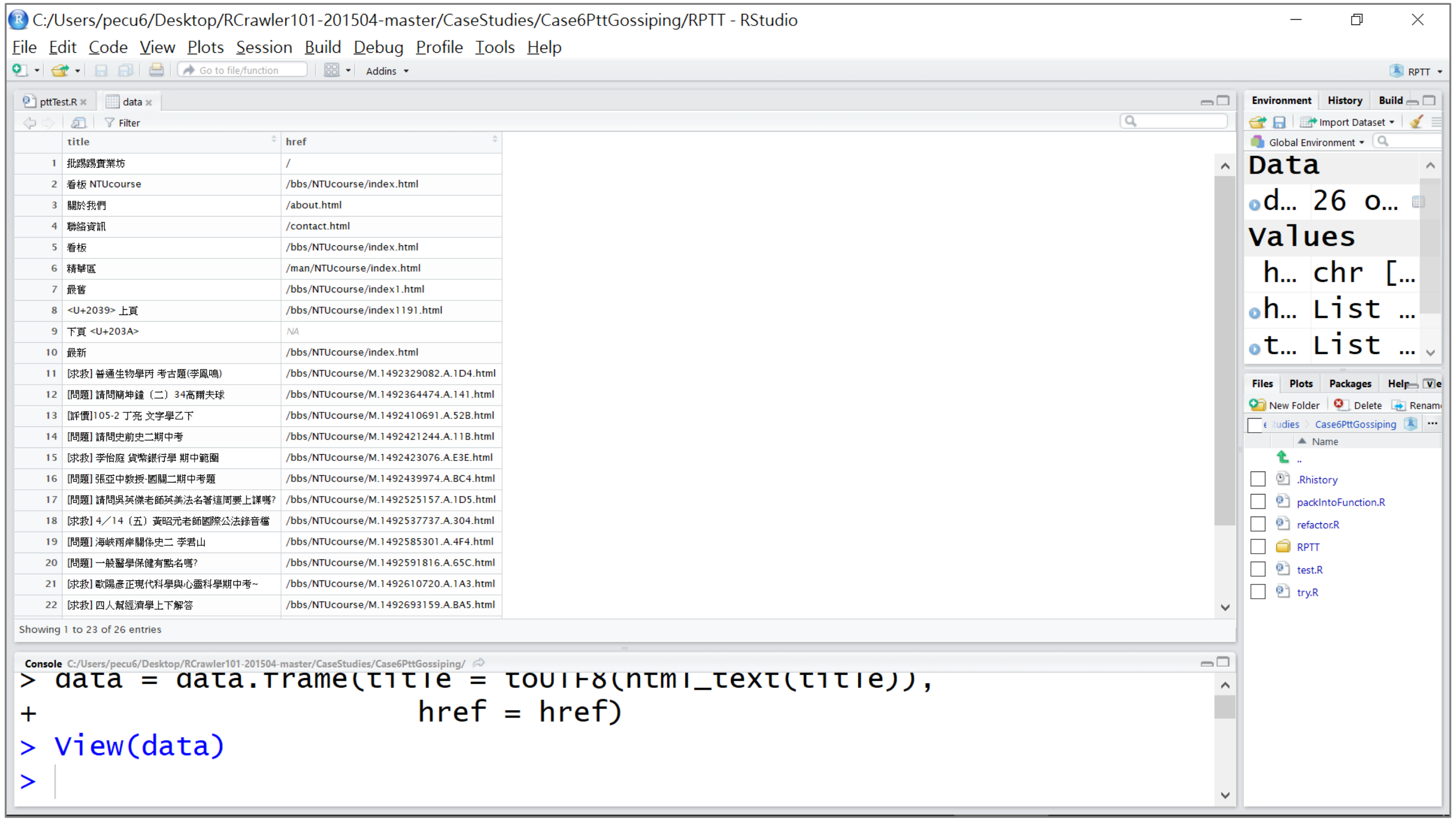
網路爬蟲文本蒐集



```r
rm(list=ls(all.names = TRUE))
#library(rvest)
#install.packages("tmcn",
#repos="http://R-Forge.R-project.org")
library(tmcn)
URL    = "https://www.ptt.cc/bbs/NTUcourse/index.html"
html   = read_html(URL)
title  = html_nodes(html, "a")
href   = html_attr(title, "href")
data = data.frame(title = toUTF8(html_text(title)),
                  href = href)

```

```
> 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

	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	[求救] 普通生物學丙 考古題(李鳳鳴)	/bbs/NTUcourse/M.1492329082.A.1D4.html
12	[問題] 請問簡坤鐘（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	[求救] 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	[求救] 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] 一般醫學保健有點名嗎?	/bbs/NTUcourse/M.1492591816.A.65C.html
21	[求救] 歐陽彥正現代科學與心靈科學期中考~	/bbs/NTUcourse/M.1492610720.A.1A3.html
22	[求救] 四人幫經濟學上下解答	/bbs/NTUcourse/M.1492693159.A.BA5.html

Showing 1 to 23 of 26 entries

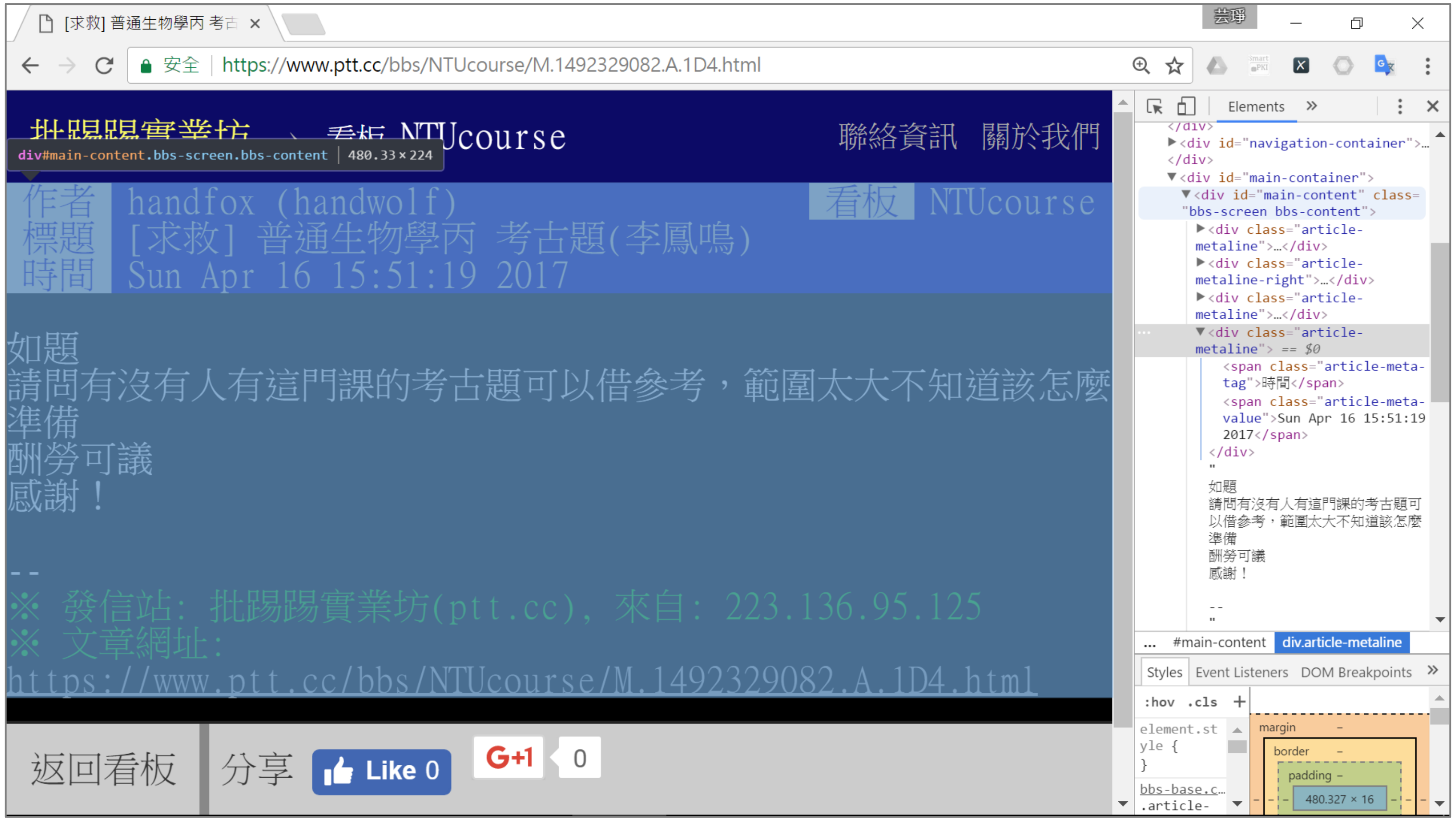
Console C:/Users/pecu6/Desktop/RCrawler101-201504-master/CaseStudies/Case6PttGossiping/

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

