成為初級資料分析師 | R 程式設計與資料科學應用

資料輸入輸出

郭耀仁

Working with data provided by R packages is a great way to learn the tools of data science, but at some point you want to stop learning and start working with your own data.

Hadley Wickham

大綱

- 內建資料
- 輸入表格式資料
- 輸入非表格式資料
- 輸出表格式資料
- 輸出非表格式資料

內建資料

豐富的內建資料

以 `data()` 函數觀察

```
In [ ]: data()
```

資料也有說明文件可以讀

```
In [ ]: ?iris # help(iris) will do
        ?cars # help(cars) will do
```

輸入表格式資料

常見的表格式（tabular）資料有哪些?

- 以不同符號分隔變數的文字檔
    - `.txt`
    - `.csv` 以逗號分隔
- Excel 試算表（`.xls`, `.xlsx`）
- Array of JSONs（`.json`）
- 資料庫表格

以不同符號分隔變數的文字檔： .txt

the_avengers.txt

```
character;hero
Tony Stark;Iron Man
Steve Rogers;Captain America
Bruce Banner;Hulk
Thor;Thor
Natasha Romanoff;Black Widow
Clint Barton;Hawkeye
```

創建 the_avengers.txt

In [ ]:

```
the_avengers <- "character;hero
Tony Stark;Iron Man
Steve Rogers;Captain America
Bruce Banner;Hulk
Thor;Thor
Natasha Romanoff;Black Widow
Clint Barton;Hawkeye"
writeLines(the_avengers, "the_avengers.txt")
```

載入 the_avengers.txt

使用 read.table() 函數

```
In [ ]: the_avengers_df <- read.table("the_avengers.txt", sep = ";", header=TRUE)
        the_avengers_df
```

以逗號分隔變數的文字檔： .csv

the_avengers.csv

```
character,hero
Tony Stark,Iron Man
Steve Rogers,Captain America
Bruce Banner,Hulk
Thor,Thor
Natasha Romanoff,Black Widow
Clint Barton,Hawkeye
```

創建 the_avengers.csv

In [ ]:

```
the_avengers <- "character,hero
Tony Stark,Iron Man
Steve Rogers,Captain America
Bruce Banner,Hulk
Thor,Thor
Natasha Romanoff,Black Widow
Clint Barton,Hawkeye"
writeLines(the_avengers, "the_avengers.csv")
```

載入 the_avengers.csv

使用 read.csv() 函數

```
In [ ]: the_avengers_df <- read.csv("the_avengers.csv", sep = ",", header=TRUE)
        the_avengers_df
```

Excel 試算表

`the_avengers.xlsx`

https://r-essentials.s3-ap-northeast-1.amazonaws.com/the_avengers.xlsx (https://r-essentials.s3-ap-northeast-1.amazonaws.com/the_avengers.xlsx)

使用 `readxl` 套件中的 `read_excel()` 函數

- 安裝 `readxl` 套件
- 載入 `readxl` 套件
- 使用 `readxl::read_excel()` 函數

安裝 readxl 套件

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

```r
install.pacakges("readxl")
```

載入 readxl 套件

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

```
library("readxl")
```

載入 the_avengers.xlsx

In [ ]:

```
library(readxl)

the_avengers_df <- read_excel("the_avengers.xlsx")
the_avengers_df
```

