Python 的 50+ 練習

> *資料的載入*

數據交點 | 郭耀仁 yaojenkuo@datainpoint.com

這個章節會登場的函數、保留字與模組

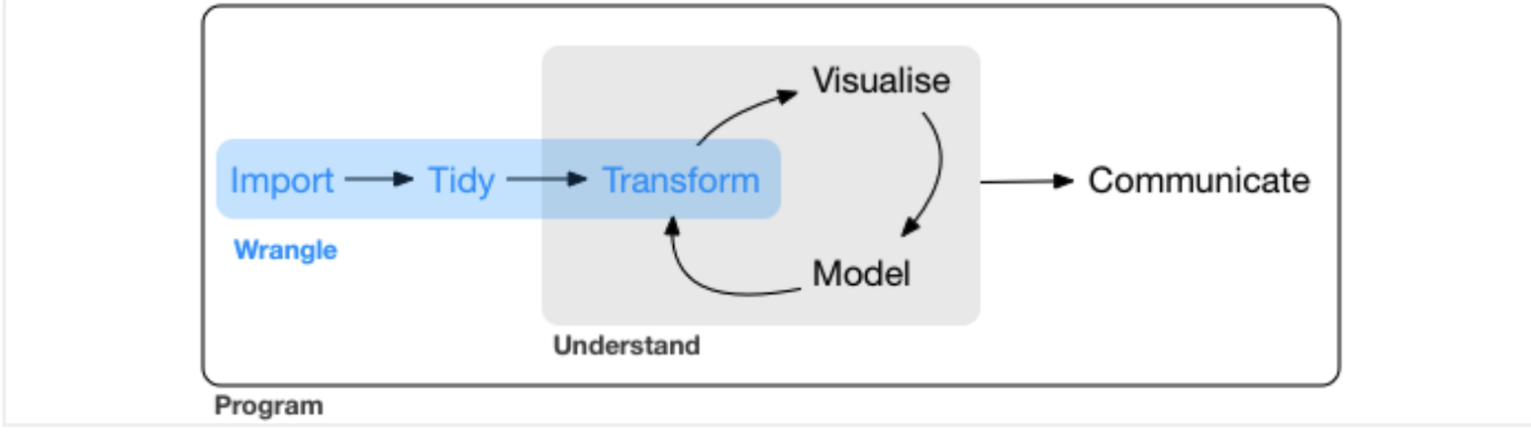
- `open()` 函數。
- `with` 保留字。
- `json` 模組。
- `sqlite3` 模組。

關於資料科學的應用場景

（複習）什麼是資料科學

資料科學是一門將資料（Data）提煉為資訊（Information）的學科，提煉過程中可能包含資料載入、資料操作、探索性分析以及監督式學習等。

（複習）現代資料科學：以程式設計做資料科學的應用



來源：R for Data Science

什麼是資料科學的應用場景

- **Import 資料的載入**。
- Tidy 資料清理。
- Transform 資料外型與類別的轉換。
- Visualise 探索性分析。
- Model 分析與預測模型。
- Communicate 溝通分享。

資料科學非技術性的工作內容

- 與使用者共同進行需求發想。
- 收斂需求並且擬定假說與敘事邏輯。
- 透過測試資料驗證假說。
- 透過溝通分享驗證敘事邏輯。

常見的來源資料格式

1. 純文字檔案。
2. 試算表。
3. 關聯式資料庫中的資料表。

純文字檔案

什麼是純文字檔案

只有文字所構成的電腦檔案，不包含字型的樣式或者段落標記，能夠使用最簡單的文字編輯器（例如 Windows 的「記事本」、macOS 的 TextEdit）直接開啟檢視。

更適合程式設計與資料科學的文字編輯器

- Visual Studio Code
- Atom
- Sublime Text
- Notepad ++(for Windows only)

不同格式的純文字檔案

- 非結構化的純文字檔案。
- JSON(JavaScript Object Notation)
- 特定符號分隔的純文字檔案。

（複習）兩種路徑的標註方式

1. 絕對路徑：從根目錄開始標註
2. 相對路徑：從工作目錄（Current working directory）開始標註

課程所使用的 JupyterLab 範例資料夾 `data` 的絕對路徑

絕對路徑 `/home/jovyan/data`

非結構化的純文字檔案

例如 `/home/jovyan/data/internet-movie-database/the_shawshank_redemption_summaries.txt`

