



# R爬蟲應用案例

## 以PTT美食版(Food版)為例

主講人：周淑樺



# About me

周淑樺 / Sharon



## 研究領域及專長

- » 文字探勘
- » 網路爬蟲
- » 關聯分析
- » R語言、Java、Python、C++、html

## 聯絡方式

- » [sharon.chou127@gmail.com](mailto:sharon.chou127@gmail.com)
- » [whtiegg01270127@gmail.com](mailto:whtiegg01270127@gmail.com)

## 學經歷

- » 臺北商業大學 資訊與決策科學研究所
- » 聯合大學 資訊管理學系



# Outline

- » 爬蟲 with R 簡介
- » R爬蟲相關套件介紹
- » PTT-美食版應用案例分享
- » Q&A





1.


# R爬蟲簡介





## 何謂網路爬蟲？

是一種用來自動瀏覽全球資訊網的網路  
機器人。其目的一般為編纂網路索引。





# 2.

## R爬蟲相關套件介紹






## 相關套件 {XML}


» 下載後的內容是HTML格式，無法直接分析。所以再利用XML套件的readHTMLTable和XPath的功能來將需要的資訊從文件中萃取出來。

» 常用函式：

- ◇ htmlParse
  - ◇ xpathSApply
- 



## 相關套件 {RCurl}

- » Rcurl是提供R使用網際網路上各種通訊協定的工具。
  - » 常用函式：
    - ◇ getURL
- 





3.

PTT-美食版應用案例分享



## 批踢踢實業坊 > 看板 Food

聯絡資訊 關於我們

看板

精華區

最舊

< 上頁

下頁 >

最新

[食記] 高雄 順億鮭魚專賣店 不錯吃的鮭魚盛合

3/27 kamgx58

[食記] 台南 六甲區 阿袍羊肉

3/27 reesion

1 [食記] 來台的韓國庶民美食--孔陵一隻雞

3/27 kenny53

[實宣] 太溪 基隆全家福海鮮餐廳 老街旁活魚餐廳

3/27 raingroup

[食記] 韓國 School Food 韓式創意料理好吃

3/27 lovecala

[食記] 台北 願意再排一次隊來吃阜杭豆漿鹹豆漿

3/27 fruittea

[食記] 台南南區 食べ才 / ta be o mu 歐姆蛋蓋飯

3/27 an765433

[食記][台中] 老芋仔芋圓，大坑超高人氣芋圓店

3/27 buuzkuo

[食記] 台北 天母巷弄美食之ISM 主義甜時

3/27 ethe10203

[食記] 台北市三道一鍋~黃金昆布湯頭和鮮美食材

3/27 j19617

(本文已被刪除) [ativan]

```
> #### Get the last page Number
> lastpage <- unlist(xpathApply(htmlParse(getURL(paste0("https://www.ptt.cc/bbs/Food/index.html"))),
  "//div[@class='btn-group btn-group-paging']/a",xmlGetAttr, "href"))[[2]]
> lastpage <- gsub(".*index", "", lastpage)
> lastpage <- as.numeric(gsub("[.]*html", "", lastpage))+1
> lastpage
[1] 6003
```

看板 Food 文章列表 x timberland - Google x 應用多重代理人系統 x fb170327152018 x 通用文字探勘技術 x PecuClub x 網路爬蟲系列 (Cra x 瀏覽

