FILE STREAMING AND PARSING

Chih-Chung Hsu (許志仲)
Institute of Data Science
National Cheng Kung University
https://cchsu.info

資料夾與檔案

- 資料可以儲存在檔案中，檔案可以放置於資料夾下
  - Python 利用資料夾與檔案相關模組，可以建立檔案、儲存資料到檔案、查詢檔案大小、從檔案中刪除資料、刪除檔案、建立資料夾、刪除資料夾、搜尋資料夾下特定副檔名的檔案等功能

- 檔案相關模組
  - 模組os 內有許多資料夾與檔案的函式可以使用，可以對資料夾與檔案進行存取與管理，重要的資料夾與檔案函式如下。

- 相關 package
  - os, glob, open

OS package

模組 os 重要函式與說明	程式碼與執行結果
os.getcwd() 回傳目前所在的資料夾。	import os print(os.getcwd())
	c:\
os.chdir(path) 改變到 path 所指定的資料夾下。	import os os.chdir('c:\\') print(os.getcwd())
	c:\
os.mkdir(path, dir_fd=None) 建立 path 所指定的資料夾，path 所指定的資料夾不能事先存在。若 dir_fd 有設定資料夾路徑，則 path 爲相對路徑，path 前須加上 dir_fd，建立的資料夾爲 dir_fd\path；若 dir_fd 爲 None 則 path 爲絕對路徑。	import os os.mkdir('c:\\test')
	建立 c 磁碟下的 test 資料夾。

資料夾與檔案相關模組

os.rmdir(path, *, dir_fd=None) 刪除 path 所指定的資料夾，但該資料夾需要是空的才能刪除。若 dir_fd 有設定資料夾路徑，則 path 爲相對路徑，path 前須加上 dir_fd，刪除的檔案爲 dir_fd\path；若 dir_fd 爲 None 則 path 爲絕對路徑。	import os os.rmdir('c:\\test') 刪除 c 磁碟下的 test 資料夾。
os.remove(path, dir_fd=None) 刪除 path 所指定的檔案，若 dir_fd 有設定資料夾路徑，則 path 爲相對路徑，path 前須加上 dir_fd，刪除的檔案爲 dir_fd\path；若 dir_fd 爲 None 則 path 爲絕對路徑。	import os os.remove('c:\\test.chm') 刪除檔案 c 磁碟下的檔案 test.chm。

資料夾與檔案相關模組

模組 os 重要函式與說明	程式碼與執行結果
os.listdir(path='.') 以串列回傳 path 所指定資料夾下的檔案與資料夾。	import os print(os.listdir('f:\\')) ['teach', 'software', 'python']
os.path.abspath(path) 將 path 所指定的相對路徑轉換成絕對路徑。	import os print(os.path.abspath('.')) d:\python
os.path.join(path, *paths) 回傳 path 結合 *paths 的資料夾或檔案路徑。	import os root = 'c:\\test' file = 'zsg.jpg' print(os.path.join(root, file)) c:\test\zsg.jpg

資料夾與檔案相關模組

os.walk(top, topdown=True, onerror=None, followlinks=False)
遞迴方式走訪 top 所指定的資料夾，預設使用 topdown 方式，由上而下，回傳一個 tuple 包含三個元素 (dirpath, dirnames, filenames)，dirpath 是字串，表示資料夾的絕對路徑，dirnames 是串列，表示 dirpath 下的所有子資料夾，filenames 是串列，表示 dirpath 下的所有非資料夾的元素。
	import os
	path = "c:\\test"
	for root, dirs, files in os.walk(path):
	  for file in files:
	  print(os.path.join(root, file))
	
	c:\test\zsg.jpg

os.path.isfile(path) 檢查 path 所指定的檔案或資料夾是否爲檔案，若是檔案則回傳 True，否則回傳 False。	import os print(os.path.isfile("c:\\Windows")) False
os.path.isdir(path) 檢查 path 所指定的檔案或資料夾是否爲資料夾，若是資料夾則回傳 True，否則回傳 False。	import os print(os.path.isdir("c:\\Windows")) True

資料夾與檔案相關模組

os.path.exists(path) 檢查 path 所指定的檔案或資料夾是否存在，若是則回傳 True，否則回傳 False。	print(os.path.exists("c:\\Windows")) True
os.path.getsize(path) 回傳 path 所指定檔案的大小，單位爲 byte。	print(os.path.getsize("f:\\zsg.jpg")) 1950