```
Data
d… 26 o…
Values
h… chr […
h… List …
t… List …
```

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

網路爬蟲文本蒐集

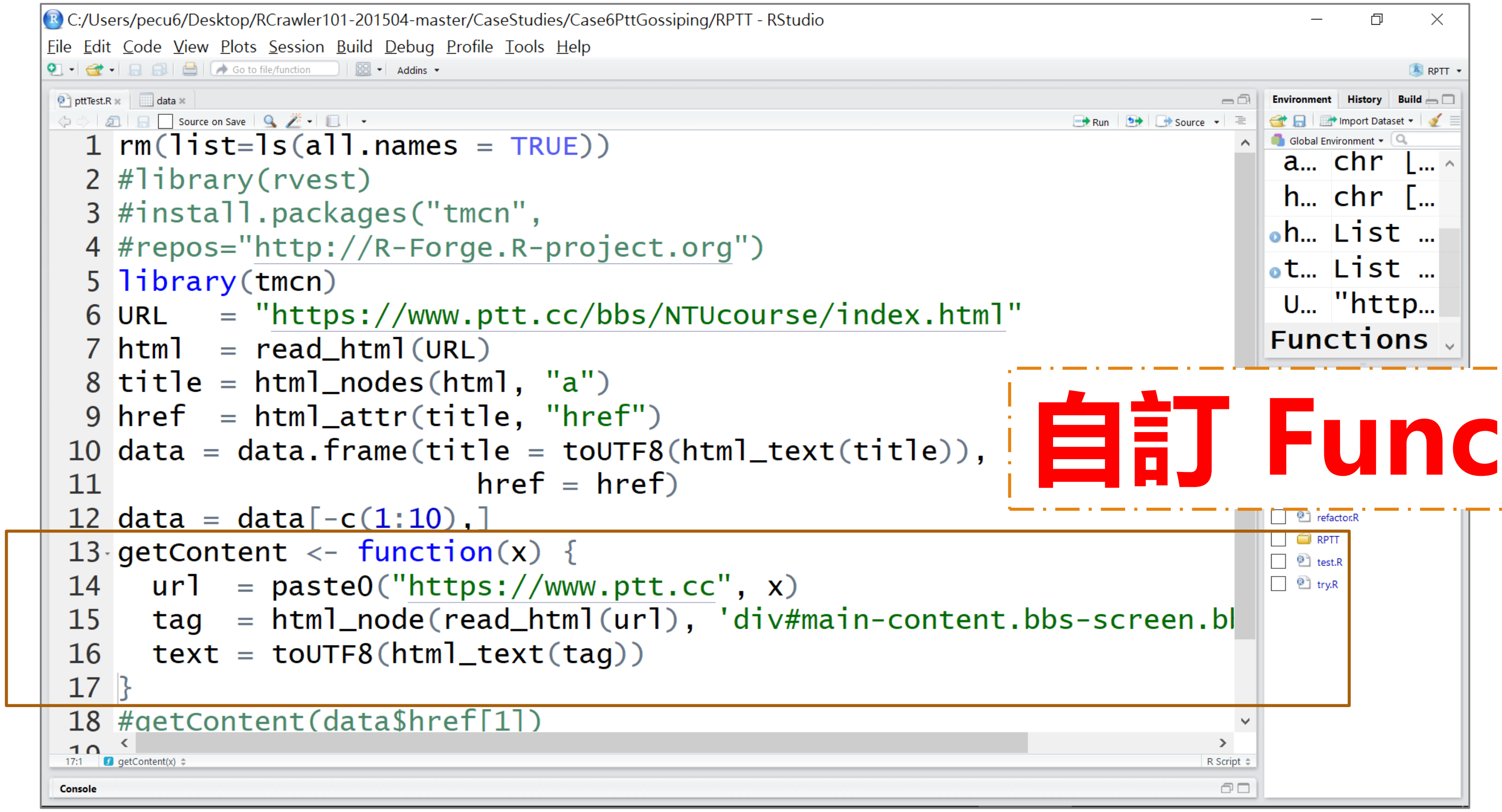


```
作者  handfox (handwolf)                     看板  NTUcourse
標題  [求救] 普通生物學丙 考古題(李鳳鳴)
時間  Sun Apr 16 15:51:19 2017

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
```

```
  </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
        <span class="article-meta-
        tag">時間</span>
        <span class="article-meta-
        value">Sun Apr 16 15:51:19
        2017</span>
      </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！