安全 | https://www.ptt.cc/bbs/Food/index.html

應用程式 Google 費用 IE Tab 程式 研究所 IOS 吃貨清單 放鬆專用 從 IE 匯入

## 批踢踢實業坊 > 看板 Food

div.btn-group.btn-group-paging | 323.41 x 40 我們

看板 精華區

最舊 < 上頁 下頁 > 最新

[食記] 高雄 順億鮭魚專賣店 不錯吃的鮭魚盛合  
3/27 kamgx58

[食記] 台南 六甲區 阿袍羊肉  
3/27 reesion

1 [食記] 來台的韓國庶民美食--孔陵一隻雞  
3/27 kenny53

[廣宣] 太溪 基隆全家福海鮮餐廳 老街旁活魚餐廳  
3/27 raindrop

[食記] 韓國 School Food 韓式創意料理好吃  
3/27 lovecala

[食記] 台北 願意再排一次隊來吃阜杭豆漿鹹豆漿  
3/27 fruittea

[食記] 台南南區 食べオムた be o mu歐姆蛋蓋飯

Elements Console Sources Network >>

```
> #shadow-root (open)
> <head>...</head>
> <body>
  > <div id="topbar-container">...</div>
  > <div id="main-container">
    > <div id="action-bar-container">
      > <div class="btn-group btn-group-dir">...</div>
      > <div class="btn-group btn-group-paging"> == $0
        > <a class="btn wide" href="/bbs/Food/index1.html">最舊</a>
        > <a class="btn wide" href="/bbs/Food/index6002.html"><上頁</a>
        > <a class="btn wide disabled">下頁 ></a>
        > <a class="btn wide" href="/bbs/Food/index.html">最新</a>
      </div>
    </div>
  </div>
  > <div class="r-list-container action-bar-margin bbs-screen">
    > <div class="r-ent">...</div>
    > <div class="r-ent">...</div>
    > <div class="r-ent">...</div>
    > <div class="r-ent">...</div>
```

看板 精華區 最舊 < 上頁 下頁 > 最新

[食記] 高雄 順億鮪魚專賣店 不錯吃的鮪魚盛合  
3/27 kamgx58

[食記] 台南 六甲區 阿袍羊肉  
3/27 reesion

1 [食記] 來台的韓國庶民美食-孔陵一隻雞  
3/27 kenny53

[廣宣] 大溪 基隆全家福海鮮餐廳 老街旁活魚餐廳  
3/27 raindrop

[食記] 韓國 School Food 韓式創意料理好吃  
3/27 lovecala

[食記] 台北 願意再排一次隊來吃阜杭豆漿鹹豆漿  
3/27 fruittea

[食記] 台南南區 食べ才 ta be o mu 歐姆蛋蓋飯  
3/27 an765433

[食記][台中] 老芋仔芋圓, 大坑超高人氣芋圓店  
3/27 buuzkuo

[食記] 台北 天母巷弄美食之ISM 主義甜時  
3/27 ethe10203

[食記] 台北市三道一鍋~黃金昆布湯頭和鮮美食材  
3/27 j19617

(本文已被刪除) [ativan]

Elements Console Sources Network Timeline Profiles Application Security

```
<a class="btn wide disabled">下頁 </a>
<a class="btn wide" href="/bbs/Food/index.html">最新</a>
</div>
</div>
</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/Food/M.1490549647.A.1D1.html">[食記] 高雄 順億鮪魚專賣店 不錯吃的鮪魚盛合</a> == $0
    </div>
    <div class="meta">...</div>
  </div>
  <div class="r-ent">...</div>
  <div class="r-ent">...</div>
  <div class="r-ent">
    <div class="nrec"></div>
    <div class="mark"></div>
    <div class="title">
      <a href="/bbs/Food/M.1490553757.A.3AB.html">[廣宣] 大溪 基隆全家福海鮮餐廳 老街旁活魚餐廳</a>
    </div>
  </div>
</div>
```

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

Styles Event Listeners DOM Breakpoints Properties

Filter :hov .cls +

```
element.style {
}
a:visited {
  margin: -
  border: -
  padding: -
}
```

bbs-base.css:132



```
#### Get link form each pages
```

```
link.Food <- NULL
```

```
for( i in (lastpage-10):lastpage){ # 先抓最新的10頁
```

```
  url <- paste0("https://www.ptt.cc/bbs/Food/index", i, ".html")
```

```
  html <- htmlParse(getURL(url))
```

```
  url.list <- xpathSApply(html, "//div[@class='title']/a[@href]", xmlAttrs)
```

```
  link.Food <- c(link.Food, paste('https://www.ptt.cc', url.list, sep=''))
```

```
  print(paste("Get url from the billboard's(Food) page :", i))
```

```
}
```



link.Food	chr [1:215] "https://www.ptt.cc/bbs/Food/M.1..."
-----------	--

```
#### Write a function to save documents
getdoc <- function(link, path){
  doc <- xpathSapply(htmlParse(getURL(link), encoding="UTF-8"), "//div[@id='main
-content']", xmlValue)
  name <- strsplit(link, '/')[[1]][6]
  write(doc, file=file.path(path, gsub('html', 'txt', name)))
}
#### Set the path where you want to save documents
system.time(sapply(1:length(link.Food), function(i) getdoc(link.Food[i], path="C
:/Users/IDS/Desktop/Food")))
```

The screenshot shows a web browser window with the address bar displaying <https://www.ptt.cc/bbs/Food/M.1490549647.A.1D1.html>. The browser's developer tools are open, showing the HTML structure of the page.