Array of JSONs

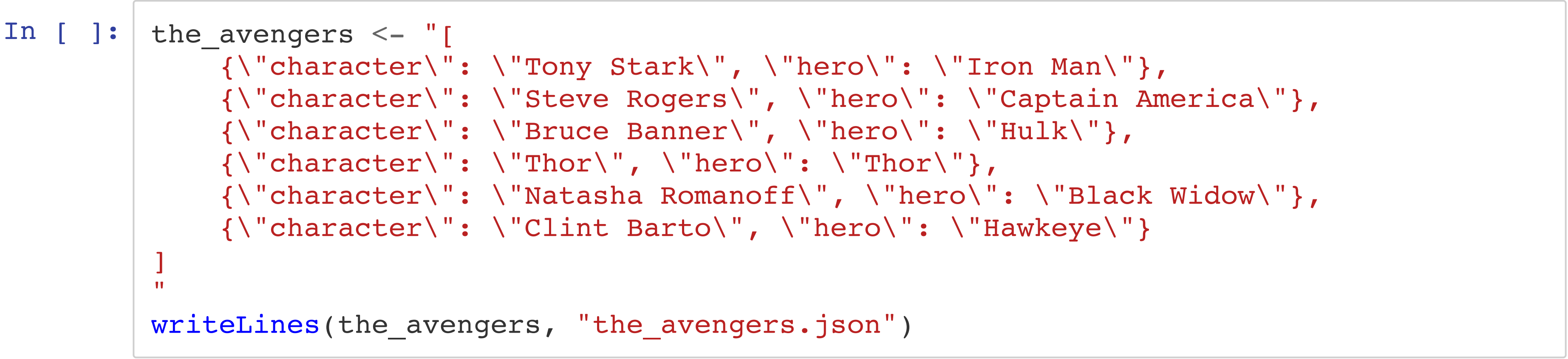
`the_avengers.json`

什麼是 JSON

- JavaScript Object Notation
- 彈性很大且常見於網站資料傳輸的檔案格式
- 它的特性是可以容納不同長度、型別並且巢狀式地（nested）包容資料

```
[
    {"character": "Tony Stark", "hero": "Iron Man"},
    {"character": "Steve Rogers", "hero": "Captain America"},
    {"character": "Bruce Banner", "hero": "Hulk"},
    {"character": "Thor", "hero": "Thor"},
    {"character": "Natasha Romanoff", "hero": "Black Widow"},
    {"character": "Clint Barto", "hero": "Hawkeye"}
]
```

創建 the_avengers.json



In [ ]:

```
the_avengers <- "[
    {\"character\": \"Tony Stark\", \"hero\": \"Iron Man\"},
    {\"character\": \"Steve Rogers\", \"hero\": \"Captain America\"},
    {\"character\": \"Bruce Banner\", \"hero\": \"Hulk\"},
    {\"character\": \"Thor\", \"hero\": \"Thor\"},
    {\"character\": \"Natasha Romanoff\", \"hero\": \"Black Widow\"},
    {\"character\": \"Clint Barto\", \"hero\": \"Hawkeye\"}
]
"
writeLines(the_avengers, "the_avengers.json")
```

使用 `jsonlite` 套件中的 `fromJSON()` 函數

- 安裝 `jsonlite` 套件
- 載入 `jsonlite` 套件
- 使用 `jsonlite::fromJSON()` 函數

安裝 `jsonlite` 套件

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

```
install.pacakges("jsonlite")
```

載入 `jsonlite` 套件

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

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

載入 the_avengers.json

In [ ]:

```
library(jsonlite)

the_avengers_df <- fromJSON("the_avengers.json")
the_avengers_df
```

資料庫表格

暫存於記憶體的 `SQLite` 表格 `the_avengers`

使用 `RSQLite` 套件

- 安裝 `RSQLite` 套件
- 載入 `RSQLite` 套件
- 使用 `DBI::dbWriteTable()` 函數創建表格
- 使用 `DBI::dbReadTable()` 函數載入表格
- 使用 `DBI::dbSendQuery()` 搭配 `DBI::dbFetch()` 函數查詢表格
- 使用 `DBI::dbClearResult()` 函數清除查詢結果
- 使用 `DBI::dbDisconnect()` 函數關閉資料庫連線

安裝 RSQLite 套件

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

```
install.pacakges("RSQLite")
```

載入 RSQLite 套件

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

```r
library("RSQLite")
```

使用 `DBI::dbWriteTable()` 函數創建表格

In [ ]:

```r
library(DBI)

con <- dbConnect(RSQLite::SQLite(), ":memory:")
dbListTables(con)
the_avengers_df <- read.csv("the_avengers.csv", sep = ",", header=TRUE)
dbWriteTable(con, "the_avengers", the_avengers_df)
dbListTables(con)
```

使用 `DBI::dbReadTable()` 函數載入表格

In [ ]:

```
the_avengers_from_db <- dbReadTable(con, "the_avengers")
the_avengers_from_db
```

使用 DBI::dbSendQuery() 搭配 DBI::dbFetch() 函數查詢表格

In [ ]:

```
sql_query <- "SELECT * FROM the_avengers WHERE hero = 'Iron Man';"
res <- dbSendQuery(con, sql_query)
dbFetch(res)
```

完成查詢之後

- 使用 `DBI::dbClearResult()` 函數清除查詢結果
- 使用 `DBI::dbDisconnect()` 函數關閉資料庫連線

In [ ]:

```
dbClearResult(res)
dbDisconnect(con)
```

輸入非表格式資料

常見的非表格式資料有哪些?