      --
      "
```

```
element.style {
}
bbs-base.c…
.article-
```

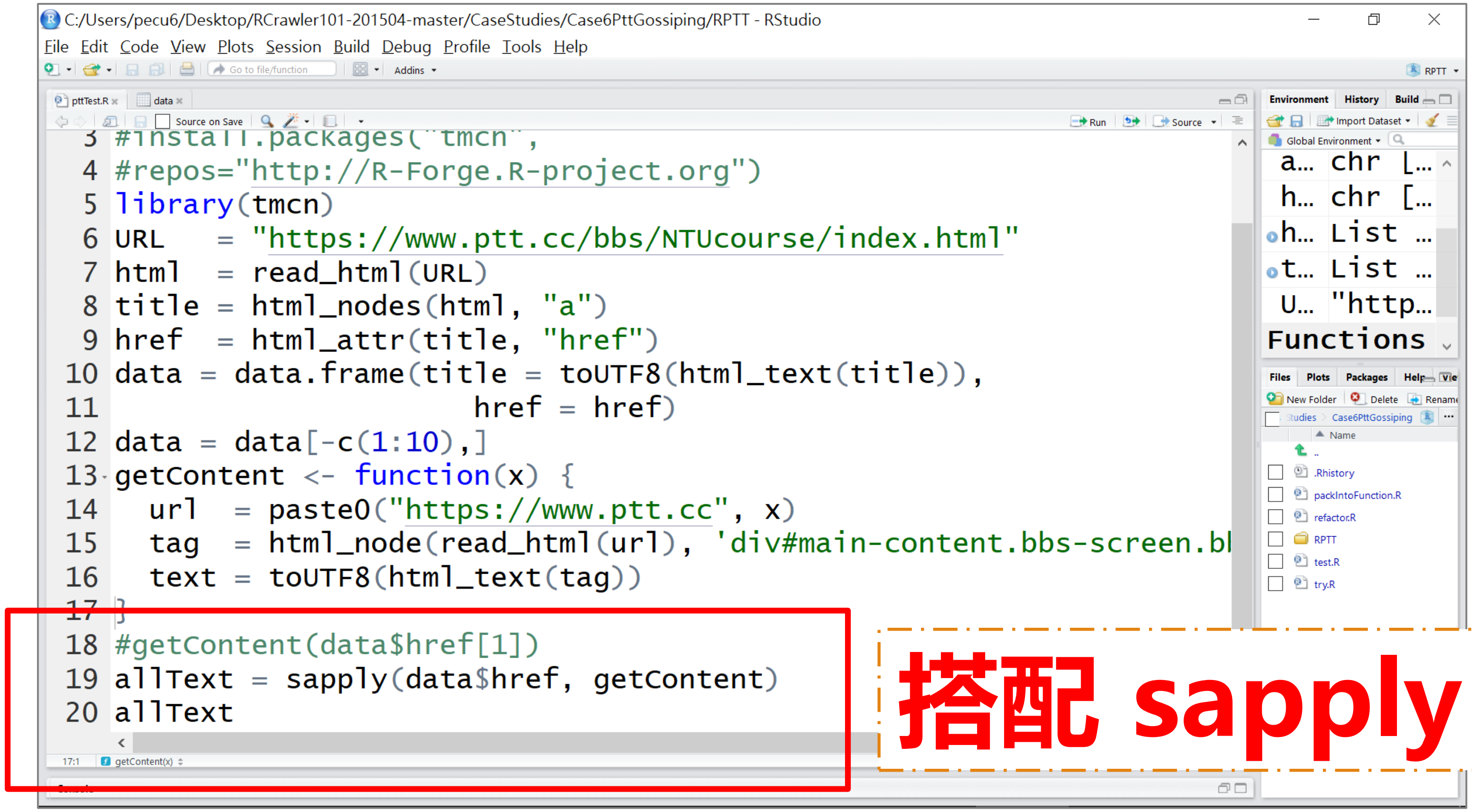
網路爬蟲文本蒐集



```r
rm(list=ls(all.names = TRUE))
#library(rvest)
#install.packages("tmcn",
#repos="http://R-Forge.R-project.org")
library(tmcn)
URL    = "https://www.ptt.cc/bbs/NTUcourse/index.html"
html   = read_html(URL)
title  = html_nodes(html, "a")
href   = html_attr(title, "href")
data  = data.frame(title = toUTF8(html_text(title)),
                   href = href)
data = data[-c(1:10),]
getContent <- function(x) {
  url  = paste0("https://www.ptt.cc", x)
  tag  = html_node(read_html(url), 'div#main-content.bbs-screen.bl
  text = toUTF8(html_text(tag))
}
#getContent(data$href[1])
```

自訂 Function

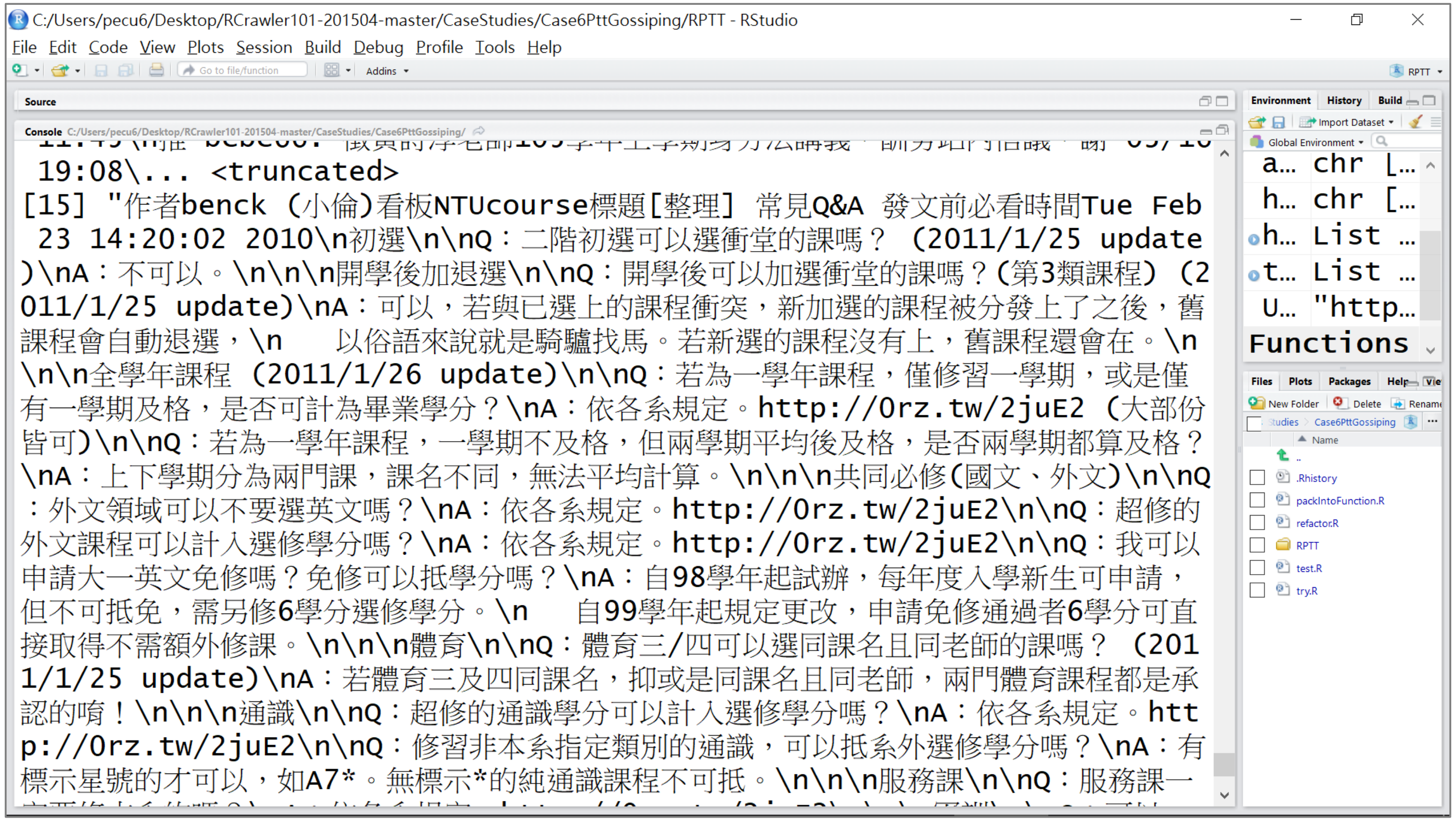
網路爬蟲文本蒐集