**Forum Post Content:**

div#main-content.bbs-screen.bbs-content | 504 x 3952 聯絡資訊 關於我們

作者 kamgx58 (壽司羊) 標題 [食記] 高雄 順億鮑魚專賣店 不錯吃的鮑魚盛合 時間 Mon Mar 27 01:34:03 2017

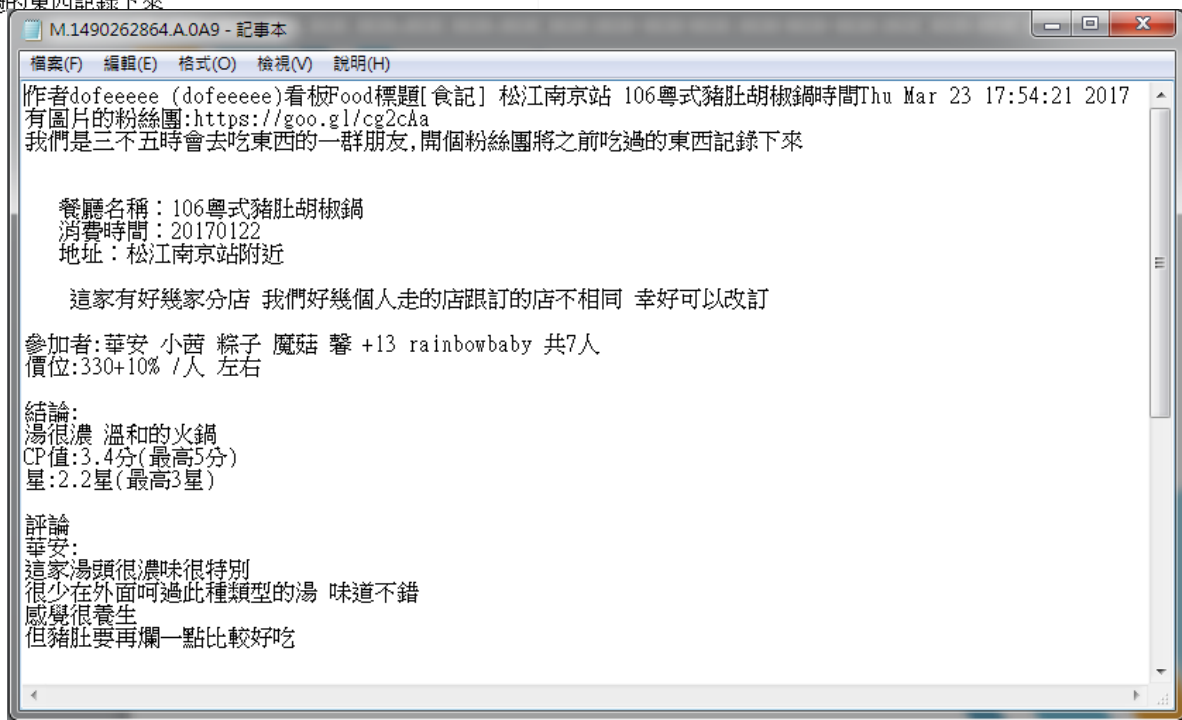
餐廳名稱：順億鮑魚專賣店  
 消費時間：2017年 / 3月  
 地址：高雄市鼓山區南屏路583號  
 電話：07 522 3738  
 營業時間：11:00~21:30  
 每人平均價位：400  
 可否刷卡：可以  
 推薦菜色：鮑盛合脂套餐