使用內建的 `open()` 函數開啟檔案

- 以 `open()` 函數開啟 `/home/jovyan/data/internet-movie-database/the_shawshank_redemption_summaries.txt`
- `file` 物件是 `TextIOWrapper` 類別的實例
- 使用 `TextIOWrapper.readlines()` 將檔案內容一列一列載入 `list`
- 使用 `TextIOWrapper.close()` 關閉檔案。

```
file = open("PATH/TO/FILE")
list = file.readlines()
file.close()
```

In [1]:

```
text_file_path = "/home/jovyan/data/internet-movie-database/the_shawshank_redemption_summaries.txt"
file = open(text_file_path)
the_shawshank_redemption_summaries = file.readlines()
file.close()
print(type(file))
print(the_shawshank_redemption_summaries)
```

```
<class '_io.TextIOWrapper'>
['Two imprisoned men bond over a number of years, finding sol
ace and eventual redemption through acts of common decenc
y.\n', "Chronicles the experiences of a formerly successful b
anker as a prisoner in the gloomy jailhouse of Shawshank afte
r being found guilty of a crime he did not commit. The film p
ortrays the man's unique way of dealing with his new, torturo
us life; along the way he befriends a number of fellow prison
ers, most notably a wise long-term inmate named Red.\n", 'Aft
er the murder of his wife, hotshot banker Andrew Dufresne is
sent to Shawshank Prison, where the usual unpleasantness occu
rs. Over the years, he retains hope and eventually gains the
respect of his fellow inmates, especially longtime convict "R
ed" Redding, a black marketeer, and becomes influential withi
n the prison. Eventually, Andrew achieves his ends on his own
terms.\n', "Andy Dufresne is sent to Shawshank Prison for the
murder of his wife and her secret lover. He is very isolated
and lonely at first, but realizes there is something deep ins
ide your body that people can't touch or get to....'HOPE'. An
dy becomes friends with prison 'fixer' Red, and Andy epitomiz
```

es why it is crucial to have dreams. His spirit and determina
tion lead us into a world full of imagination, one filled wit
h courage and desire. Will Andy ever realize his dreams?\n",
'Bank Merchant Andy Dufresne is convicted of the murder of hi
s wife and her lover, and sentenced to life imprisonment at S
hawshank prison. Life seems to have taken a turn for the wors
e, but fortunately Andy befriends some of the other inmates,
in particular a character known only as Red. Over time Andy f
inds ways to live out life with relative ease as one can in a
prison, leaving a message for all that while the body may be
locked away in a cell, the spirit can never be truly imprison
ed.']

容易因為忘記 `file.close()` 而造成錯誤

- 透過 `with` 敘述建立資源管理器的程式區塊。
- 可以省略 `TextIOWrapper.close()`

```
with open("PATH/TO/FILE") as file:
    list = file.readlines()
```

In [2]:

```
with open(text_file_path) as file:
    the_shawshank_redemption_summaries = file.readlines()
print(the_shawshank_redemption_summaries)
```

```
['Two imprisoned men bond over a number of years, finding sol
ace and eventual redemption through acts of common decenc
y.\n', "Chronicles the experiences of a formerly successful b
anker as a prisoner in the gloomy jailhouse of Shawshank afte
r being found guilty of a crime he did not commit. The film p
ortrays the man's unique way of dealing with his new, torturo
us life; along the way he befriends a number of fellow prison
ers, most notably a wise long-term inmate named Red.\n", 'Aft
er the murder of his wife, hotshot banker Andrew Dufresne is
sent to Shawshank Prison, where the usual unpleasantness occu
rs. Over the years, he retains hope and eventually gains the
respect of his fellow inmates, especially longtime convict "R
ed" Redding, a black marketeer, and becomes influential withi
n the prison. Eventually, Andrew achieves his ends on his own
terms.\n', "Andy Dufresne is sent to Shawshank Prison for the
murder of his wife and her secret lover. He is very isolated
and lonely at first, but realizes there is something deep ins
ide your body that people can't touch or get to....'HOPE'. An
dy becomes friends with prison 'fixer' Red, and Andy epitomiz
es why it is crucial to have dreams. His spirit and determina
tion lead us into a world full of imagination, one filled wit
h courage and desire. Will Andy ever realize his dreams?\n",
```

'Bank Merchant Andy Dufresne is convicted of the murder of hi
s wife and her lover, and sentenced to life imprisonment at S
hawshank prison. Life seems to have taken a turn for the wors
e, but fortunately Andy befriends some of the other inmates,
in particular a character known only as Red. Over time Andy f
inds ways to live out life with relative ease as one can in a
prison, leaving a message for all that while the body may be
locked away in a cell, the spirit can never be truly imprison
ed.']

