shp2csv 程式設計練習

目的: 使用 **python** 以及相關 **libary** 練習撰寫地理工具箱**(Geoprocessing toolbox)**
開發環境:

- Python 3.7.0 (Anaconda)
- pyshp (latest) (https://github.com/GeospatialPython/pyshp)
- csv (built-in) (https://docs.python.org/3/library/csv.html#module-csv) or pandas (https://pandas.pydata.org/pandas-docs/stable/)

作者: 正瑋 **2019/06**

In [ ]:
```
# 引入Libraries
## 內建
import csv # 檔案輸出使用
```

In [1]:
```
## 第三方
import shapefile as shp # 讀取shapefile使用
import pandas as pd # 資料處理與檔案輸出使用
```

In [ ]:
```
import numpy as np #練習的檔案使用
```

如果出現以下錯誤(**ModuleNotFoundError**)，代表系統中沒有安裝對應的模組:

```
---------------------------------------------------------------------------
-
ModuleNotFoundError
Traceback (most recent call last)
<ipython-input-1-c05e705b3105> in <module>()
      1 ## 第三方
----> 2 import shapefile as shp

ModuleNotFoundError: No module named 'shapefile'
```

確認 **Library** 版本

在開發工具時了解自己的程式版本是很重要的，因為不同的版本函數的表現可能會不一樣（會因為過時而刪除或是新版本有新的函數），在檔案開頭就要確認自己的工具的版本並記載清楚。

有兩個方法：

1. 使用 `conda list` 看所有以安裝的程式版本
2. 使用各Library內的屬性

In [2]:
```
# pandas 模組
pd.show_versions()
```

```
INSTALLED VERSIONS
------------------
commit: None
python: 3.7.0.final.0
python-bits: 64
OS: Windows
OS-release: 10
machine: AMD64
processor: Intel64 Family 6 Model 158 Stepping 10, GenuineIntel
byteorder: little
LC_ALL: None
LANG: None
LOCALE: None.None

pandas: 0.23.4
pytest: 3.8.0
pip: 10.0.1
setuptools: 40.2.0
Cython: 0.28.5
numpy: 1.15.1
scipy: 1.1.0
pyarrow: None
xarray: None
IPython: 6.5.0
sphinx: 1.7.9
patsy: 0.5.0
dateutil: 2.7.3
pytz: 2018.5
blosc: None
bottleneck: 1.2.1
tables: 3.4.4
numexpr: 2.6.8
feather: None
matplotlib: 2.2.3
openpyxl: 2.5.6
xlrd: 1.1.0
xlwt: 1.3.0
xlsxwriter: 1.1.0
lxml: 4.2.5
bs4: 4.6.3
html5lib: 1.0.1
sqlalchemy: 1.2.11
pymysql: None
psycopg2: None
jinja2: 2.10
s3fs: None
fastparquet: None
pandas_gbq: None
pandas_datareader: None
```

In [3]:
```
# pyshp
shp.__version__
```

Out[3]: '2.1.0'

讀檔模組

目標：開啟 shapefile 檔，並取出其中幾何或是屬性資料
使用： GDAL 或 pyshp
備註： GDAL可以支援更廣泛的格式，未來應該會使用 GDAL 來讀取 FileGDB 中的 Feature Class

輸入：

- 檔案路徑（String）: shapefile 的檔案路徑。
  輸出：
- 資料表（pandas.DataFrame）：屬性資料。

流程圖：

- 開啟 shapefile
- 把座標或是屬性讀取並存成 **pandas.DataFrame (http://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html)** 物件

以下請自行練習：

In [ ]:
```
# 範例檔的 SHP
CSV_path = r".\InstitutesInAS.shp"
# 請練習寫出讀檔模組，把裡面的表格轉成 pd.DataFrame 物件
```

In [ ]:
```

```

資料處理模組

目的： 將讀取出來的資料進行處理或運算
使用： pandas

輸入：
輸入：

- 資料表（pandas.DataFrame）：屬性資料。
  輸出：
- 資料表（pandas.DataFrame）：處理完畢的資料。

流程圖（本練習）：

- 讀取範例檔（CSV） 建立成 pd.DataFrame 物件
- 以 Map_ID 排序

參考資料：
pandas.DataFrame.sort_values (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sort_values.html)