圖文網誌  
 版：http://lntzuyang79.pixnet.net/blog/post/340836465

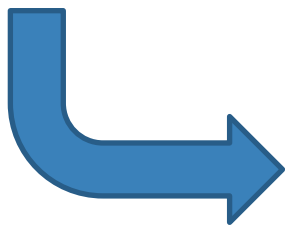
**HTML Source Code Snippets:**

```
<div id="fb-root" class=" fb_reset">...</div>
<script src="https://apis.google.com/_scs/apps-static/_js/k-oz.gapi.zh_TW.3Pd...d=1/ed=1/
am=AQ/rs=AGLTcCMpqtVgDFMoza6-V2qEgetXMSuZA/cb=gapi.loaded_1" async></script>
<script src="https://apis.google.com/_scs/apps-static/_js/k-oz.gapi.zh_TW.3Pd...d=1/ed=1/
am=AQ/rs=AGLTcCMpqtVgDFMoza6-V2qEgetXMSuZA/cb=gapi.loaded_0" async></script>
<script async src="https://www.google-analytics.com/analytics.js"></script>
<script type="text/javascript" async src="https://apis.google.com/js/plusone.js"
gapi_processed="true"></script>
<script id="facebook-jssdk" src="//connect.facebook.net/en_US/all.js#xfbml=1"></script>
<script>...</script>
<div id="topbar-container">...</div>
<div id="navigation-container">...</div>
<div id="main-container">
...
<div id="main-content" class="bbs-screen bbs-content">...</div> == $0
<div id="article-polling" data-pollurl="/poll/Food/M.1490549647.A.1D1.html?cacheKey=2051-
193379362&offset=4927&offset-sig=743eef6b2e48301fa4a897538b9819ab96d5abfc" data-longpollurl=
"/v1/longpoll?id=84db1bfec8fb8e0dcca555506c37b27c8783ef3" data-offset="4927">推文自動更新已
關閉</div>
```

```
> #####  
> #       Read Files       #  
> #####  
> library(tmcn)           # require tm 0.1-4 version  
> library(tm)  
> ##### Put the documents' directory  
> d.corpus <- Corpus(DirSource("C:/Users/IDS/Desktop/Food"), list(language = NA))  
> content(d.corpus[[1]])  
[1] "作者dofeeeee (dofeeeee)看板Food標題[食記] 松江南京站 106粵式豬肚胡椒鍋時間Thu Mar 23 17:54  
:21 2017"  
[2] "有圖片的粉絲團:https://goo.gl/cg2cAa"  
  
[3] "我們是三不五時會去吃東西的一群朋友,開個粉絲團將之前吃過的東西記錄下來"  
  
[4] ""  
[5] ""  
[6] " 餐廳名稱:106粵式豬肚胡椒鍋"  
[7] " 消費時間:20170122"  
[8] " 地址:松江南京站附近"  
[9] ""  
[10] " 這家有好幾家分店 我們好幾個人走的店跟訂的店不相同
```



```
#### Remove Punctuation and Numbers from corpus
d.corpus <- tm_map(d.corpus, removePunctuation) # 移除標點符號
d.corpus <- tm_map(d.corpus, removeNumbers) # 移除數字
d.corpus <- tm_map(d.corpus, function(word) {
  gsub("[A-Za-z0-9]", "", word)
}) # 移除大小寫英文
inspect(d.corpus) # 預覽內文
```



```
[[211]]
[1] 作者 五樓愛自宮X看板標題請益 想問各位吃過海底撈的平均一人多少錢時間
[2] 小弟前幾天約了一群朋友吃海底撈
[3]
[4] 位置已經訂好了 但有人覺得單點店會太貴
[5]
[6] 想知道有吃過海底撈的鄉民
[7]
[8] 平均分擔下來一個人大概多少錢
[9]
[10]
[11] 發信站 批踢踢實業坊 來自
[12] 文章網址
[13] 推 越多人分湯底越便宜 我跟朋友兩個人一個人
[14] 小弟一行人出頭
[15] 編輯
[16] 推 個人人千
[17] 推
[18] 推 上次跟我女友兩人吃了兩千八 不過我跟他食量都算大
[19] 推 我跟男友也是一人一千左右
[20]
```



```
#### Using Rwordseg or jiebaR package to break down Chines. Here, we using Rwordseg
d.corpus <- sapply(1:length(d.corpus), function(u) {
  segmentCN(as.String(unlist(d.corpus[u])), nosymbol=F)})
```

```
Sentence| <- sapply(1:length(d.corpus), function(u) paste(d.corpus[[u]], collapse=" "))
```

```
d.corpus <- Corpus(VectorSource(d.corpus))
```

```
> Sentence[1]
```

[1] "作者 看板 標題 食記 松江 南京 站 粵 式 豬肚 胡椒 鍋 時間 有 圖片 的 粉絲 團 我們 是 三 不 五 時 會 去 吃 東西 的 一 群 朋友 開 個 粉絲 團 將 之前 吃 過 的 東西 記錄 下來 餐廳 名稱 粵 式 豬肚 胡椒 鍋 消費 時間 地址 松江 南京 站 附近 這 家 有 好 幾 家 分店 我們 好 幾 個 人 走 的 店 跟 訂 的 店 不 相同 幸好 可以 改 訂 參加 者 華 安 小 茜 粽子 魔 菇 馨 共 人 價位 人 左右 結論 湯 很 濃 溫和 的 火鍋 值 分 最高 分 星星 最高 星 評論 華 安 這 家 湯 頭 很 濃 味 很 特別 很少 在 外面 呵 過 此 種 類型 的 湯 味道 不錯 感覺 很 養生 但 豬肚 要 再 爛 一 點 比較 好吃 小 茜 湯 頭 不錯 很 濃 湯 和 料理 吃 起來 不錯 粽子 食材 什麼 都 不錯 缺點 就是 對於 口味 較 淡 的 人 比較 不 適合 蘑菇 手工 蛋 餃 厚 而且 胖 呼呼 的 很好 吃 華 安 也 大 推 蛋 餃 適合 冬天 吃 的 豬肚 胡椒 湯 吃 完 全身 熱 呼呼 馨 湯 濃 味 溫 醇 味道 好 發 信 站 批 踢 踢 實 業 坊 來自 文章 網址"

```
#### Set the stopwords and remove them
```

```
myStopWords <- c(toTrad(stopwordsCN()), stopwords("english"), "編輯", "時間", "標題",  
"發信", "實業", "作者", "看板")
```

```
d.corpus <- tm_map(d.corpus, removeWords, myStopWords)  
content(d.corpus[[1]])
```

```
> content(d.corpus[[1]])
```