什麼是 JSON

> *JSON(JavaScript Object Notation) 是一種輕量的資料交換格式，對於人類而言是容易閱讀和寫作的格式，對於電腦而言是容易解析和建立的格式，簡言之，是一種對人類與電腦都友善的純文字格式，JSON 雖然源於 JavaScript，但卻是獨立於該程式語言之外，能夠被眾多程式語言輕鬆解析和建立的一種理想資料交換格式。JSON 可能由兩種資料結構組成：鍵值對應關係與有序列表，可以用 Python 的 `dict` 與 `list` 來理解。*

來源：https://www.json.org/json-en.html

使用標準模組 `json` 的 `load()` 函數載入

- 以 `open()` 函數搭配 `with` 敘述開啟 `/home/jovyan/data/nba/teams_rowbased.json`
- 使用 `json.load()` 函數將檔案內容載入。
- 因為 JSON 可能由兩種資料結構組成，輸出物件可能是 `dict` 或 `list` 類別的實例。

```
import json

with open("PATH/TO/FILE") as file:
    dict|list = json.load(file)
```

`teams_rowbased.json` 由兩種資料結構組成

`list` 中包含 34 個有 12 組鍵值對應關係的 `dict`

In [3]:

```
import json

json_file_path = "/home/jovyan/data/nba/teams_rowbased.json"
with open(json_file_path) as file:
    teams_json = json.load(file)
print(type(teams_json))
print(len(teams_json))
print(type(teams_json[0]))
print(len(teams_json[0]))
```

```
<class 'list'>
34
<class 'dict'>
12
```

假如 JSON 中的每一組鍵值長度相同表示為表格式資料

使用 `pd.read_json()` 載入為 `DataFrame`

表格式資料常見有兩種 JSON 格式

1. 基於列的 JSON(Row-based JSON)
2. 基於欄的 JSON(Column-based JSON)

基於列的 JSON(Row-based JSON)

例如 `teams_rowbased.json`

```
[{"column_1": value_1, "column_2": value_2, ...},
 {"column_1": value_1, "column_2": value_2, ...},
 ...]
```

In [4]:

```
import pandas as pd

teams_dataframe = pd.read_json(json_file_path)
teams_dataframe.head() # just show the first 5 rows
```

Out[4]:

	city	fullName	isNBAFranchise	confName	tricode	teamShortName	divName	isAllStar	nickname	urlName	teamId	altCity
0	Atlanta	Atlanta Hawks	True	East	ATL	Atlanta	Southeast	False	Hawks	hawks	1610612737	
1	Boston	Boston Celtics	True	East	BOS	Boston	Atlantic	False	Celtics	celtics	1610612738	
2	Brooklyn	Brooklyn Nets	True	East	BKN	Brooklyn	Atlantic	False	Nets	nets	1610612751	Br
3	Charlotte	Charlotte Hornets	True	East	CHA	Charlotte	Southeast	False	Hornets	hornets	1610612766	Ch
4	Chicago	Chicago Bulls	True	East	CHI	Chicago	Central	False	Bulls	bulls	1610612741	C

基於欄的 JSON(Column-based JSON)

例如 `teams_columnbased.json`

```
{"column_1": [value_1, value_2, ...],
 "column_2": [value_1, value_2, ...],
 ...}
```

In [5]:

```
teams_dataframe = pd.read_json(json_file_path)
teams_dataframe.head() # just show the first 5 rows
```