```r
#install.packages("tmcn",
#repos="http://R-Forge.R-project.org")
library(tmcn)
URL   = "https://www.ptt.cc/bbs/NTUcourse/index.html"
html  = read_html(URL)
title = html_nodes(html, "a")
href  = html_attr(title, "href")
data = data.frame(title = toUTF8(html_text(title)),
                  href = href)
data = data[-c(1:10),]
getContent <- function(x) {
  url  = paste0("https://www.ptt.cc", x)
  tag  = html_node(read_html(url), 'div#main-content.bbs-screen.bl
  text = toUTF8(html_text(tag))
}
#getContent(data$href[1])
allText = sapply(data$href, getContent)
allText
```

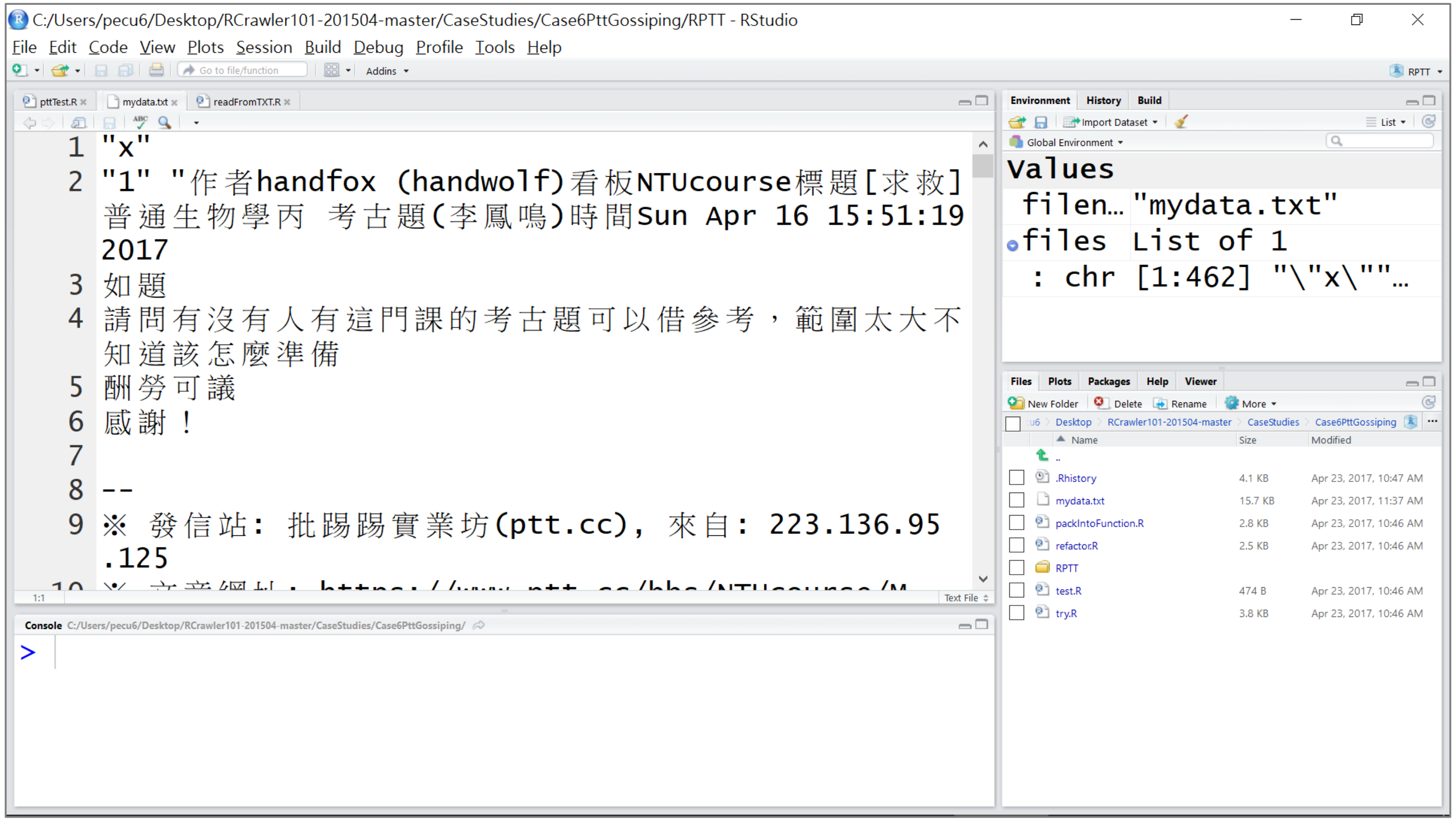
搭配 sapply

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
RPTT
Source
Console C:/Users/pecu6/Desktop/RCrawler101-201504-master/CaseStudies/Case6PttGossiping/
Environment History Build
Import Dataset
Global Environment
a… chr [...
h… chr [...
h… List ...
t… List ...
U… "http...
Functions
Files Plots Packages Help Vie
New Folder Delete Rename
Case6PttGossiping
Name
.Rhistory
packIntoFunction.R
refactor.R
RPTT
test.R
try.R

```
 19:08\... <truncated>
[15] "作者benck (小倫)看板NTUcourse標題[整理] 常見Q&A 發文前必看時間Tue Feb
 23 14:20:02 2010\n初選\n\nQ：二階初選可以選衝堂的課嗎？ (2011/1/25 update
)\nA：不可以。\n\n\n開學後加退選\n\nQ：開學後可以加選衝堂的課嗎？(第3類課程) (2
011/1/25 update)\nA：可以，若與已選上的課程衝突，新加選的課程被分發上了之後，舊
課程會自動退選，\n    以俗語來說就是騎驢找馬。若新選的課程沒有上，舊課程還會在。\n
\n\n全學年課程 (2011/1/26 update)\n\nQ：若為一學年課程，僅修習一學期，或是僅
有一學期及格，是否可計為畢業學分？\nA：依各系規定。http://0rz.tw/2juE2 (大部份
皆可)\n\nQ：若為一學年課程，一學期不及格，但兩學期平均後及格，是否兩學期都算及格？
\nA：上下學期分為兩門課，課名不同，無法平均計算。\n\n\n共同必修(國文、外文)\n\nQ
：外文領域可以不要選英文嗎？\nA：依各系規定。http://0rz.tw/2juE2\n\nQ：超修的
外文課程可以計入選修學分嗎？\nA：依各系規定。http://0rz.tw/2juE2\n\nQ：我可以
申請大一英文免修嗎？免修可以抵學分嗎？\nA：自98學年起試辦，每年度入學新生可申請，
但不可抵免，需另修6學分選修學分。\n    自99學年起規定更改，申請免修通過者6學分可直
接取得不需額外修課。\n\n\n體育\n\nQ：體育三/四可以選同課名且同老師的課嗎？ (201
1/1/25 update)\nA：若體育三及四同課名，抑或是同課名且同老師，兩門體育課程都是承
認的唷！\n\n\n通識\n\nQ：超修的通識學分可以計入選修學分嗎？\nA：依各系規定。htt
p://0rz.tw/2juE2\n\nQ：修習非本系指定類別的通識，可以抵系外選修學分嗎？\nA：有
標示星號的才可以，如A7*。無標示*的純通識課程不可抵。\n\n\n服務課\n\nQ：服務課一
```

本機端純文字檔本蒐集


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
RPTT
pttTest.R
mydata.txt
readFromTXT.R
1:1
Text File
Environment History Build
Import Dataset
List
Global Environment
Files Plots Packages Help Viewer
New Folder Delete Rename More
Console