資料夾與檔案相關模組

其他檔案與資料夾相關模組的重要函式，如下。

其他相關模組與說明	程式碼與執行結果
glob.glob(pathname) 找出 pathname 所指定類型的檔案。	import os,glob path = "f:\\teach\\python" os.chdir(path) print(glob.glob('*.py')) 註：「*」表示任何檔案名稱。
	['test.py', 'hello.py']
fnmatch.fnmatch(filename, pattern) 檢查 filename 是否滿足 pattern，若滿足 pattern 則回傳 True，否則回傳 False。	import os,glob,fnmatch path = "f:\\teach\\python" os.chdir(path) files = glob.glob('*.py') for file in files: print(fnmatch.fnmatch(file,'*.py')) 註：「*」表示任何檔案名稱。
	True True

資料夾與檔案相關模組

fnmatch.filter(names, pattern) 可以使用 pattern 過濾 names 所匯入的檔案名稱串列，只找出 pattern 所指定的副檔名。	import os,glob,fnmatch path = "f:\\teach\\python" os.chdir(path) files = glob.glob('*.py') print(fnmatch.filter(files, '*.py')) 註：「*」表示任何檔案名稱。
	['test.py', 'hello.py']

str.endswith(suffix)
字串是否以 suffix 結尾。
	path = "f:\\teach\\python"
	os.chdir(path)
	files = glob.glob('*.py')
	for file in files:
	  if file.endswith('*.py'):
	  print(file)
	註：「*」表示任何檔案名稱。

	test.py hello.py

使用os.walk 列出所有Python 檔案

- Sample code

```
1  import os
2  path = "f:\\python"
3  for root, dirs, files in os.walk(path):
4      for file in files:
5          if file.endswith(".py"):
6              print(os.path.join(root, file))
```

使用os.walk 列出所有JPG 與PNG 檔案

```
1   import fnmatch, os
2   path = "f:\\python"
3   exts = ['*.jpg', '*.jpeg', '*.png']
4   matches = []
5   for root, dirs, files in os.walk(path):
6       for ext in exts:
7           for file in fnmatch.filter(files, ext):
8               matches.append(os.path.join(root, file))
9   for image in matches:
10      print(image)
```

READING CONTENTS FROM FILES

存取文字檔

- 這一節要取出檔案內的每一行，首先要學會開啟檔案、讀取檔案、寫入檔案與關閉檔案，處理的檔案類型可以是文字檔、csv 檔或二進位檔

- 基本package不須額外 import

存取文字檔

<table>
<tr><th>檔案存取函式與說明</th><th>程式碼與執行結果</th></tr>
<tr><td>open(filename, mode, encoding)
開啓檔案 filename，設定模式爲 mode，使用 encoding 爲文字編碼。
mode 所提供的模式，如下表。
<table><tr><th>mode</th><th>說明</th></tr><tr><td>r</td><td>讀取</td></tr><tr><td>w</td><td>寫入，先刪除原先檔案的內容。</td></tr><tr><td>x</td><td>寫入，但檔案不能已經存在，防止複寫檔案。</td></tr><tr><td>a</td><td>寫入，若檔案已經存在，保留原本的資料，將新增的內容加到檔案的最後。</td></tr><tr><td>t</td><td>文字檔</td></tr><tr><td>b</td><td>二進位檔</td></tr><tr><td>+</td><td>可以讀取與寫入檔案。</td></tr></table></td><td>fin = open('poem.txt','rt',encoding='utf-8')

開啓檔案 poem.txt，開啓檔案模式設定爲文字模式，且設定爲讀取模式，文字檔爲 utf-8 編碼。</td></tr>
</table>

存取文字檔

read(size) 若不指定 size，則會讀取整個檔案，否則會讀取大小爲 size 位元組的資料。	fin = open('poem.txt','rt',encoding='utf-8') s = fin.read() print(s) fin.close() 昔人已乘黃鶴去，此地空余黃鶴樓。
readline() 從檔案中讀取一行。	fin = open('poem.txt','rt',encoding='utf-8') s = fin.readline() print(s) fin.close() 昔人已乘黃鶴去，此地空余黃鶴樓。