[1]	""	"看"	"板"	""	"食"	"記"	"松江"	"南京"	"站"
[10]	"粵"	"式"	"豬肚"	"胡椒"	"鍋"	""	""	"圖片"	""
[19]	"粉絲"	"團"	""	""	"三"	"不"	"五"	"時"	"會"
[28]	"去"	"吃"	"東西"	""	"一群"	"朋友"	"開"	""	"粉絲"
[37]	"團"	""	"之前"	"吃"	""	""	"東西"	"記錄"	"下來"
[46]	"餐廳"	"名稱"	"粵"	"式"	"豬肚"	"胡椒"	"鍋"	"消費"	""
[55]	"地址"	"松江"	"南京"	"站"	"附近"	"這家"	""	"好"	"幾家"
[64]	"分店"	""	"好幾個"	"人"	"走"	""	"店"	""	"訂"
[73]	""	"店"	"不"	"相同"	"幸好"	""	"改訂"	"參加者"	"華"
[82]	"安"	"小"	"茜"	"粽子"	"魔"	"菇"	"馨"	"共"	"人"
[91]	"價位"	"人"	"左右"	"結論"	"湯"	"很"	"濃"	"溫和"	""
[100]	"火鍋"	"值"	"分"	"最高分"	"星星"	"最高"	"星"	"評論"	"華"
[109]	"安"	"這家"	"湯頭"	"很"	"濃"	"味"	"很"	"特別"	"很"
[118]	"少"	""	"外面"	""	""	""	""	"類型"	""
[127]	"湯"	"味道"	"不錯"	"感覺"	"很"	"養生"	""	"豬肚"	""
[136]	"再"	"爛"	"一點"	"比較"	"好"	"吃"	"小"	"茜"	"湯頭"
[145]	"不錯"	"很"	"濃湯"	""	"料理"	"吃"	"起來"	"不錯"	"粽子"
[154]	"食"	"材"	""	"都"	"不錯"	"缺"	"點"	""	""
[163]	""	""	"口味"	""	"淡"	""	"人"	"比較"	"不"
[172]	"適合"	"蘑菇"	"手工"	"蛋"	"餃"	"厚"	""	"胖"	"呼呼"
[181]	""	"很"	"好"	"吃"	"華"	"安"	""	"大"	"推"
[190]	"蛋"	"餃"	"適合"	"冬天"	"吃"	""	"豬肚"	"胡椒"	"湯"
[199]	"吃"	"完"	"全身"	"熱呼呼"	"馨"	"湯"	"濃"	"味"	"溫"
[208]	"醇"	"味道"	"好"	""	"站"	"批"	"踢"	"踢"	""
[217]	"坊"	"來自"	"文章"	"網址"	""	""	""	""	""

**TermDocumentMatrix** 指的是關鍵字為列，文件是行的矩陣。  
儲存的數字是關鍵字在這些文件中出現的次數

```
> tdm <- TermDocumentMatrix(d.corpus,  
+                             control = list(wordLengths = c(2, Inf),  
+                             weighting = function(x)  
+                             weightTfIdf(x, normalize = FALSE)))  
> tdm  
<<TermDocumentMatrix (terms: 23801, documents: 211)>>  
Non-/sparse entries: 47820/4974191  
Sparsity           : 99%  
Maximal term length: 153  
Weighting          : term frequency - inverse document frequency (tf-idf)  
  
#### wordcloud  
library(wordcloud)  
m <- as.matrix(tdm)  
v <- sort(rowSums(m), decreasing = TRUE)  
d <- data.frame(word = names(v), freq = v)  
#### you can using rWordCloud package for D3 wordcloud  
library(rWordCloud)  
#require(devtools)  
#install_github('adymimos/rWordCloud')  
library(htmlwidgets)  
d3Cloud(text = d$word, size = d$freq)
```

表示我們挑至少兩個字的詞。

## 文字雲{rwordcloud}輸出





4.

Q&A



Thank you for  
your listening