```
"x"
"1" "作者handfox (handwolf)看板NTUcourse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
```

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

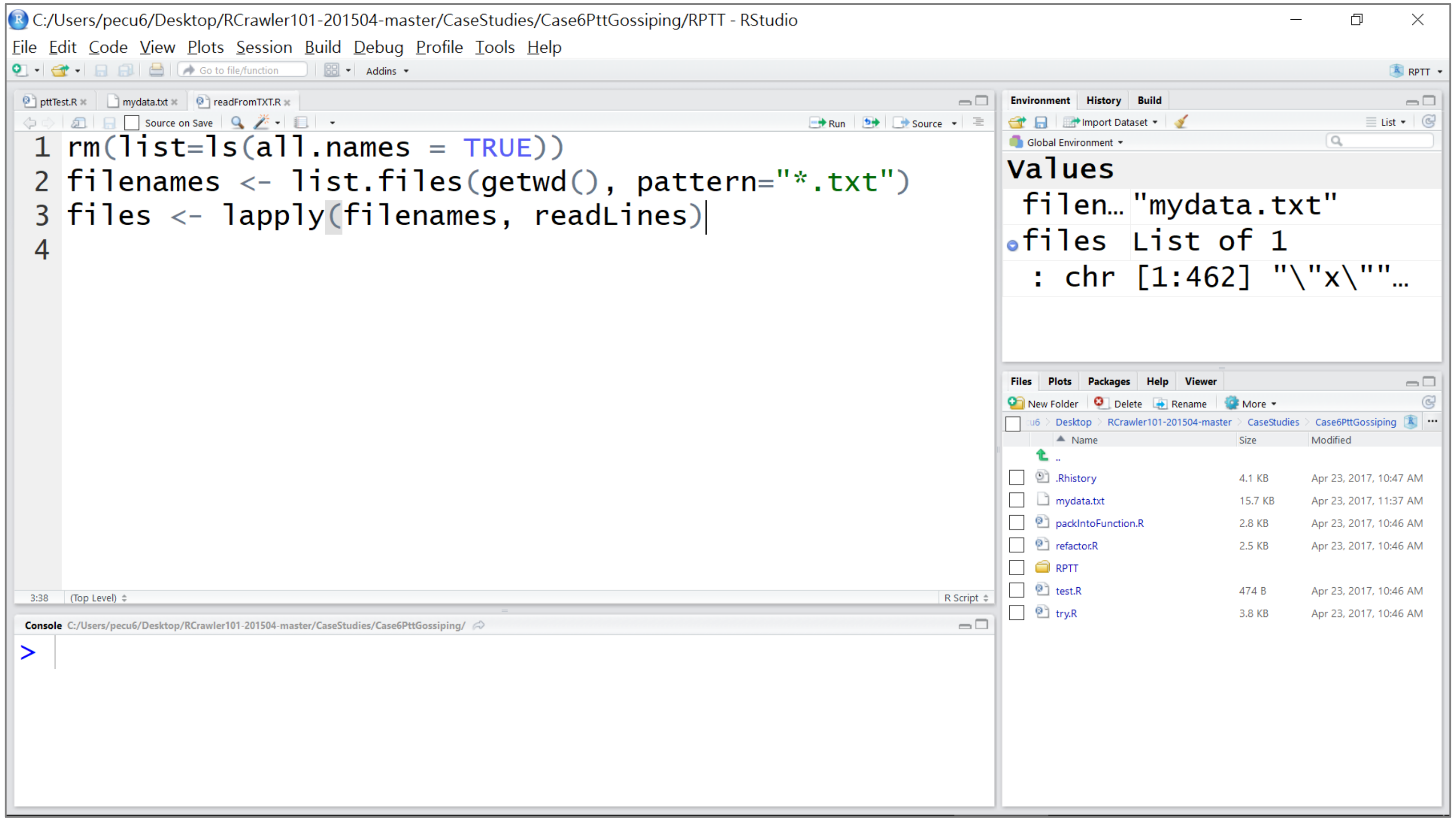
```
>
```

```
Values
 filen… "mydata.txt"
 files  List of 1
  : chr [1:462] "\"x\""…
```

u6 > Desktop > RCrawler101-201504-master > CaseStudies > Case6PttGossiping

Name	Size	Modified
..		
.Rhistory	4.1 KB	Apr 23, 2017, 10:47 AM
mydata.txt	15.7 KB	Apr 23, 2017, 11:37 AM
packIntoFunction.R	2.8 KB	Apr 23, 2017, 10:46 AM
refactor.R	2.5 KB	Apr 23, 2017, 10:46 AM
RPTT		
test.R	474 B	Apr 23, 2017, 10:46 AM
try.R	3.8 KB	Apr 23, 2017, 10:46 AM

本機端純文字檔本蒐集



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

```r
rm(list=ls(all.names = TRUE))
filenames <- list.files(getwd(), pattern="*.txt")
files <- lapply(filenames, readLines)
```

Values

```
filen… "mydata.txt"
files  List of 1
 : chr [1:462] "\"x\""…
```

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

