

作者	ifcherng@gmail.com
原始碼	https://github.com/ifcherng/CAAC-Toolkit

目錄

一、	環境安裝.....	2
	1. 依照作業系統的位元數下載 Python 3.6	2
	2. 安裝 Python	2
	3. 雙擊 install_requirements.bat 安裝所需的 Python 套件	2
二、	第一階段-篩選結果	3
	1. 修改 do_crawl.bat 中的參數設定	3
	2. 抓取網站內容並建立本地資料庫.....	3
	3. 修改 do_lookup.bat 中的參數設定	4
	4. 從資料庫取出資料.....	4
三、	第二階段-交叉查榜	6
	1. 先執行過 第一階段-篩選結果	6
	2. 將要查詢的准考證號碼寫入 admission_ids.txt 中	6
	3. 雙擊 do_cross.bat.....	6
	4. 等待抓取當年度的網站內容，直到看見 [Done] It takes XXX seconds. 後關閉	6
	5. result_xxx.xlsx 即為結果，准考證的順序將與 admission_ids.txt 中的順序相同。	6

一、環境安裝

● Python 3

1. 依照作業系統的位元數下載 Python 3.6 或 3.7

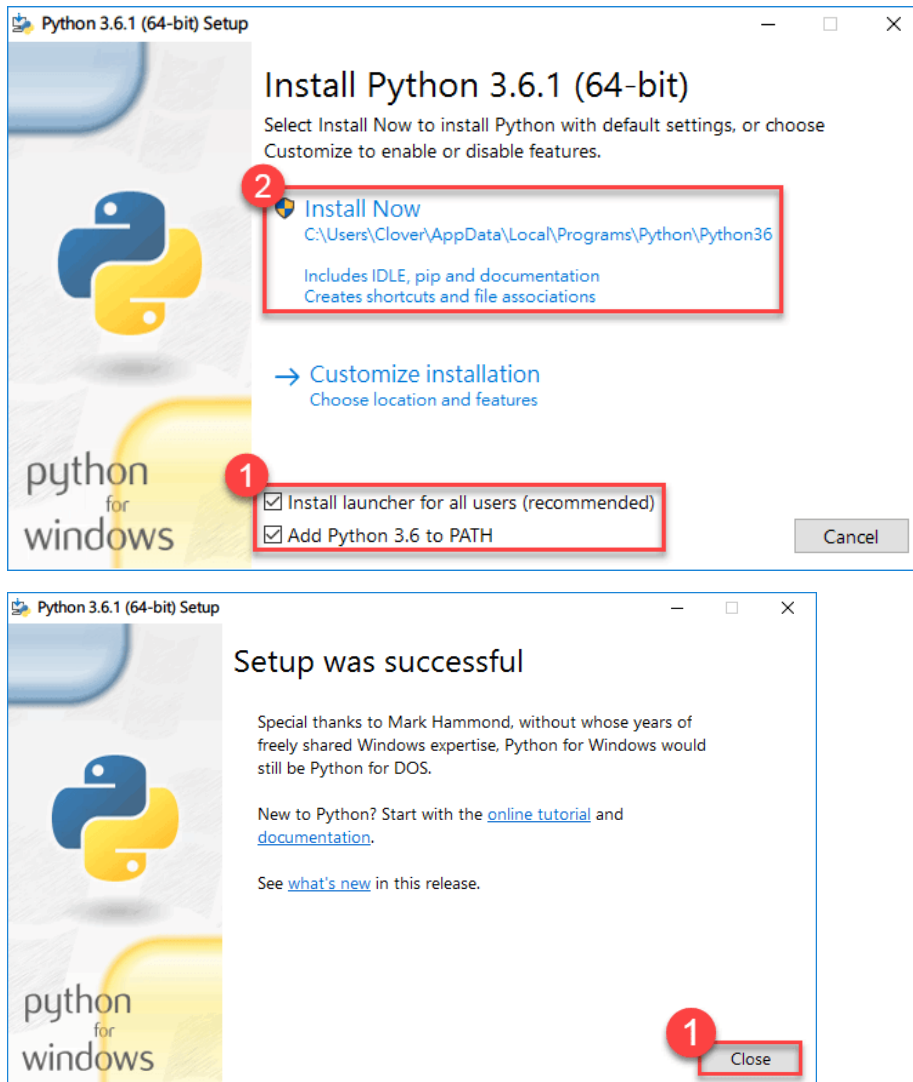