- 非表格式的文字檔
- JSON（`.json`）

非表格式的文字檔

endgame_summaries.txt

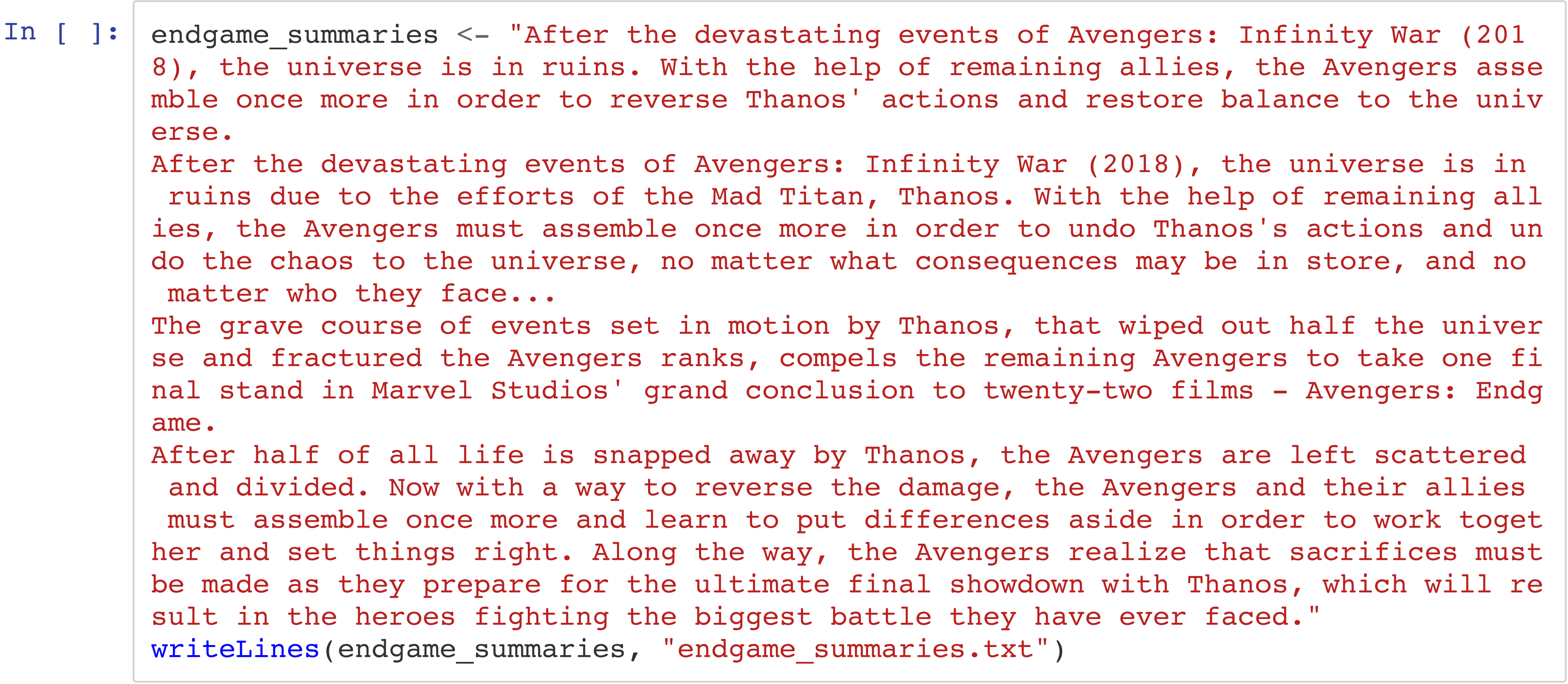
```
After the devastating events of Avengers: Infinity War (2018), the universe is in
ruins. With the help of remaining allies, the Avengers assemble once more in orde
r to reverse Thanos' actions and restore balance to the universe.

After the devastating events of Avengers: Infinity War (2018), the universe is in
ruins due to the efforts of the Mad Titan, Thanos. With the help of remaining all
ies, the Avengers must assemble once more in order to undo Thanos's actions and u
ndo the chaos to the universe, no matter what consequences may be in store, and n
o matter who they face...

The grave course of events set in motion by Thanos, that wiped out half the unive
rse and fractured the Avengers ranks, compels the remaining Avengers to take one
 final stand in Marvel Studios' grand conclusion to twenty-two films - Avengers:
 Endgame.

After half of all life is snapped away by Thanos, the Avengers are left scattered
and divided. Now with a way to reverse the damage, the Avengers and their allies
 must assemble once more and learn to put differences aside in order to work toge
ther and set things right. Along the way, the Avengers realize that sacrifices mu
st be made as they prepare for the ultimate final showdown with Thanos, which wil
l result in the heroes fighting the biggest battle they have ever faced.
```