```
>
```

u6 > Desktop > RCrawler101-201504-master > CaseStudies > Case6PttGossiping

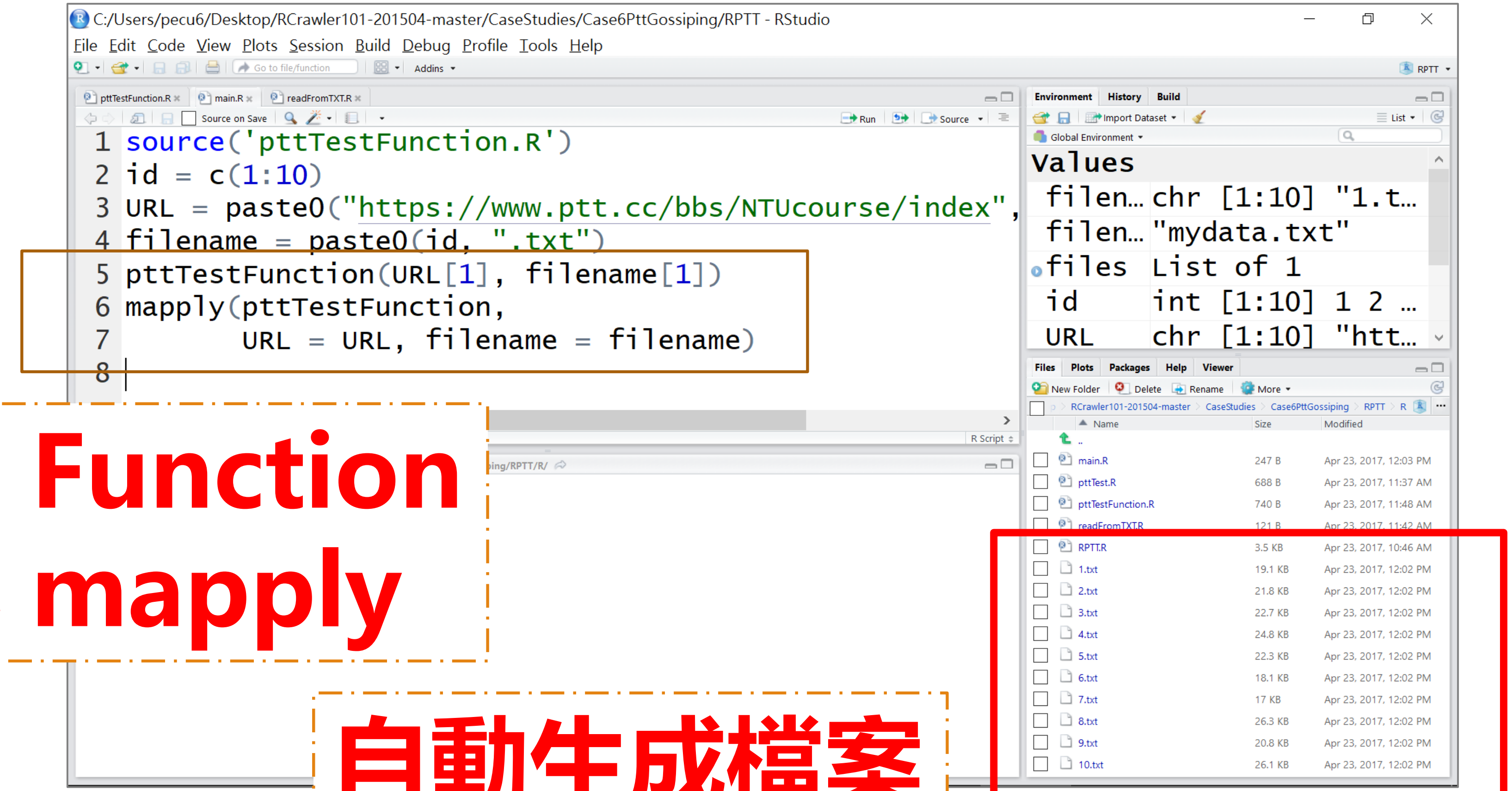
Name	Size	Modified
..		
.Rhistory	4.1 KB	Apr 23, 2017, 10:47 AM
mydata.txt	15.7 KB	Apr 23, 2017, 11:37 AM
packIntoFunction.R	2.8 KB	Apr 23, 2017, 10:46 AM
refactor.R	2.5 KB	Apr 23, 2017, 10:46 AM
RPTT		
test.R	474 B	Apr 23, 2017, 10:46 AM
try.R	3.8 KB	Apr 23, 2017, 10:46 AM

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



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
RPTT
pttTestFunction.R
main.R
readFromTXT.R
Source on Save
Run
Source
R Script
Environment
History
Build
Import Dataset
List
Global Environment
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
RCrawler101-201504-master > CaseStudies > Case6PttGossiping > RPTT > R

```r
source('pttTestFunction.R')
id = c(1:10)
URL = paste0("https://www.ptt.cc/bbs/NTUcourse/index",
filename = paste0(id, ".txt")
pttTestFunction(URL[1], filename[1])
mapply(pttTestFunction,
         URL = URL, filename = filename)
```

```
Values
 filen… chr [1:10] "1.t…
 filen… "mydata.txt"
 files  List of 1
 id     int [1:10] 1 2 …
 URL    chr [1:10] "htt…
```

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.R	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

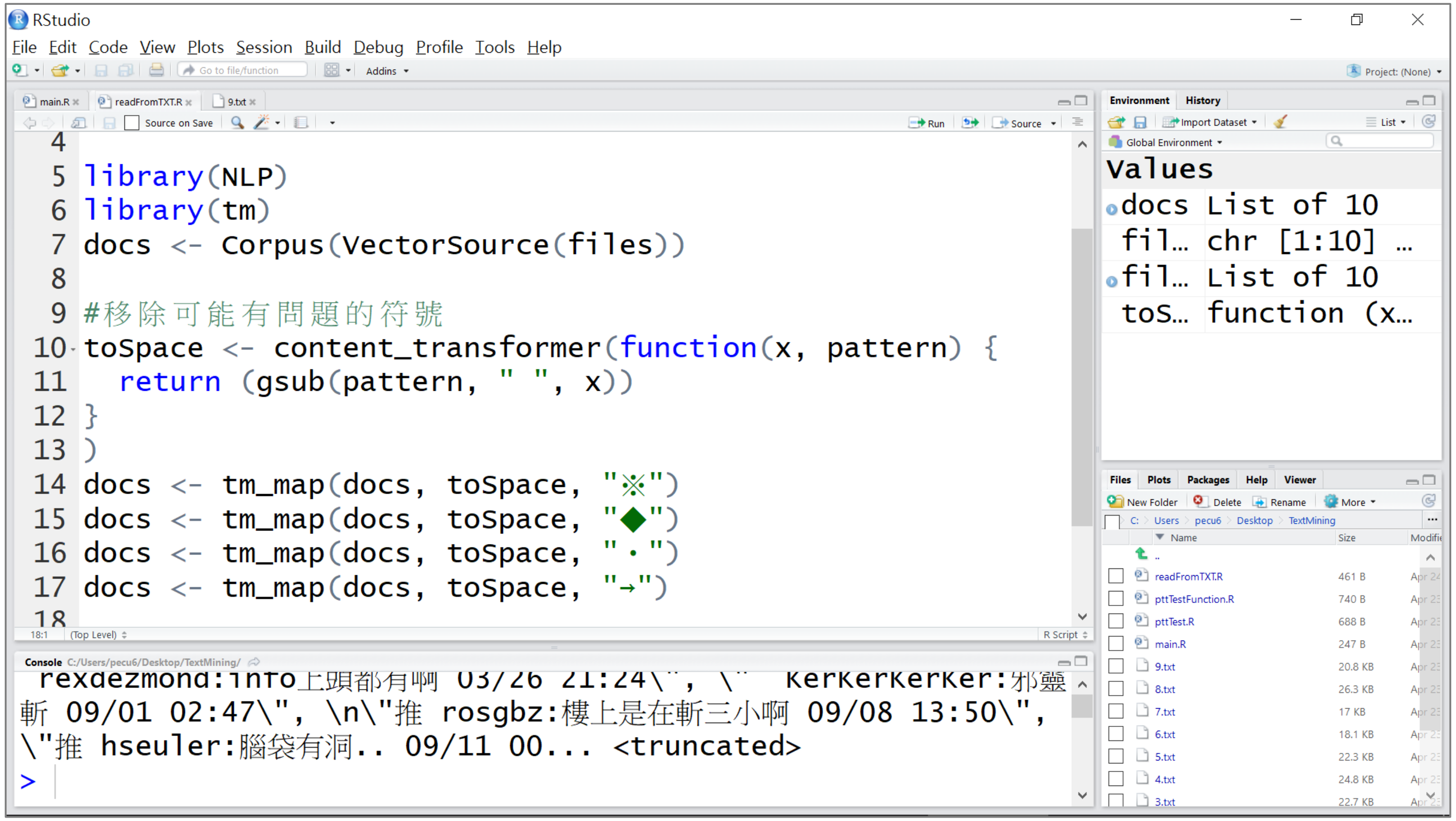
自訂 Function
搭配 mapply

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



```r
library(NLP)
library(tm)
docs <- Corpus(VectorSource(files))