※備註：其他版本的 Python 3 或許也可以，但我沒做過測試。

- 32 位元：<https://www.python.org/ftp/python/3.6.8/python-3.6.8.exe>
- 64 位元：<https://www.python.org/ftp/python/3.6.8/python-3.6.8-amd64.exe>

如果不確定作業系統的位元數，你總是可以安裝 **32 位元** 的版本。

但現代的電腦應該都已經是使用 **64 位元** 的作業系統了。

2. 安裝 Python



3. 雙擊 install_requirements.bat 安裝所需的 Python 套件

二、 第一階段-篩選結果

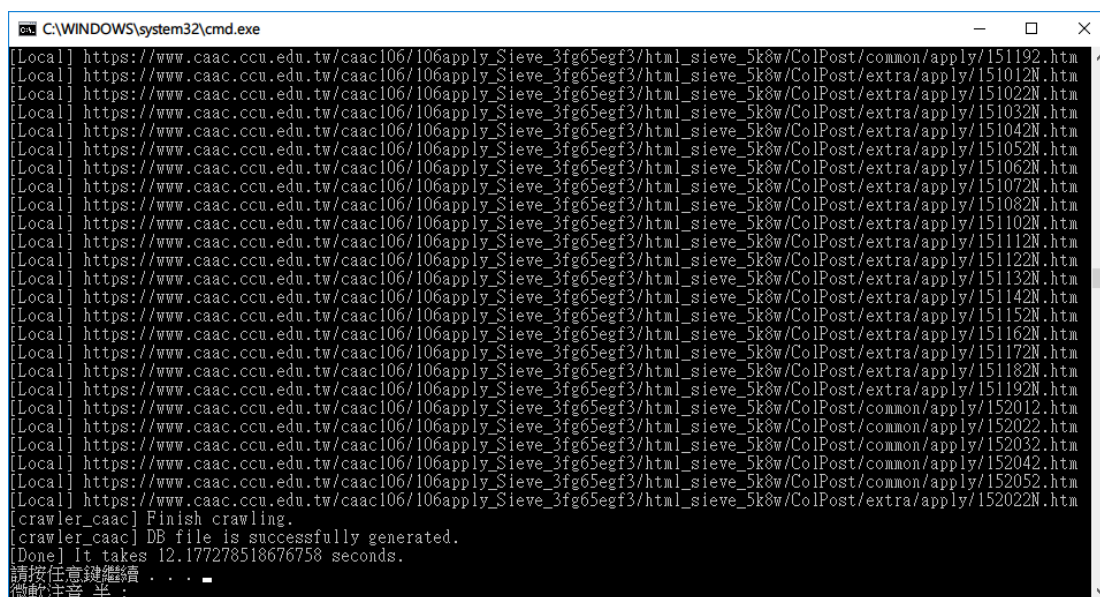
1. 修改 do_crawl.bat 中的參數設定
 - i. 對 do_crawl.bat 按 右鍵→編輯
 - ii. 打開網頁瀏覽器，找到當年度 第一階段篩選結果 的網頁



- iii. 複製該網頁的網址，貼上到 do_crawl.bat 中的 projectBaseUrl



- iv. 儲存並關閉檔案
2. 抓取網站內容並建立本地資料庫
 - i. 雙擊 do_crawl.bat
 - ii. 等待抓取當年度的網站內容，直到看見 **[Done] It takes XXX seconds.** 後關閉



