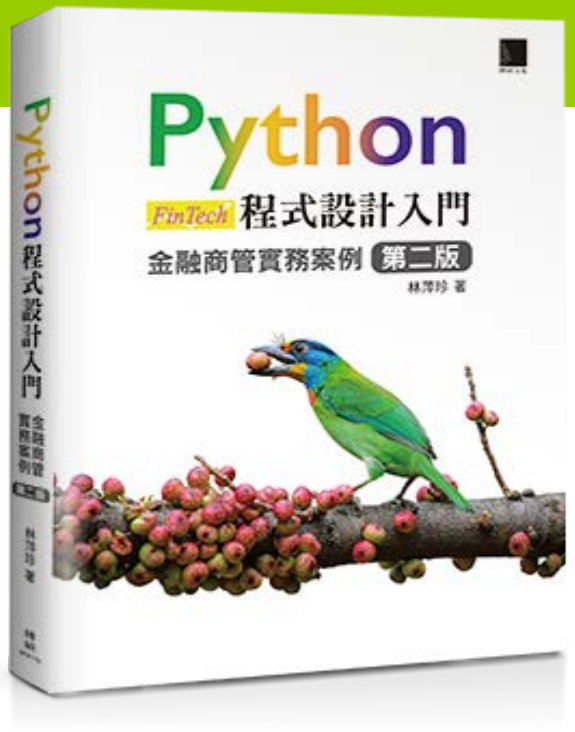
【南區Fintech研習營】Python 程式設計基礎: Google finance股價爬蟲應用