#移除可能有問題的符號
toSpace <- content_transformer(function(x, pattern) {
  return (gsub(pattern, " ", x))
}
)
docs <- tm_map(docs, toSpace, "※")
docs <- tm_map(docs, toSpace, "◆")
docs <- tm_map(docs, toSpace, "・")
docs <- tm_map(docs, toSpace, "→")
```

Console C:/Users/pecu6/Desktop/TextMining/

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

Values

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

C: › Users › pecu6 › Desktop › TextMining

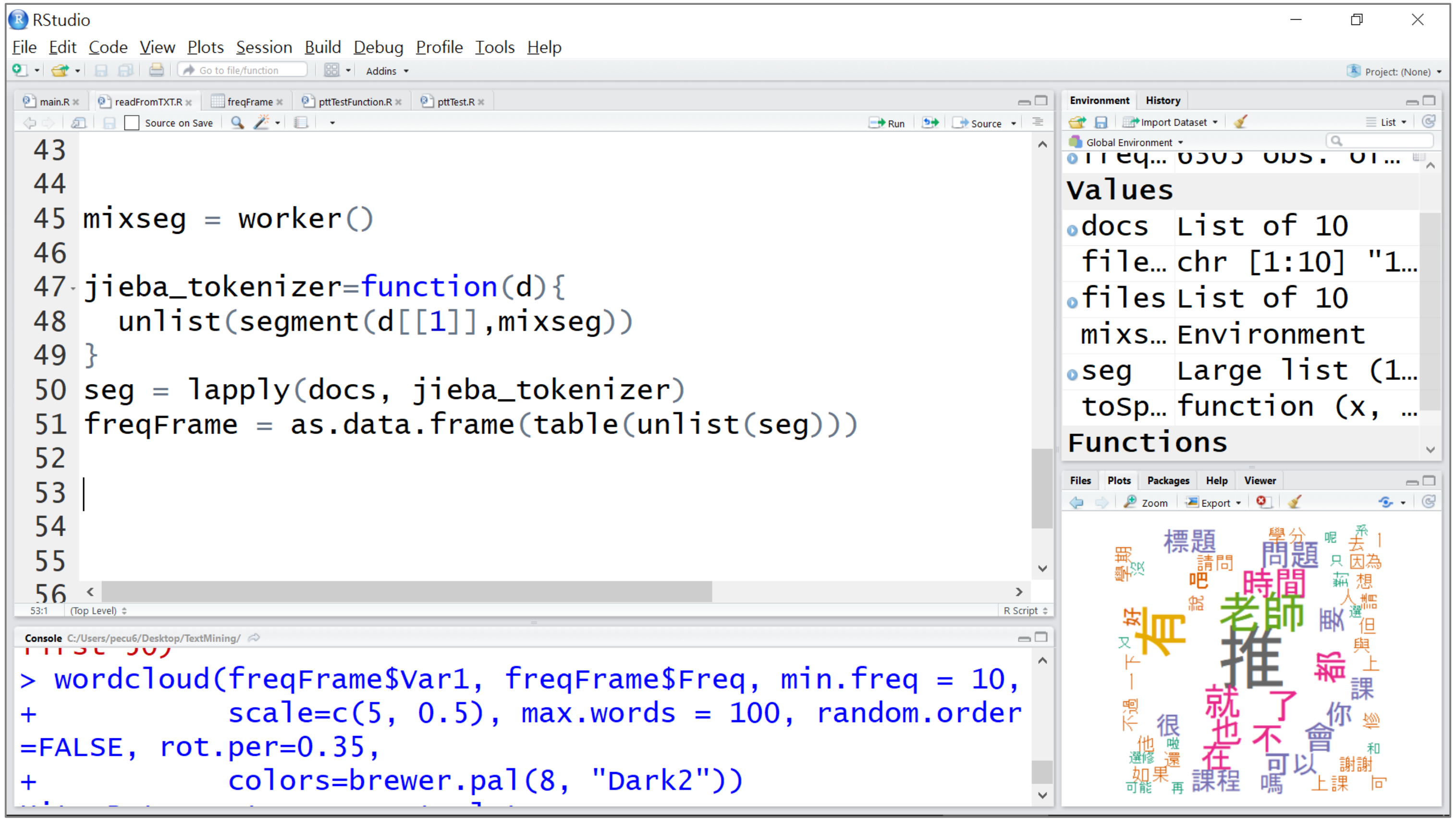
Name	Size	Modifie
..		
readFromTXT.R	461 B	Apr 24
pttTestFunction.R	740 B	Apr 23
pttTest.R	688 B	Apr 23
main.R	247 B	Apr 23
9.txt	20.8 KB	Apr 23
8.txt	26.3 KB	Apr 23
7.txt	17 KB	Apr 23
6.txt	18.1 KB	Apr 23
5.txt	22.3 KB	Apr 23
4.txt	24.8 KB	Apr 23
3.txt	22.7 KB	Apr 23

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

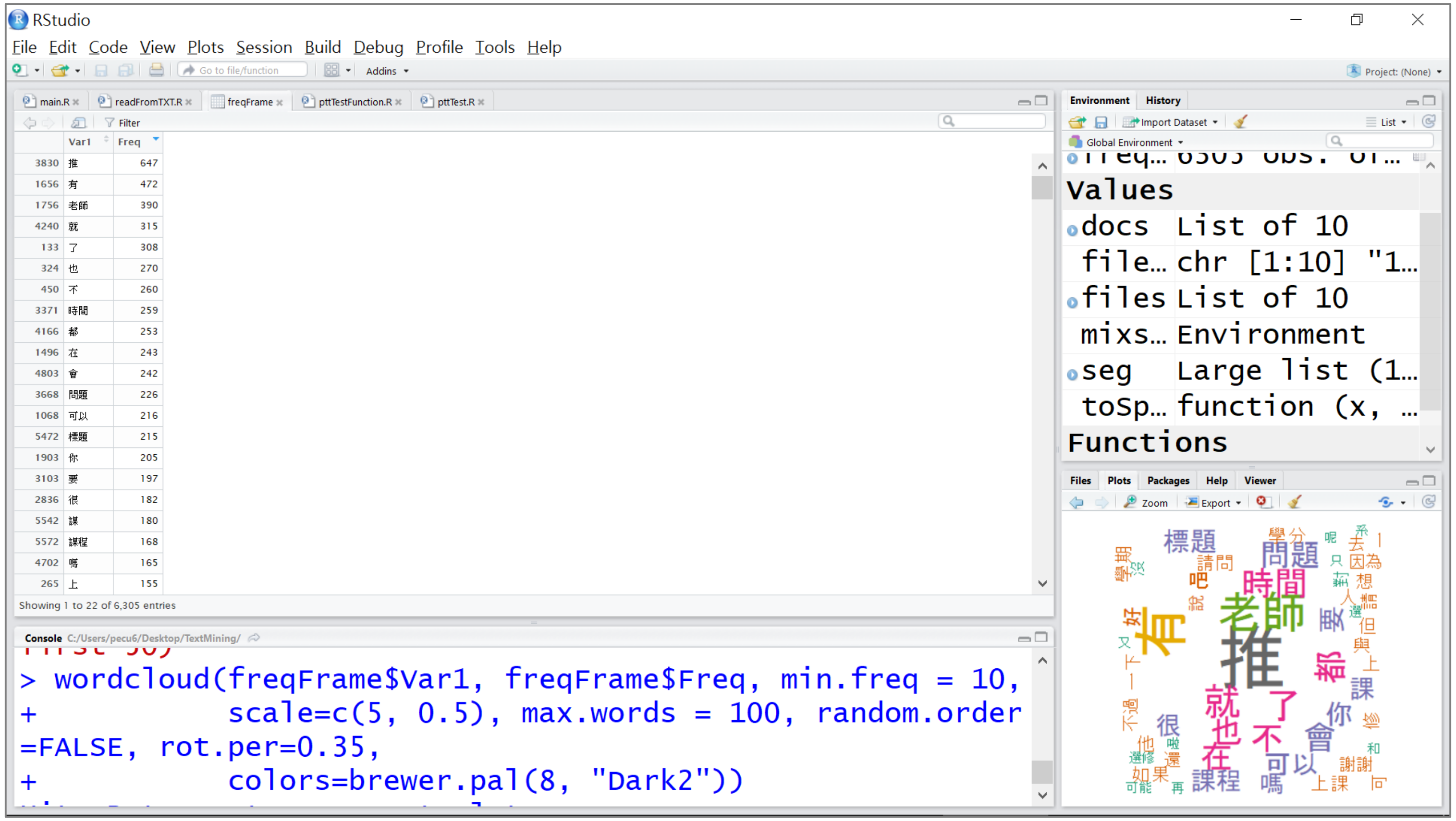