Out[5]:

	city	fullName	isNBAFranchise	confName	tricode	teamShortName	divName	isAllStar	nickname	urlName	teamId	altCity
0	Atlanta	Atlanta Hawks	True	East	ATL	Atlanta	Southeast	False	Hawks	hawks	1610612737	
1	Boston	Boston Celtics	True	East	BOS	Boston	Atlantic	False	Celtics	celtics	1610612738	
2	Brooklyn	Brooklyn Nets	True	East	BKN	Brooklyn	Atlantic	False	Nets	nets	1610612751	Br
3	Charlotte	Charlotte Hornets	True	East	CHA	Charlotte	Southeast	False	Hornets	hornets	1610612766	Ch
4	Chicago	Chicago Bulls	True	East	CHI	Chicago	Central	False	Bulls	bulls	1610612741	C

原則上 `pd.read_json()` 函數會自行判斷基於列或者基於欄

如果沒有判斷正確，可以指定參數 `orient="records"|"columns"`

來源：https://pandas.pydata.org/docs/reference/api/pandas.read_json.html

特定符號分隔的純文字檔案

- 透過特定符號分隔欄位的結構化資料。
- 常見的有逗號（`,`）、分號（`;`）、Tab 鍵（`\t`）等。
- 最廣泛使用的是逗號，因此有特定的副檔名 `.csv` 意指逗號分隔值（Comma-separated values）。

使用 `pd.read_csv()` 函數載入

In [6]:

```
movies_csv = pd.read_csv("/home/jovyan/data/internet-movie-database/movies.csv")
movies_csv.head() # just show the first 5 rows
```

Out[6]:

	id	title	release_year	rating	director	runtime
0	1	The Shawshank Redemption	1994	9.3	Frank Darabont	142
1	2	The Godfather	1972	9.2	Francis Ford Coppola	175
2	3	The Godfather: Part II	1974	9.0	Francis Ford Coppola	202
3	4	The Dark Knight	2008	9.0	Christopher Nolan	152
4	5	12 Angry Men	1957	9.0	Sidney Lumet	96

其他特定符號分隔的純文字檔依然可以使用 `pd.read_csv()` 函數

指定參數 `sep=";"|"\t"`

來源：https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html

試算表

什麼是試算表

> *試算表是類比紙本計算表格的電腦軟體，由許多列與欄所構成的儲存格，其中可以存放數值、算式或文字。常見的試算表軟體有 Microsoft Excel、macOS Numbers 與 Google Sheets*

來源：https://en.wikipedia.org/wiki/Spreadsheet

使用 `pd.read_excel()` 函數載入

In [7]:

```
excel_file_path = "/home/jovyan/data/internet-movie-database/imdb.xlsx"
movies_xlsx = pd.read_excel(excel_file_path)
movies_xlsx.head() # just show the first 5 rows
```

Out[7]:

	id	title	release_year	rating	director	runtime
0	1.0	The Shawshank Redemption	1994.0	9.3	Frank Darabont	142.0
1	2.0	The Godfather	1972.0	9.2	Francis Ford Coppola	175.0
2	3.0	The Godfather: Part II	1974.0	9.0	Francis Ford Coppola	202.0
3	4.0	The Dark Knight	2008.0	9.0	Christopher Nolan	152.0
4	5.0	12 Angry Men	1957.0	9.0	Sidney Lumet	96.0

試算表檔案的組成

- 試算表的檔案為一個活頁簿（Workbook）、活頁簿中可以有多個試算表（Spreadsheets）
- `pd.read_excel()` 函數預設讀取第零張試算表（左邊數來第一張）。
- 可以透過 `pd.ExcelFile.sheet_names` 屬性檢視活頁簿中有幾個試算表。

In [8]:

```
excel_file = pd.ExcelFile(excel_file_path)
excel_file.sheet_names
```

Out[8]:

```
['movies', 'casting', 'actors']
```

指定參數 `sheet_name` 載入特定試算表

In [9]:

```
movies_xlsx = pd.read_excel(excel_file_path)
casting_xlsx = pd.read_excel(excel_file_path, sheet_name="casting")
actors_xlsx = pd.read_excel(excel_file_path, sheet_name="actors")
print(movies_xlsx.shape)
print(casting_xlsx.shape)
print(actors_xlsx.shape)
```

```
(250, 6)
(3584, 3)
(3108, 2)
```