此過程視網路環境的不同，可能耗費 5 至 40 分鐘。

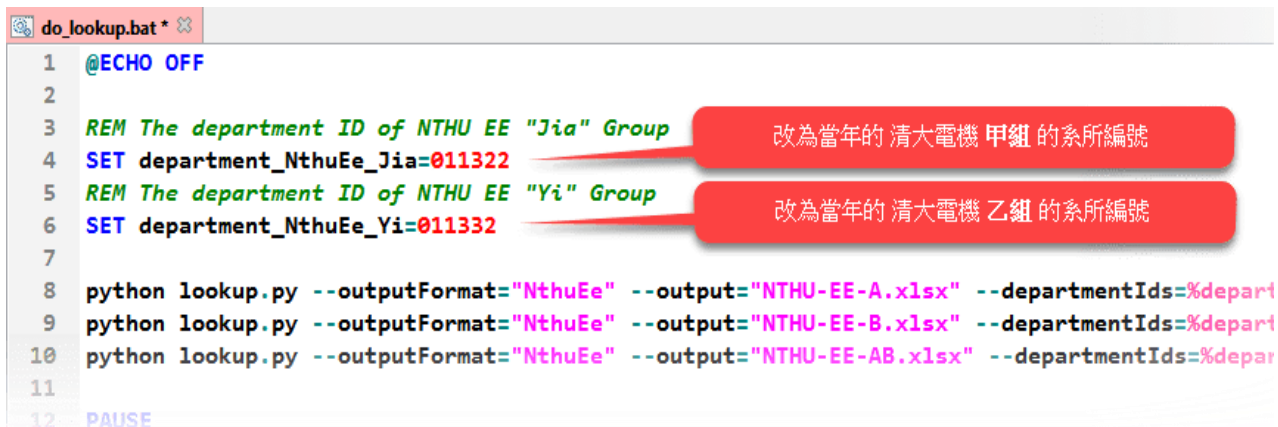
- iii. 如果想要**完全**重新抓取網站內容，可以先刪除 **crawler XXX** 資料夾

3. 修改 do_lookup.bat 中的參數設定

i. 對 do_lookup.bat 按 右鍵→編輯

ii. 修改參數設定

- department_NthuEe_Jia 為當年度的 清大電機甲組 的系所編號（6 位數）
- department_NthuEe_Yi 為當年度的 清大電機乙組 的系所編號（6 位數）



```
1 @ECHO OFF
2
3 REM The department ID of NTHU EE "Jia" Group
4 SET department_NthuEe_Jia=011322
5 REM The department ID of NTHU EE "Yi" Group
6 SET department_NthuEe_Yi=011332
7
8 python lookup.py --outputFormat="NthuEe" --output="NTHU-EE-A.xlsx" --departmentIds=%department_NthuEe_Jia%
9 python lookup.py --outputFormat="NthuEe" --output="NTHU-EE-B.xlsx" --departmentIds=%department_NthuEe_Yi%
10 python lookup.py --outputFormat="NthuEe" --output="NTHU-EE-AB.xlsx" --departmentIds=%department_NthuEe_Jia% %department_NthuEe_Yi%
11
12 PAUSE
```

iii. 儲存並關閉檔案

4. 從資料庫取出資料

i. 雙擊 do_lookup.bat



如果跑出如上圖一堆數字，那麼表示正確執行，可關閉該視窗。

ii. 產生的三個 .csv 檔案即為結果

- NTHU-EE-A.xlsx 為報名 清大電機甲組 者
- NTHU-EE-B.xlsx 為報名 清大電機乙組 者
- NTHU-EE-AB.xlsx 為報名 清大電機甲組 或 清大電機乙組 者