In [4]:
```
# 讀入 CSV檔的路徑
CSV_Path = r".\InstitutesInAS.csv"
```

In [5]:
```
# 將 CSV 轉為 DataFrame 物件
IIS_df = pd.read_csv(CSV_Path)
```

In [6]:
```
# 來看看讀進來的表格是什麼樣子
IIS_df
```

Out[6]:

	id	Name	Map_Index	X	Y
0	1	國際研究生教學研究大樓/幼稚園	60.0	311858.315374	2.771187e+06
1	2	跨領域大樓	8.0	311944.899157	2.770908e+06
2	3	細胞個體生物研究所	4.0	311937.719038	2.770836e+06
3	4	國家動物中心	7.0	311865.917852	2.770839e+06
4	5	生物醫學科學研究所	2.0	312014.588543	2.770781e+06
5	6	分子生物研究所	5.0	311934.340159	2.770738e+06
6	7	環安衛小組	3.0	311845.222216	2.770742e+06
7	8	生物化學研究所	6.0	311861.694253	2.770683e+06
8	9	生物體研究中心	18.0	311991.358748	2.770611e+06
9	10	農業生物科技研究中心	19.0	312051.756216	2.770683e+06
10	11	生物多樣性研究博物館	NaN	312104.551205	2.770706e+06
11	12	中央研究院溫室	27.0	311626.439780	2.770641e+06
12	13	植物分子育種溫室	26.0	311763.706753	2.770646e+06
13	14	中國文哲研究所	21.0	311783.980029	2.770563e+06
14	15	生物多樣性研究博物館	11.0	312148.898996	2.770644e+06
15	16	動物房	NaN	312039.930138	2.770569e+06
16	17	蔡元培紀念館	14.0	311973.197271	2.770549e+06
17	18	統計所	15.0	311941.942637	2.770509e+06
18	19	歸國學人宿舍	NaN	311765.396192	2.770466e+06
19	20	人文社會科學館	24.0	311699.508045	2.770458e+06
20	21	中央研究院活動中心	20.0	311829.172540	2.770444e+06
21	22	中央研究院綜合體育館	23.0	311897.594846	2.770403e+06
22	23	地球科學研究所	22.0	311959.681754	2.770392e+06
23	24	環境變遷研究中心	25.0	311955.035795	2.770354e+06
24	25	資訊科學研究所	32.0	312031.060580	2.770462e+06
25	26	人文社會科學研究中心	31.0	312080.476690	2.770482e+06
26	27	化學研究所	30.0	312120.178522	2.770493e+06
27	28	資訊科技創新研究中心	13.0	312104.551205	2.770566e+06
28	29	行政大樓	10.0	312176.352391	2.770591e+06
29	30	郵局/福利社	16.0	312212.675344	2.770646e+06
30	31	近史所檔案館	42.0	312066.116453	2.770428e+06
31	32	嶺南美術館/歐美所	41.0	312046.687897	2.770393e+06
32	33	近史所	35.0	312129.048080	2.770393e+06
33	34	近代史研究所	35.0	312147.631917	2.770409e+06
34	35	物理學研究所	33.0	312255.333695	2.770461e+06
35	36	胡適紀念館	34.0	312194.513867	2.770409e+06
36	37	經濟學研究所	39.0	312282.364730	2.770380e+06
37	38	傅斯年圖書館	38.0	312230.836820	2.770388e+06
38	39	亭灣考古館	43.0	312188.178469	2.770363e+06

In [7]:
```
# 資料處理：排序
## 函數用法：pandas.DataFrame.sort_values -> https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sort_values.html
IIS_sorted_df = IIS_df.sort_values(by="Map_Index",ascending=True)
```