講者：林萍珍



第七章 檔案處理

7-1　檔案串流物件

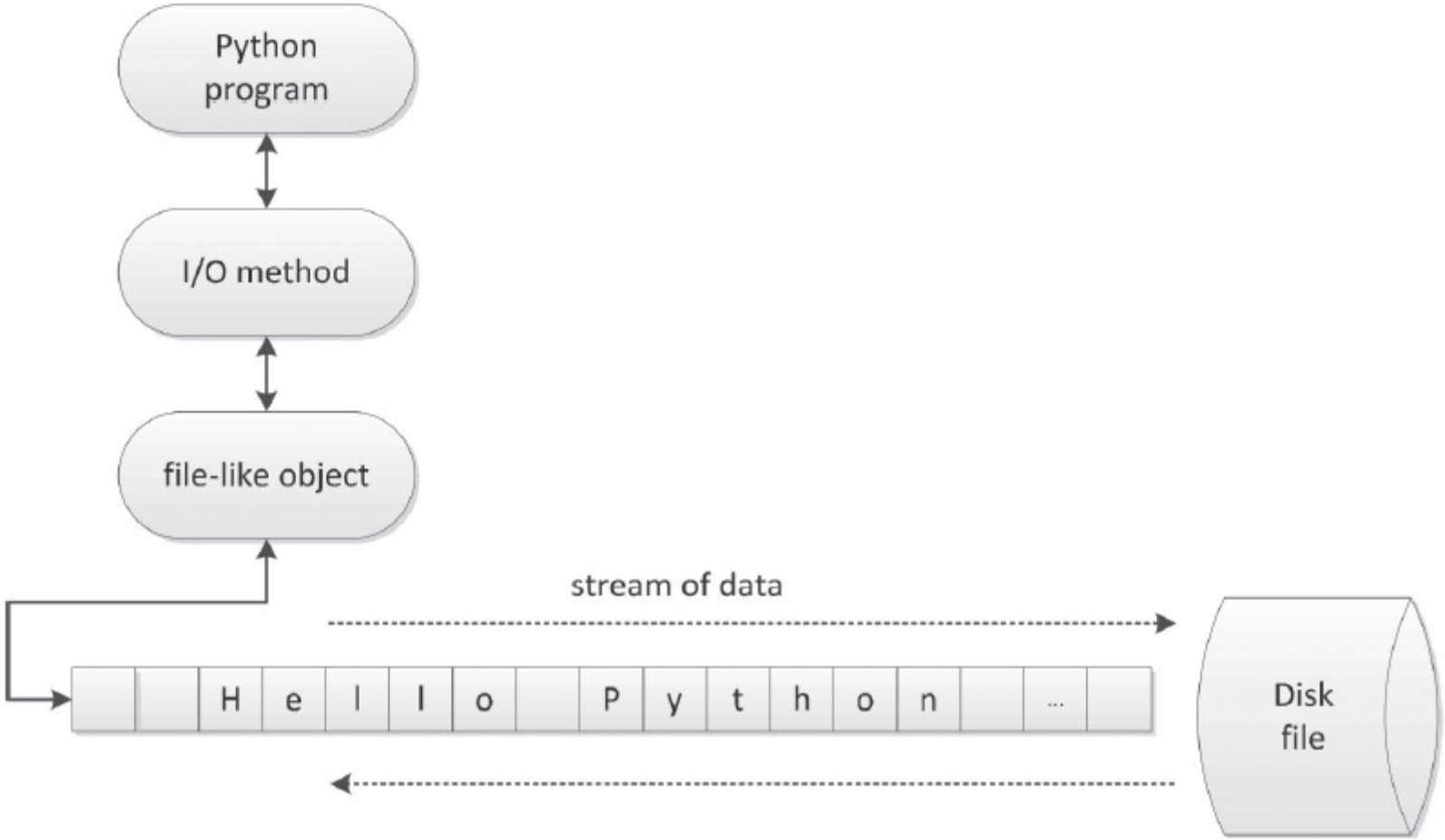
7-2　檔案存取方法

7-3　檔案路徑處理

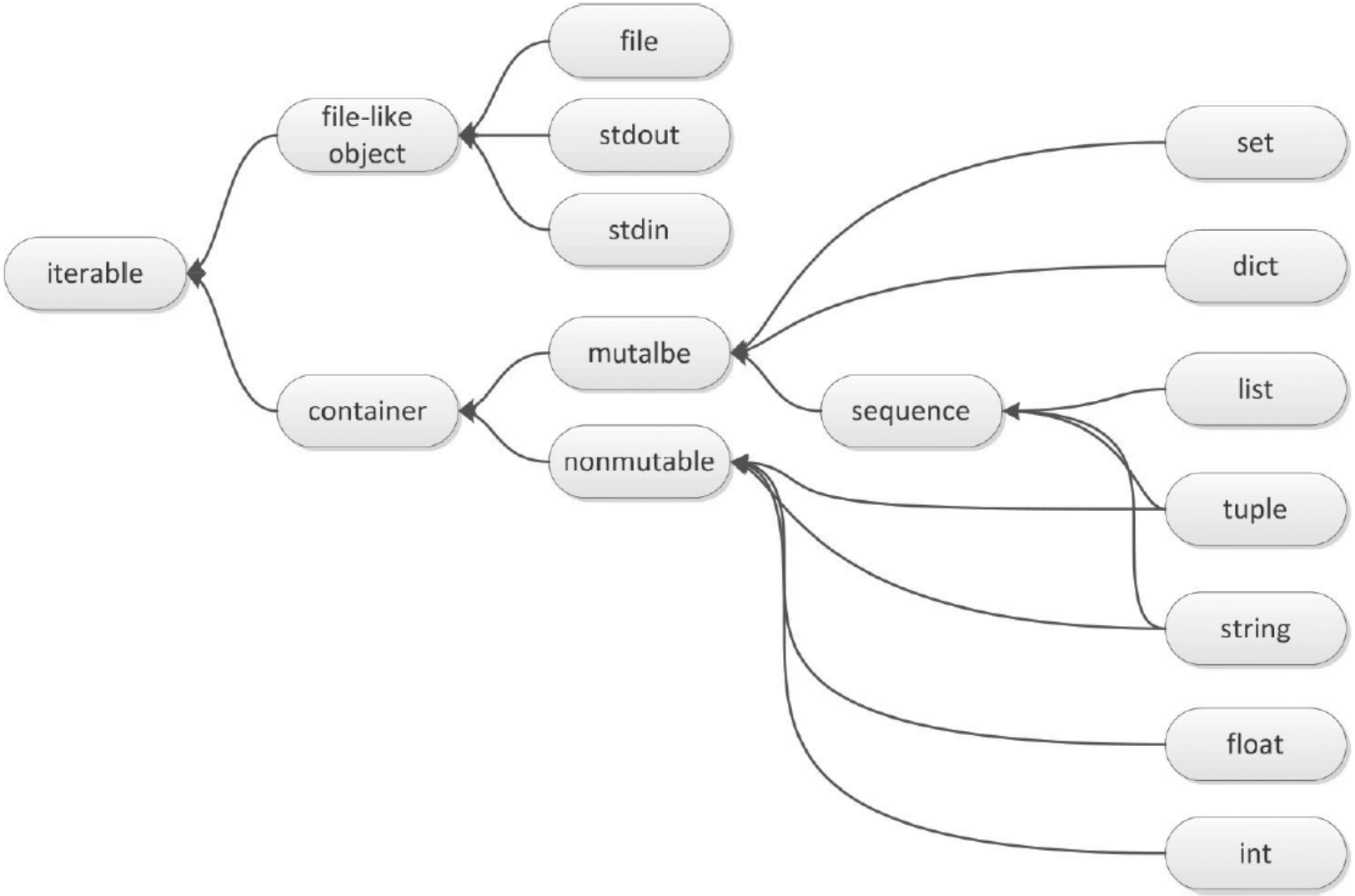
7-4　網路取得資料

7-5　實務案例

檔案與串流的運作



類檔案物件的抽象物件架構



檔案處理步驟

步驟 1 開啟檔案

利用 open() 函數開啟指定路徑與檔名的檔案，其回傳值是檔案物件（file object），不同檔案獨立指派給不同的檔案物件，後續可依指定的檔案物件進行檔案讀寫動作。

步驟 2 讀寫檔案等處理程序

利用特定的檔案物件之 read(), readlines(), write() 等方法讀寫檔案，處理檔案的標準輸入與輸出（file I/O）。

步驟 3 關閉檔案

當一個檔案物件被重新指派給其他檔案時，Python 會自動關閉原檔案，但工作結束後使用 close() 函數關閉檔案才是良好的做法。利用 close() 函數關閉某個檔案物件，即代表關閉某特定的路徑與檔名的檔案，並釋放記憶體資源。

開啟txt文字檔

open 函數可以用以開啟檔案讀取資料，其回傳值是代表該檔案物件，使用者一定要輸入第 1 個參數為檔案名稱，第 2 個參數為讀寫模式，第 2 個參數以後可以省略依預設值處理，其使用語法如下：