存取文字檔

檔案存取函式與說明	程式碼與執行結果
readlines() 從檔案中讀取每一行資料，最後將每一行資料製作成串列。	fin = open('poem.txt','rt',encoding='utf-8') lines = fin.readlines() for line in lines: print(line) fin.close()
	昔人已乘黃鶴去，此地空余黃鶴樓。
write(string) 將 string 寫入檔案。	s = 'Python' fout = open('my.txt','wt') fout.write(s) fout.close()
	開啓 my.txt 發現內容爲「Python」。
print(*objects, sep='', end='\n', file=sys.stdout) 函式 print 也可以寫入檔案，使用 file 指定要寫入的檔案，就可以使用函式 print 將 *objects 寫入檔案，而 sep 用於設定每個資料間的間隔字元，end 用於設定行與行之間的換行字元。	s = 'Python' fout = open('my.txt','wt') print(s, file=fout) fout.close()
	開啓 my.txt 發現內容爲「Python」。

純文字檔讀取

- 解題流程：
  - 使用函式open 開啟檔案，接著利用函式read 一次讀取整個檔案
  - 最後使用函式close 關閉檔案，需事先使用文字編輯器新增文字到文字檔

- Sample code: colab預設資料

```
1   fp  =  open('sample_data/california_housing_test.csv',  'r')
2   print(fp.readlines())
3   fp.close()
```

```
['"longitude","latitude","housing_median_age","total_rooms","total_bedrooms",
```

10-2-2　使用for 迴圈讀取純文字檔

- 使用函式open 開啟檔案，接著利用for 迴圈一行接著一行讀取檔案內容
- 最後使用函式close 關閉檔案，需事先使用文字編輯器新增文字到文字檔，檔名命名為「poem.txt」，若使用其他檔名，需修改函式open 內所指定的檔案名稱。

搭配 For-loop 讀取

```
1   fp  =  open('sample_data/california_housing_test.csv',  'r')
2   data  =  fp.readlines()
3   fp.close()
4
5   for  line  in  data:
6       print(line)
```

串流輸出內容已截斷至最後 5000 行。

```
-118.410000,34.030000,36.000000,3053.000000,635.000000,1234.000000,577.000000,5.163700,500001.000000

-121.450000,38.610000,32.000000,2436.000000,612.000000,1509.000000,618.000000,1.042400,81400.000000

-117.250000,32.830000,17.000000,2075.000000,262.000000,704.000000,241.000000,10.952900,500001.000000

-119.800000,36.820000,24.000000,5377.000000,1005.000000,2010.000000,982.000000,3.454200,121200.000000

-121.310000,38.010000,22.000000,2101.000000,514.000000,1304.000000,511.000000,2.834800,101600.000000

-118.180000,34.050000,41.000000,762.000000,147.000000,817.000000,176.000000,3.750000,123100.000000

-122.130000,37.370000,30.000000,2139.000000,260.000000,742.000000,242.000000,11.806000,500001.000000
```

讀取指定資料夾下所有檔案內容的程式

- 使用「glob.glob」找出指定資料夾下的所有Python 檔，使用「with open as」開啟檔案
- 接著利用for 迴圈一行接著一行讀取檔案內容，將每一行程式顯示在螢幕上。
- 使用「with open as」開啟檔案會自動關閉檔案，不須再加上函式close 關閉檔案。

```
1  import glob
2  python_files = glob.glob('f:\\teach\\python\\*.py')
3  for file_name in python_files:
4      print(' 檔案爲 ' + file_name)
5      with open(file_name) as f:
6          for line in f:
7              print(line.rstrip())
8      print()
```

將字串寫入檔案

- 使用「with open as」開啟檔案進行寫入，接著利用函式print 與函式write 將字串寫入檔案。

```
1  s=' 昔人已乘黃鶴去，此地空余黃鶴樓。\
2  黃鶴一去不復返，白雲千載空悠悠。'
3  with open('poem.txt','wt',encoding='utf-8') as fout:
4      print(s,file=fout)
5      fout.write(s)
```