創建 endgame_summaries.txt



In [ ]:

```r
endgame_summaries <- "After the devastating events of Avengers: Infinity War (2018), the universe is in ruins. With the help of remaining allies, the Avengers assemble once more in order to reverse Thanos' actions and restore balance to the universe.
After the devastating events of Avengers: Infinity War (2018), the universe is in ruins due to the efforts of the Mad Titan, Thanos. With the help of remaining allies, the Avengers must assemble once more in order to undo Thanos's actions and undo the chaos to the universe, no matter what consequences may be in store, and no matter who they face...
The grave course of events set in motion by Thanos, that wiped out half the universe and fractured the Avengers ranks, compels the remaining Avengers to take one final stand in Marvel Studios' grand conclusion to twenty-two films - Avengers: Endgame.
After half of all life is snapped away by Thanos, the Avengers are left scattered and divided. Now with a way to reverse the damage, the Avengers and their allies must assemble once more and learn to put differences aside in order to work together and set things right. Along the way, the Avengers realize that sacrifices must be made as they prepare for the ultimate final showdown with Thanos, which will result in the heroes fighting the biggest battle they have ever faced."
writeLines(endgame_summaries, "endgame_summaries.txt")
```

使用 readLines() 函數「依列」載入為文字向量，長度同段落數量

In [ ]:

```
endgame_summary_char <- readLines("endgame_summaries.txt")
class(endgame_summary_char)
length(endgame_summary_char)
endgame_summary_char
```

JSON

avengers_endgame.json

```
{
    "title": "Avengers: Endgame",
    "rating": 8.7,
    "genre": ["Action", "Adventure", "Sci-Fi"]
}
```

創建 avengers_endgame.json

In [ ]:

```
avengers_endgame <- "
{
    \"title\": \"Avengers: Endgame\",
    \"rating\": 8.7,
    \"genre\": [\"Action\", \"Adventure\", \"Sci-Fi\"]
}
"
writeLines(avengers_endgame, "avengers_endgame.json")
```

載入 `avengers_endgame.json`

與 array of JSONs 不同的是載入後資料結構為 `list`

In [ ]:

```
library(jsonlite)

avengers_endgame <- fromJSON("avengers_endgame.json")
class(avengers_endgame)
length(avengers_endgame)
names(avengers_endgame)
avengers_endgame
```

輸出表格式資料

常用作表格式資料輸出的檔案格式

- 表格式文字檔 `.txt` 或 `.csv`
- Array of JSONs `.json`

表格式文字檔 `.txt` 或 `.csv`

使用 `write.table()` 函數

In [ ]:

```r
write.table(iris, file = "iris.txt", row.names = FALSE) # 不輸出資料框的列索引
write.table(iris, file = "iris.csv", sep = ",", row.names = FALSE) # 指定分隔符號為，
```

Array of JSONs `.json`

使用 `jsonlite::toJSON()` 搭配 `writeLines()` 函數

In [ ]:

```
library(jsonlite)

json_char <- toJSON(iris)
writeLines(json_char, "iris.json")
```

輸出非表格式資料

常用作非表格式資料輸出的檔案格式： `.json`

使用 `jsonlite::toJSON()` 搭配 `writeLines()` 函數

In [ ]:

```r
json_str <- toJSON(avengers_endgame)
writeLines(json_str, "avengers_endgame_out.json")
```