```
fin=open(filename, accessmode)
```

1. filename：指檔案名稱包含指定路徑，此參數是一個字串資料型別。若要讀取的檔案與程式檔 .py 是在同一路徑下，則可以不用指定路徑，只寫檔名即可，否則要寫相對路徑或絕對路徑。
2. accessmode：指讀取模式，決定檔案開啟時的模式如 read(r), write(w), append(a) 等，完整的模式整理如下表 7-1。

讀寫模式

表 7-1　讀寫模式

模式	功能說明
r	開啟檔案指定讀取模式（預設）。
rb	開啟檔案指定讀取二進位檔案模式。
r+	開啟檔案讀取與寫入模式並存，檔案須已存在。
rb+	開啟檔案讀取與寫入二進位檔案模式。
w	開啟檔案指定寫入模式，若檔案已存在則覆寫，若檔案不存在則開新檔案做為後續的寫入動作。
wb	開啟檔案指定寫入二進位檔案模式，若檔案已存在則覆寫，若檔案不存在則開新檔案做為後續的寫入動作。
w+	開啟檔案指定同時讀取與寫入模式，若檔案已存在則覆寫，若檔案不存在則開新檔案做為後續的寫入動作。
wb+	開啟檔案指定同時讀取與寫入二進位檔案模式，若檔案已存在則覆寫，若檔案不存在則開新檔案做為後續的寫入動作。