需要開啓檔案 poem.txt，看看檔案內是否有字串 s 的內容。

存取csv 檔

存取 csv 檔的函式與說明	程式碼與執行結果
csv.writer(csvfile) 將 csvfile 所指定的檔案轉換成 csv.writer 物件。	import csv with open('99.csv', 'wt', newline='') as fout: writer = csv.writer(fout)
	開啓檔案 99.csv，允許寫入資料，並轉換成 csv.writer 物件。
csv.reader(csvfile) 將 csvfile 所指定的檔案轉換成 csv.reader 物件。	import csv with open('99.csv', 'rt') as fin: reader = csv.reader(fin)
	開啓檔案 99.csv，允許讀取資料，並轉換成 csv.reader 物件。

writerows(rows) 將 rows 寫入 csv 檔。	import csv with open('test.csv', 'wt', newline='') as fout: writer = csv.writer(fout) writer.writerows([(1,2,3)])
	開啓 test.csv，查看內容是否爲「1,2,3」
csv.DictReader(csvfile) 將 csvfile 所指定的檔案轉換成 csv.DictReader 物件。	import csv with open('test.csv', 'wt', newline='') as fout: writer = csv.writer(fout) writer.writerows([(1,2,3)]) with open('test.csv', 'rt') as fin: reader = csv.DictReader(fin, fieldnames=['a','b','c']) rows = [row for row in reader] print(rows)
	[{'b': '2', 'a': '1', 'c': '3'}]

<table>
<tr><th>存取 csv 檔的函式與說明</th><th>程式碼與執行結果</th></tr>
<tr><td>csv.DictWriter(csvfile)
將 csvfile 所指定的檔案轉換成 csv.DictWriter 物件。</td><td rowspan="2">import csv
with open('test.csv', 'wt', newline='') as fout:
writer = csv.writer(fout)
writer.writerows([(1,2,3)])
with open('test.csv', 'rt') as fin:
reader = csv.DictReader(fin, fieldnames=['a','b','c'])
rows = [row for row in reader]
print(rows)
fout = open('test2.csv', 'wt',newline='')
writer = csv.DictWriter(fout, fieldnames=['a','b','c'])
writer.writeheader()
writer.writerows(rows)
fout.close()</td></tr>
<tr><td>writeheader()
寫入 csv 檔的標題列。</td></tr>
<tr><td colspan="2">將檔案 test.csv 拷貝到檔案 test2.csv，加上標題「a,b,c」，開啓 test2.csv 會發現檔案內容爲「a,b,c
1,2,3」</td></tr>
</table>

使用模組csv 對csv 檔進行寫入與讀取

```
1  import csv
2  with open('99.csv', 'wt', newline='') as fout:
3      writer = csv.writer(fout)
4      for i in range(1,10):
5          for j in range(1,10):
6              writer.writerows([(str(i),str(j),str(i*j))])
7  with open('99.csv', 'rt') as fin:
8      reader = csv.reader(fin)
9      rows = [row for row in reader]
10     print(rows)
```

```
[['1', '1', '1'], ['1', '2', '2'], ['1',
'3', '3'], ['1', '4', '4'], ['1', '5',
'5'], ['1', '6', '6'], ['1', '7', '7'],
['1', '8', '8'], ['1', '9', '9'], ['2',
'1', '2'], ['2', '2', '4'], ['2', '3',
'6'],… , ['9', '1', '9'], ['9', '2',
'18'], ['9', '3', '27'], ['9', '4',
'36'], ['9', '5', '45'], ['9', '6',
'54'], ['9', '7', '63'], ['9', '8',
'72'], ['9', '9', '81']]
```

使用模組csv 寫入與讀取csv 檔並加上標題

```
1   import csv
2   with open('99.csv', 'wt', newline='') as fout:
3       writer = csv.writer(fout)
4       for i in range(1,10):
5           for j in range(1,10):
6               writer.writerows([(str(i),str(j),str(i*j))])
7   with open('99.csv', 'rt') as fin:
8       reader = csv.DictReader(fin, fieldnames=[' 被乘數 ',' 乘數 ',' 積 '])
9       rows = [row for row in reader]
10      print(rows)
11      fout = open('99b.csv', 'wt',newline='',encoding='utf-8')
12      writer = csv.DictWriter(fout, fieldnames=[' 被乘數 ',' 乘數 ',' 積 '])
13      writer.writeheader()
14      writer.writerows(rows)
15      fout.close()
```