mylibs	2017/4/20 下午 09:...	檔案資料夾	
crawler.py	2017/4/19 下午 04:...	Python File	1 KB
do_crawl.bat	2017/4/12 下午 04:...	Windows 批次檔案	1 KB
do_lookup.bat	2017/4/20 下午 10:...	Windows 批次檔案	1 KB
lookup.py	2017/4/20 下午 10:...	Python File	4 KB
NTHU-EE-A.xlsx	2017/4/20 下午 10:...	Microsoft Excel 工...	7 KB
NTHU-EE-AB.xlsx	2017/4/20 下午 10:...	Microsoft Excel 工...	14 KB
NTHU-EE-B.xlsx	2017/4/20 下午 10:...	Microsoft Excel 工...	14 KB

	A	B	C	D	E
1	准考證號	校名與系所			
2	10000903	國立臺灣大學 化學工程學系	國立臺灣大學 材料科學與工程學系	國立臺灣大學 財務金融學系	中山醫學大學 醫學系
3	10004236	國立臺灣大學 機械工程學系	國立臺灣大學 材料科學與工程學系	國立清華大學 物理學系物理組	
4	10004334	國立臺灣大學 機械工程學系	國立臺灣大學 資訊工程學系	國立交通大學 電子工程學系	國立交通大學 電子工程學系
5	10004405	國立臺灣大學 機械工程學系	國立臺灣大學 電機工程學系	國立成功大學 機械工程學系	國立交通大學 電機工程學系
	國立臺灣大學	國立臺灣大學	國立交通大學	國立清華大學

三、 第二階段-交叉查榜

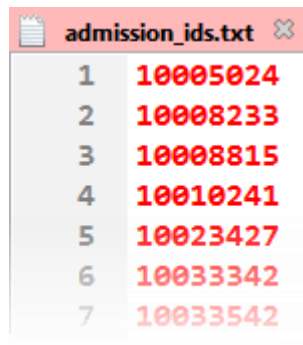
本階段的資料來源為：新鮮人查榜 <http://freshman.tw>

1. 先執行過 第一階段-篩選結果

如果以前已經做過 第一階段-篩選結果 可以跳過此步驟。

※ 主要是 data/crawler_xxx/sqlite3.db 必須已經產生

2. 將要查詢的准考證號碼寫入 admission_ids.txt 中



如左圖，一行一個准考證號。

可以直接從 Excel 那邊選取整列後直接複製過來。

3. 雙擊 do_cross.bat

4. 等待抓取當年度的網站內容，直到看見 **[Done] It takes XXX seconds.** 後關閉

5. result_xxx.xlsx 即為結果，准考證的順序將與 admission_ids.txt 中的順序相同。

若 admission_ids.txt 中有重複的准考證號，此處也會有重複以保證資料筆數相同。

※備註：因「落榜」與「未知狀態」的在網頁上看起來一樣，因此「落榜」實際上可能是「未知狀態」。也可能在 Excel 裡看到被分發到「落榜」的校系，實際上該分發結果（皇冠）是正確的，只是該「落榜」實際上是「未知狀態」。這可能與臺灣大學的個資保護政策有關。

	A	B	C	D	E	F	G	H
1	准考證號	考生姓名	校系名稱	榜單狀態				
2	10004829		國立成功大學 電機工程學系	備62	國立中央大學 電機工程學系	正12	國立清華大學 資訊工程學系乙組(資訊工程組)	備123
3	10007501		國立臺灣大學 機械工程學系	備6	國立臺灣大學 生物產業機電工程學系	落	國立交通大學 資訊工程學系資訊工程組	備1
4	10007629		國立臺灣大學 電機工程學系	備45	國立臺灣大學 資訊工程學系	落	國立交通大學 電子工程學系(乙組)	落
5	10008840		國立臺灣大學 物理學系	落	國立臺灣大學 工商管理學系科技管理組	正15	國立臺灣大學 電機工程學系	落
6	10008927		中國醫藥大學 醫學系	備83	中山醫學大學 醫學系	備37	長庚大學 醫學系	備201