In [8]:
```
# 來看看結果：
IIS_sorted_df
```

Out[8]:

	id	Name	Map_Index	X	Y
47	48	院區大門	1.0	312165.614116	2.770765e+06
4	5	生物醫學科學研究所	2.0	312014.588543	2.770781e+06
6	7	環安衛小組	3.0	311845.222216	2.770742e+06
2	3	細胞個體生物研究所	4.0	311937.719038	2.770836e+06
5	6	分子生物研究所	5.0	311934.340159	2.770738e+06
7	8	生物化學研究所	6.0	311861.694253	2.770683e+06
3	4	國家動物中心	7.0	311865.917852	2.770839e+06
1	2	跨領域大樓	8.0	311944.899157	2.770908e+06
28	29	行政大樓	10.0	312176.352391	2.770591e+06
14	15	生物多樣性研究博物館	11.0	312148.898996	2.770644e+06
44	45	植物暨微生物學研究所	12.0	312082.367880	2.770611e+06
27	28	資訊科技創新研究中心	13.0	312104.551205	2.770566e+06
16	17	蔡元培紀念館	14.0	311973.197271	2.770549e+06
17	18	統計所	15.0	311941.942637	2.770509e+06
29	30	郵局/福利社	16.0	312212.675344	2.770646e+06
8	9	生物體研究中心	18.0	311991.358748	2.770611e+06
9	10	農業生物科技研究中心	19.0	312051.756216	2.770683e+06
20	21	中央研究院活動中心	20.0	311829.172540	2.770444e+06
13	14	中國文哲研究所	21.0	311783.980029	2.770563e+06
22	23	地球科學研究所	22.0	311959.681754	2.770392e+06
21	22	中央研究院綜合體育館	23.0	311897.594846	2.770403e+06
19	20	人文社會科學館	24.0	311699.508045	2.770458e+06
23	24	環境變遷研究中心	25.0	311955.035795	2.770354e+06
12	13	植物分子育種溫室	26.0	311763.706753	2.770646e+06
11	12	中央研究院溫室	27.0	311626.439780	2.770641e+06
45	46	環境變遷研究大樓	28.0	311948.175998	2.770322e+06
26	27	化學研究所	30.0	312120.178522	2.770493e+06
25	26	人文社會科學研究中心	31.0	312080.476690	2.770482e+06
24	25	資訊科學研究所	32.0	312031.060580	2.770462e+06
34	35	物理學研究所	33.0	312255.333695	2.770461e+06
35	36	胡適紀念館	34.0	312194.513867	2.770409e+06
39	40	近代史研究所	35.0	312127.358641	2.770363e+06
32	33	近史所	35.0	312129.048080	2.770393e+06
33	34	近代史研究所	35.0	312147.631917	2.770409e+06
40	41	歐美所圖書館	36.0	312076.253091	2.770308e+06
41	42	歐美研究所	36.0	312122.290322	2.770303e+06
42	43	歷史語言研究所	37.0	312181.420710	2.770308e+06
37	38	傅斯年圖書館	38.0	312230.836820	2.770388e+06
36	37	經濟學研究所	39.0	312282.364730	2.770380e+06

資料輸出模組

目的： 將屬性資料輸出成表格
使用： io,pandas

輸入：
輸入：

- 資料表（pandas.DataFrame）：屬性資料。
  輸出：
- CSV檔檔案路徑（String）：輸出CSV的檔案路徑。

參考資料：

- pandas.DataFrame.to_csv (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.to_csv.html)
- python io模組 (https://docs.python.org/3/library/io.html)
- python 基礎教學 (https://openhome.cc/Gossip/CodeData/PythonTutorial/index.html)

io 基礎使用：

```
with open('spam.txt', 'w') as file:
    file.write('Spam and eggs!')
```

In [9]:
```
# 指定輸出的路徑：
Out_CSV_Path = r".\DataExport.csv"
```

In [10]:
```
# 先試試看如果使用 pandas.DataFrame.to_csv 預設會輸出什麼：
IIS_sorted_df.to_csv()
```