```
mixseg = worker()

jieba_tokenizer=function(d){
  unlist(segment(d[[1]],mixseg))
}
seg = lapply(docs, jieba_tokenizer)
freqFrame = as.data.frame(table(unlist(seg)))
```

```
freq… 6565 obs. of…
Values
docs   List of 10
file…  chr [1:10] "1…
files  List of 10
mixs…  Environment
seg    Large list (1…
toSp…  function (x, …
Functions
```

```
> wordcloud(freqFrame$Var1, freqFrame$Freq, min.freq = 10,
+           scale=c(5, 0.5), max.words = 100, random.order
=FALSE, rot.per=0.35,
+           colors=brewer.pal(8, "Dark2"))
```

詞頻矩陣



	Var1	Freq
3830	推	647
1656	有	472
1756	老師	390
4240	就	315
133	了	308
324	也	270
450	不	260
3371	時間	259
4166	都	253
1496	在	243
4803	會	242
3668	問題	226
1068	可以	216
5472	標題	215
1903	你	205
3103	要	197
2836	很	182
5542	課	180
5572	課程	168
4702	嗎	165
265	上	155

Showing 1 to 22 of 6,305 entries

```
Values
docs    List of 10
file…   chr [1:10] "1…
files   List of 10
mixs…   Environment
seg     Large list (1…
toSp…   function (x, …
Functions
```

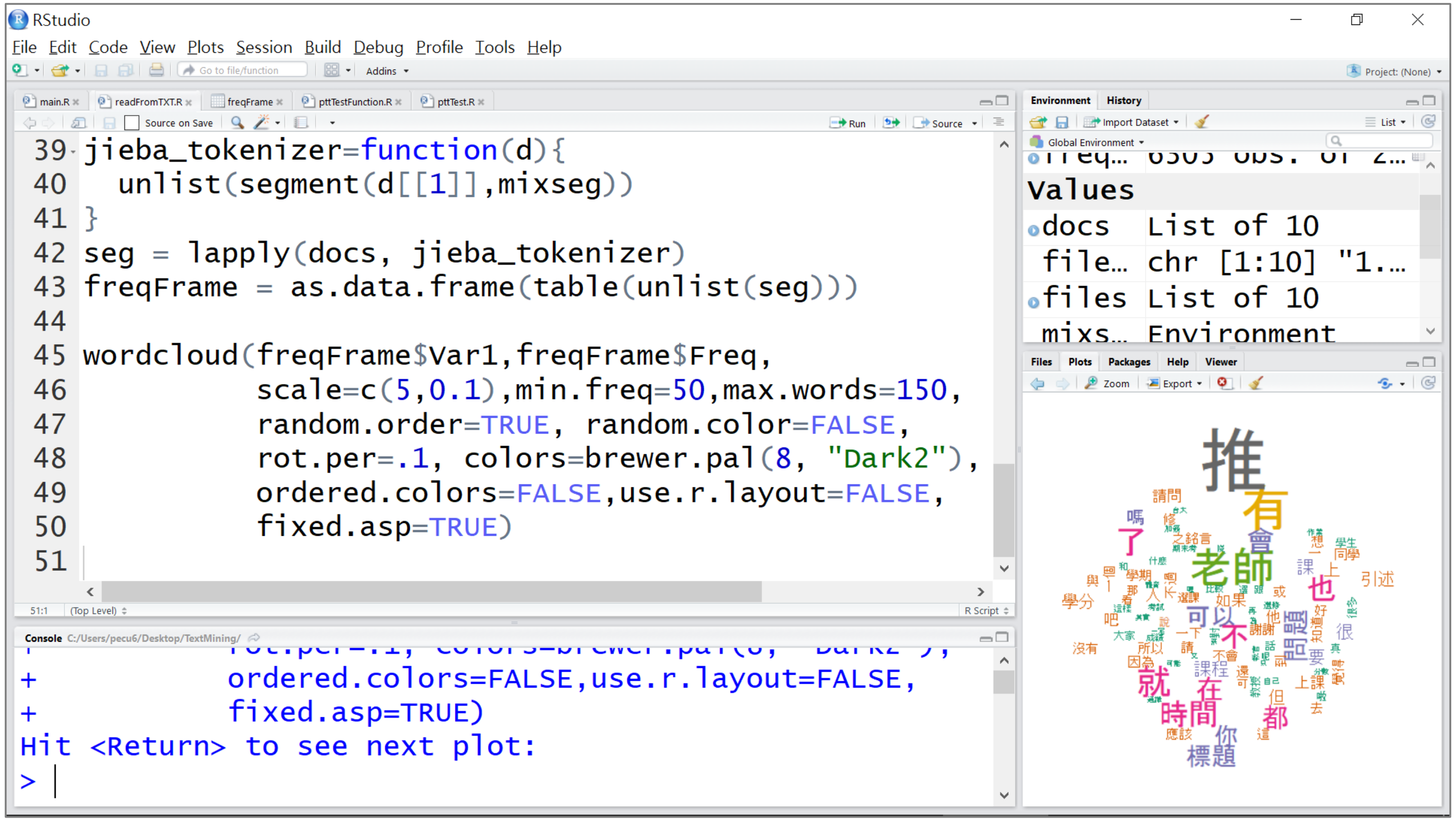
```
> wordcloud(freqFrame$Var1, freqFrame$Freq, min.freq = 10,
+           scale=c(5, 0.5), max.words = 100, random.order
=FALSE, rot.per=0.35,
+           colors=brewer.pal(8, "Dark2"))
```

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



```
jieba_tokenizer=function(d){
  unlist(segment(d[[1]],mixseg))
}
seg = lapply(docs, jieba_tokenizer)
freqFrame = as.data.frame(table(unlist(seg)))

wordcloud(freqFrame$Var1,freqFrame$Freq,
          scale=c(5,0.1),min.freq=50,max.words=150,
          random.order=TRUE, random.color=FALSE,
          rot.per=.1, colors=brewer.pal(8, "Dark2"),
          ordered.colors=FALSE,use.r.layout=FALSE,
          fixed.asp=TRUE)

```

```
Values
docs    List of 10
file…   chr [1:10] "1.…
files   List of 10
mixs…   Environment
```

```
+           ordered.colors=FALSE,use.r.layout=FALSE,
+           fixed.asp=TRUE)
Hit <Return> to see next plot:
> 
```

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