指定參數 `sheet_name` 亦能夠接受整數輸入，左邊數來第幾張試算表（由零開始）

In [10]:

```
movies_xlsx = pd.read_excel(excel_file_path)
casting_xlsx = pd.read_excel(excel_file_path, sheet_name=1)
actors_xlsx = pd.read_excel(excel_file_path, sheet_name=2)
print(movies_xlsx.shape)
print(casting_xlsx.shape)
print(actors_xlsx.shape)
```

```
(250, 6)
(3584, 3)
(3108, 2)
```

關聯式資料庫中的資料表

什麼是關聯式資料庫中的資料表

> *關聯式資料庫（Relational database）是基於關聯模型所建構的資料庫，是儲存在電腦中的資料集合。一個資料庫中通常會有多個資料表（Table）能夠藉助集合運算（Set operation）等數學方法來處理資料表之間的操作和關聯，也能夠用來表示現實世界中的商業邏輯，並能夠接受結構化查詢語言 SQL 來進行檢索和操作。*

來源：https://en.wikipedia.org/wiki/Relational_database

如何載入關聯式資料庫中的資料表

- 使用 `sqlite3.connect()` 函數建立與資料庫的連線。
- 運用 `str` 儲存結構化查詢語言 SQL
- 使用 `pd.read_sql()` 函數載入。
- 使用 `Connection.close()` 關閉資料庫連線。

```
import sqlite3
import pandas as pd

connection = sqlite3.connect("PATH/TO/SQLITE/DATABASE")
query = """SQL QUERY"""
pd.read_sql(query, connection)
connection.close()
```

關聯式資料庫 `/home/jovyan/data/internet-movie-database/imdb.db` 的資料表

- `movies`
- `casting`
- `actors`

建立與資料庫的連線

In [11]:

```
import sqlite3

connection = sqlite3.connect("/home/jovyan/data/internet-movie-database/imdb.db")
type(connection)
```

Out[11]:

```
sqlite3.Connection
```

載入 `movies` 資料表

In [12]:

```
query = """
SELECT *
  FROM movies;
"""
movies_db = pd.read_sql(query, connection)
print(movies_db.shape)
movies_db.head() # just show the first 5 rows
```

```
(250, 6)
```

Out[12]:

	id	title	release_year	rating	director	runtime
0	1	The Shawshank Redemption	1994	9.3	Frank Darabont	142
1	2	The Godfather	1972	9.2	Francis Ford Coppola	175
2	3	The Godfather: Part II	1974	9.0	Francis Ford Coppola	202
3	4	The Dark Knight	2008	9.0	Christopher Nolan	152
4	5	12 Angry Men	1957	9.0	Sidney Lumet	96

載入 `casting` 資料表

In [13]:

```
query = """
SELECT *
  FROM casting;
"""
casting_db = pd.read_sql(query, connection)
print(casting_db.shape)
casting_db.head() # just show the first 5 rows
```

```
(3584, 3)
```

Out[13]:

	movie_id	actor_id	ord
0	1	2853	1
1	1	2097	2
2	1	333	3
3	1	3029	4
4	1	533	5

載入 `actors` 資料表

In [14]:

```
query = """
SELECT *
  FROM actors;
"""
actors_db = pd.read_sql(query, connection)
print(actors_db.shape)
actors_db.head() # just show the first 5 rows
```

```
(3108, 2)
```

Out[14]:

	id	name
0	1	Aamir Khan
1	2	Aaron Eckhart
2	3	Abbas-Ali Roomandi
3	4	Abbey Lee
4	5	Abbie Cornish

關閉資料庫連線

In [15]:

```
connection.close()
```

重點統整

- 資料科學非技術性的工作內容：
    - 與使用者共同進行需求發想。
    - 收斂需求並且擬定假說與敘事邏輯。
    - 透過測試資料驗證假說。
    - 透過溝通分享驗證敘事邏輯。

重點統整（續）

- 常見的來源資料格式：
    - 純文字檔案。
    - 試算表。
    - 關聯式資料庫中的資料表。

重點統整（續）

- 使用 `TextIOWrapper.readlines()` 載入非結構化的純文字檔案。
- 使用 `json.load()` 函數載入 JSON。
- 使用 `pd.read_csv()` 函數載入特定符號分隔的結構化純文字檔案。
- 使用 `pd.read_json()` 函數載入每組鍵值長度相同的 JSON。
- 使用 `pd.read_excel()` 函數載入試算表。
- 使用 `pd.read_sql()` 函數載入關聯式資料庫的資料表。