使用 with as

- Python 的物件若支援文章脈絡管理協定（context management protocol），則此物件即可稱為文章脈絡管理者（context manager）
- 支援文章脈絡管理協定的物件，必須自我管理，自行建構以 with as 為關鍵字的進入與離開的方法或時點，這類物件會適時執行自我管理或清理的動作。
- with as 語法如下

```
with 運算式 as 物件名稱 :
指令 1
… 指令 n
```

文字檔的讀寫方法

表 7-2　文字檔案讀寫方法

方法	功能說明
read()	一次讀取所有的檔案內容。
readline()	逐列讀取。
readlines()	傳回一 list，使用 list 收集讀取的每一列資料，可配合 for 迴圈每次取出串列的 1 個字串元素進行處理。
write()	資料寫入檔案。

範例 7-7 文字檔寫入 write() 方法

使用 write() 方法，可以同時做讀取、寫入檔案的動作，資料來源同【範例 7-4】，去掉表頭後寫入檔案。

示範程式碼

```
1  #E_7_7.py 功能 : readlines() 配合 for 每次取出串列並忽略表頭
2  with open('./file/oil.txt', 'r') as fin:
3      with open ('oil_write.txt', 'w') as fout:
4          for line in fin.readlines():
5          if line.startswith(' 日 '): # 判斷表頭的第 1 個字
6              continue
7          print(line, end='')
8          fout.write(line)
```

讀取csv檔案常用的方法

表 7-4　CSV 套件常用的方法

方法	功能說明
csv.reader()	讀取檔案物件指向所有列的資料，並回傳給可迭代的閱讀器物件 (reader object)。
csv.writer ()	傳回一個寫入器物件 (writer object)。
next()	取出檔案物件內一列元素。
writerow()	依指定格式 (分隔字完) 轉成字串。

使用with讀取csv檔

示範程式碼

```
1   #E_7_9.py 功能 : 使用 CSV 套件讀寫檔案
2   import csv
3   with open('./file/SP3008_201511.csv', 'r') as fin:
4       with open('./file/SP3008_csv_out.csv', 'w') as fout:
5           csvreader = csv.reader(fin, delimiter=',')
6           csvwriter = csv.writer(fout, delimiter=',')
7           header = next(csvreader)
8           print(header)
9           csvwriter.writerow(header)
10           for row in csvreader:
11              row[0] = row[0].replace('/', '-')
12              print(','.join(row))
13              csvwriter.writerow(row)
```

pandas 的 xlsxwriter

- pandas 是 Python 提供資料分析的套件。它可以讀取、篩選、重組與分析大小樣本 資料並可以依指定資料格式或檔案型態輸出。
- 若是要進一 步做複雜的資料分析或上網抓股價資料，則推薦用 pandas 套件。
- pandas 有一個資料框 架（DataFrame）功能可以協助資料分析的有用功能語法 如下：
  - pandas.DataFrame(data=None, index=None, columns=None)
    - data：可以是字典、numpy 的陣列（ndarray）等資料型別或 dataframe 本身。
    - index：列的索引值。
    - columns：欄的索引值。

xlsxwriter 常用方法

表 7-7　xlsxwriter 套件常用的方法與屬性

方法	功能說明
read_excel()	開啟一個要讀取的 excel 檔案試算表。
describe()	回傳一組敘述統計指標的數據。
ExcelWriter()	開啟一個要寫入的 excel 檔案試算表。
to_excel()	寫入到指定工作表。
workbook.add_chart()	新增一個統計圖物件。
add_series()	新增一組數列到統計圖物件中。
set_x_axis()	設定 x 軸標題。
set_y_axis()	設定 y 軸標題。
set_legend()	設定圖例。
insert_chart()	插入一個統計圖物作。
conditional_format()	統計圖的其他屬性設定包含顏色。
save()	儲存試算表檔案。