Out[10]:
```
',id,Name,Map_Index,X,Y\n47,48,院區大
門,1.0,312165.6141163437,2770765.4763012654\n4,5,生物醫學科學研究
所,2.0,312014.58854297985,2770781.0833038767\n6,7,環安衛小
組,3.0,311845.22221647046,2770741.803831644\n2,3,細胞個體生物研究
所,4.0,311937.71903818013,2770835.567733103\n5,6,分子生物研究
所,5.0,311934.34015884827,2770737.580232479\n7,8,生物化學研究
所,6.0,311861.6942532133,2770683.095803253\n3,4,國家動物中
心,7.0,311865.91785237816,2770838.5242525185\n1,2,跨領域大
樓,8.0,311944.8991567603,2770907.7912788214\n28,29,行政大
樓,10.0,312176.3523909927,2770590.598981543\n14,15,生物多樣性研究博物
館,11.0,312148.8989964212,2770643.81633102\n44,45,植物暨微生物學研究
所,12.0,312082.36788022734,2770611.0637244936\n27,28,資訊科技創新研究中
心,13.0,312104.5512051906,2770566.1021063877\n16,17,蔡元培紀念
館,14.0,311973.1972711647,2770548.7853498105\n17,18,統計
所,15.0,311941.9426373448,2770509.0835176613\n29,30,郵局/福利
社,16.0,312212.6753438102,2770645.928130602\n8,9,生物體研究中
心,18.0,311991.35874757334,2770611.294617451\n9,10,農業生物科技研究中
心,19.0,312051.75621563045,2770682.673443336\n20,21,中央研究院活動中
心,20.0,311829.1725396441,2770444.4624504405\n13,14,中國文哲研究
所,21.0,311783.98002858047,2770563.1455869726\n22,23,地球科學研究
所,22.0,311959.6817538372,2770391.66746088\n21,22,中央研究院綜合體育
館,23.0,311897.5948461144,2770402.6488187085\n19,20,人文社會科學
館,24.0,311699.5080452841,2770458.400327684\n23,24,環境變遷研究中
心,25.0,311955.03579475597,2770353.6550683966\n12,13,植物分子育種溫
室,26.0,311763.7067525894,2770646.350490518\n11,12,中央研究院溫
室,27.0,311626.4397797326,2770641.2821715204\n45,46,環境變遷研究大
樓,28.0,311948.1759980335,2770322.4592655287\n26,27,化學研究
所,30.0,312120.1785221005,2770493.4562007515\n25,26,人文社會科學研究中
心,31.0,312080.47668995126,2770481.63012309\n24,25,資訊科學研究
所,32.0,312031.06057972275,2770462.201566932\n34,35,物理學研究
所,33.0,312255.33369537484,2770461.3568470995\n35,36,胡適紀念
館,34.0,312194.51386740146,2770408.984217456\n39,40,近代史研究
所,35.0,312127.35864068073,2770363.369346476\n32,33,近史
所,35.0,312129.04808034666,2770392.512180713\n33,34,近代史研究
所,35.0,312147.63191667193,2770408.984217456\n40,41,歐美所圖書
館,36.0,312076.25309078646,2770307.6178375003\n41,42,歐美研究
所,36.0,312122.2903216829,2770303.394238336\n42,43,歷史語言研究
所,37.0,312181.4207099905,2770308.462557333\n37,38,傅斯年圖書
館,38.0,312230.83682021894,2770388.2885815483\n36,37,經濟學研究
所,39.0,312282.3647300298,2770380.263743135\n43,44,民族學研究
所,40.0,312259.13493462326,2770267.91600535\n31,32,嶺南美術館/歐美
所,41.0,312046.6878966326,2770392.934540629\n30,31,近史所檔案
館,42.0,312066.1164527908,2770427.990413697\n38,39,臺灣考古
館,43.0,312188.1784686542,2770362.946986559\n46,47,中研院宿舍
群,50.0,312270.65668557567,2770745.5182131124\n0,1,國際研究生教學研究大
樓/幼稚園,60.0,311858.3153738815,2771187.3935435326\n10,11,生物多樣性研
究博物館,,312104.5512051906,2770706.3255986595\n15,16,動物
房,,312039.93013796885,2770569.480985719\n18,19,歸國學人宿
舍,,311765.39619225526,2770466.002806181\n'
```

上面嘗試了以後，發現資料會包含 DataFrame 的 index（列的編號，輸出開頭是逗點，代表 header 該行的第一欄是空白的），我們參考說明文件 (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.to_csv.html)並調整輸入的參數：

```
IIS_sorted_df.to_csv(index=False)
```

