成為初級資料分析師 IR 程式設計與資料科學應用

網頁爬蟲

郭耀仁

The world's most valuable resource is no longer oil, but data.

The Economist

大綱

- 網站爬蟲的核心任務
- JSON 格式資料
- HTML 格式資料

網站爬蟲的核心任務

核心任務可以簡單區分為兩個

- 請求資料（requesting data）
- 解析資料（parsing data）

請求資料的運作就像在瀏覽器中輸入網址一般，不過送出請求的管道由瀏覽器改變成為 R 語言程式碼

解析資料的運作則是將請求所得之不同格式資料轉換為 R 資料結構，常見的有

- JSON 格式
- HTML 格式

JSON 格式資料

使用 `jsonlite` 套件中的 `fromJSON()` 函數

`fromJSON()` 函數一次就處理了兩個核心任務：

- 對網站發出請求
- 將請求的回應解析為
    - `data.frame`（Array of JSONs）
    - `list`（JSON）

安裝 jsonlite 套件

- 透過 RStudio 的 Packages 功能頁籤
- 透過 install.packages() 函數

```
install.packages("jsonlite")
```

載入 jsonlite 套件

- 透過 RStudio 的 Packages 功能頁籤
- 透過 library() 函數

```
library("jsonlite")
```

安裝 Chrome 瀏覽器外掛：JSONView

- https://chrome.google.com/webstore/detail/jsonview/chklaanhfefbnpoihckbnefhakgc
  (https://chrome.google.com/webstore/detail/jsonview/chklaanhfefbnpoihckbnefhakg
- 更漂亮地在瀏覽器上預覽 JSON 格式資料

安裝好 JSONView 後用瀏覽器預覽
https://opendata.epa.gov.tw/ws/Data/AQI/?$format=json
(https://opendata.epa.gov.tw/ws/Data/AQI/?$format=json)

擷取 https://opendata.epa.gov.tw/ws/Data/AQI/?$format=json (https://opendata.epa.gov.tw/ws/Data/AQI/?$format=json) 資料

In [ ]:
```
library("jsonlite")

aqi_url <- "https://opendata.epa.gov.tw/ws/Data/AQI/?$format=json"
aqi <- fromJSON(aqi_url)
class(aqi)
```

隨堂練習：全台灣有幾個空氣品質測站?

In [2]:
```
ans
```

81

隨堂練習：列出位於臺北市與新北市的空氣品質測站

In [4]:
```
ans
```

County	SiteName
新北市	富貴角
新北市	永和
新北市	三重
臺北市	陽明
臺北市	大同
臺北市	松山
臺北市	古亭
臺北市	萬華
臺北市	中山
臺北市	士林
新北市	淡水
新北市	林口
新北市	菜寮
新北市	新莊
新北市	板橋
新北市	土城
新北市	新店
新北市	萬里
新北市	汐止

HTML 格式資料

使用 `rvest` 套件中的

- `read_html()` 函數對網站發出請求
- `html_nodes(CSS = )` 定位 HTML 中指定的標記
- `html_text()` 解析標記中的文字
- `html_attr(ATTR)` 解析標記中的屬性

安裝 rvest 套件

- 透過 RStudio 的 Packages 功能頁籤
- 透過 install.packages() 函數

```
install.packages("rvest")
```

載入 `rvest` 套件

- 透過 RStudio 的 `Packages` 功能頁籤
- 透過 `library()` 函數

```
library("rvest")
```

以 `read_html()` 請求 https://www.imdb.com/title/tt4154796 (https://www.imdb.com/title/tt4154796) 資料

In [ ]:
```
library("rvest")

html_doc <- read_html("https://www.imdb.com/title/tt4154796")
class(html_doc)
```

安裝 Chrome 瀏覽器外掛：Selector Gadget

- https://chrome.google.com/webstore/detail/selectorgadget/mhjhnkcfbdhnjickkkdbjoe
(https://chrome.google.com/webstore/detail/selectorgadget/mhjhnkcfbdhnjickkkdbjo
- 查詢 HTML 標記的 CSS Selector

以 html_nodes() 解析 html_doc

In [ ]:
```
# rating
html_doc %>%
  html_nodes(css = "strong span")
```

以 html_text() 解析標記中的文字

In [ ]:

```
# rating
html_doc %>%
  html_nodes(css = "strong span") %>%
  html_text() %>%
  as.numeric()
```

以 html_attr(ATTR) 解析標記中的屬性

In [ ]:
```
# poster image link
html_doc %>%
  html_nodes(css = ".poster img") %>%
  html_attr("src")
```

隨堂練習：寫作一個函數 get_movie_info()

In [6]:
```
movie_url <- "https://www.imdb.com/title/tt4154796"
endgame_movie_info <- get_movie_info(movie_url)
print(endgame_movie_info$title)
print(endgame_movie_info$rating)
print(endgame_movie_info$genre)
print(endgame_movie_info$posterLink)
print(endgame_movie_info$cast)
```

```
[1] "Avengers: Endgame (2019)"
[1] 8.7
[1] "Action"    "Adventure" "Sci-Fi"
[1] "https://m.media-amazon.com/images/M/MV5BMTc5MDE2ODcwNV5BMl5BanBnXkFtZTgwM
zI2NzQ2NzM@._V1_UX182_CR0,0,182,268_AL_.jpg"
 [1] "Robert Downey Jr."    "Chris Evans"          "Mark Ruffalo"
 [4] "Chris Hemsworth"      "Scarlett Johansson"   "Jeremy Renner"
 [7] "Don Cheadle"          "Paul Rudd"            "Benedict Cumberbatch"
[10] "Chadwick Boseman"     "Brie Larson"          "Tom Holland"
[13] "Karen Gillan"         "Zoe Saldana"          "Evangeline Lilly"
```

In [7]:

```
movie_url <- "https://www.imdb.com/title/tt6320628/"
spiderman_movie_info <- get_movie_info(movie_url)
print(spiderman_movie_info$title)
print(spiderman_movie_info$rating)
print(spiderman_movie_info$genre)
print(spiderman_movie_info$posterLink)
print(spiderman_movie_info$cast)
```

```
[1] "Spider-Man: Far from Home (2019)"
[1] 8.2
[1] "Action"    "Adventure" "Sci-Fi"
[1] "https://m.media-amazon.com/images/M/MV5BMGZlNTY1ZWUtYTMzNC00ZjUyLWE0MjQtM
TMxN2E3ODYxMWVmXkEyXkFqcGdeQXVyMDM2NDM2MQ@@._V1_UX182_CR0,0,182,268_AL_.jpg"
 [1] "Tom Holland"         "Samuel L. Jackson"   "Jake Gyllenhaal"
 [4] "Marisa Tomei"        "Jon Favreau"         "Zendaya"
 [7] "Jacob Batalon"       "Tony Revolori"       "Angourie Rice"
[10] "Remy Hii"            "Martin Starr"        "J.B. Smoove"
[13] "Jorge Lendeborg Jr." "Cobie Smulders"      "Numan Acar"
```