上網爬股價資料畫K線圖

實務案例 7-3 上網路抓蘋果（Apple）股價資料畫 K 線圖

上美國 yahoo 網站抓蘋果（代號 AAPL）股價，即時畫成 K 線圖。

示範程式碼

```
1   #RC_7_3: 上網抓資料並畫 K 線圖
2   import matplotlib.pyplot as plt
3   from matplotlib.dates import DateFormatter
4   from matplotlib.finance import quotes_historical_yahoo_ohlc, candlestick_ohlc
5   # 下載資料起迄日 , 日期格式與股票代號
6   start = (2016, 4, 1)
7   end = (2016, 4, 25)
8   weekFormatter = DateFormatter('%b %d') # 例如 , Jan 03 2016
9   quotes = quotes_historical_yahoo_ohlc('AAPL', start, end)
10  # 若抓取的資料是空字串則離開系統
11  if len(quotes) == 0:
12      raise SystemExit
13  # 設定繪圖區域的格式化
14  fig, ax = plt.subplots()
15  ax.xaxis_date()
16  plt.setp(plt.gca().get_xticklabels(), rotation=45, horizontalalignment='right')
```

執行結果

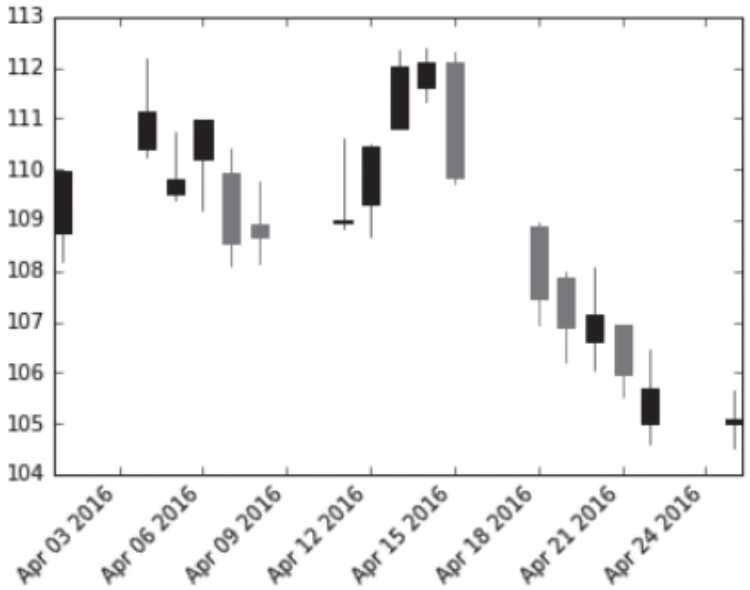


圖 7-9　上網路抓蘋果（Apple）股價資料畫 K 線圖

結果說明

執行結果會顯示在 Ipython console 視窗（見圖 7-9），可以點選此圖按右鍵（Save images as）即可以將顯示的圖形儲存成 *.png。

處理微巨量資料效能

實務案例 7-2 處理微巨量資料效能

使用 pandas 套件開啟文字檔 TXF2003.txt 容量約 59M 共約 82 萬筆。計算敘述統計並計算執行時間。檔案內容為過去幾年大盤指數每分鐘資料。

示範程式碼

```
1  #RC_7_2: 大盤指數每分鐘資料 82 萬筆
2  import pandas as pd
3  import time
4  starttime = time.clock()
5  df = pd.read_csv('./file/TXF2003.txt', sep=",")
6  df.columns = ['Date', 'Time', 'Open', 'High', 'Low','Close', 'Vol']
7  print(df.describe())
8  endtime = time.clock()
9  print(' 程式執行時間 = %d %s' %(round(endtime - starttime), ' 秒 '))
```

本章講解完畢

現場同學們如有不懂的地方，
請提出問題。