In [11]:
```
# 引入第一個方法需要的 io 模組
import io
```

In [12]:
```
# 資料輸出： 這邊示範使用 io 模組

with io.open(Out_CSV_Path,"wt",encoding="UTF-8") as CSV_out:
    CSV_out.write("\ufeff" ) # Only for Windows, comment this line on other OS
    CSV_out.write(IIS_sorted_df.to_csv(index=False,encoding="UTF-8"))
    CSV_out.flush()
    print("File saved at '{}'.".format(Out_CSV_Path))
```

```
File saved at '.\DataExport.csv'.
```

In [ ]:
```
# 資料輸出：使用 pandas.DataFrame.to_csv ，請自行練習：
```

後記：

使用 Python 3 時預設編碼為 UTF-8，基本上不會遇到中文字無法解析的問題，不過 Python 2 預設使用 ascii，無法處理非英文字元，需使用「unicode 類別」處理，編碼處理可參考下面的作法：資料來源 (https://stackoverflow.com/questions/17912307/u-ufeff-in-python-string)

```
#!python2
#coding: utf8
u = u'ABC'
e8 = u.encode('utf-8')         # encode without BOM
e8s = u.encode('utf-8-sig')    # encode with BOM
e16 = u.encode('utf-16')       # encode with BOM
e16le = u.encode('utf-16le')   # encode without BOM
e16be = u.encode('utf-16be')   # encode without BOM
print 'utf-8     %r' % e8
print 'utf-8-sig %r' % e8s
print 'utf-16    %r' % e16
print 'utf-16le  %r' % e16le
print 'utf-16be  %r' % e16be
print
print 'utf-8  w/ BOM decoded with utf-8     %r' % e8s.decode('utf-8')
print 'utf-8  w/ BOM decoded with utf-8-sig %r' % e8s.decode('utf-8-sig')
print 'utf-16 w/ BOM decoded with utf-16    %r' % e16.decode('utf-16')
print 'utf-16 w/ BOM decoded with utf-16le  %r' % e16.decode('utf-16le')
```

Windows 的小麻煩：程式預設的編碼為系統語系的編碼
我們使用 UTF-8 存檔時，若使用 Excel 等程式開啟時，因為使用系統預設編碼（ 繁體中文系統為：CP950），因此中文字元會解為亂碼，若再檔頭加上一個位元順序記號 BOM (https://zh.wikipedia.org/wiki/%E4%BD%8D%E5%85%83%E7%B5%84%E9%A0%86%E5%BA%8F%E8%A8%98%E8%99%9F)：「 **\ufeff** 」才可以正確解碼。「\u」代表後面的字元是使用 Unicode 編碼，而feff為「EF BB BF」（16進位）的序列，Unicode 代號：U+FEFF，為代表使用 UTF-8 的編碼。
注意！ POSIX （Linux 系統）請「不要加上」上面的BOM，否則在執行過程會出現錯誤。

延伸練習：
在寫檔模組加入一個條件判斷（if述句）OS的種類，若是Windows則寫入 UTF-8 的 BOM。
參考資料： **sys.platform (https://docs.python.org/3/library/sys.html#sys.platform)**

OS	*platform* value
Linux	'linux'
Windows	'win32'
Windows/Cygwin	'cygwin'
Mac OS X	'darwin'
FreeBSD	'freebsd'

使用範例：（取自這邊 (https://docs.python.org/3/library/sys.html#sys.platform)）

```
if sys.platform.startswith('freebsd'):
    # FreeBSD-specific code here...
elif sys.platform.startswith('linux'):
```

In [ ]:
```
# 延伸練習：資料輸出，使用 io 模組
import sys
Out_CSV_Path = "DataExport_BOM.csv"
with io.open(Out_CSV_Path,"wt",encoding="UTF-8") as CSV_out:
    # ----------
    if : # Complete this line
    CSV_out.write("\ufeff" ) # Only for Windows, comment this line on
other OS
    # ----------
    CSV_out.write(IIS_sorted_df.to_csv(index=False,encoding="UTF-8"))
    CSV_out.flush()
    print("File saved at '{}'.".format(Out_CSV_Path))
```