

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 3.6.1 （或更高版本）	2
2.	安裝 Python	2
二、	第一階段-篩選結果	3
1.	抓取網站內容並建立本地資料庫.....	3
2.	修改 do_lookup.bat 中的參數設定	3
3.	從資料庫取出資料.....	3
三、	第二階段-交叉查榜	5
1.	先執行過 第一階段-篩選結果	5
2.	將要查詢的准考證號碼寫入 admission_ids.txt 中	5
3.	雙擊 do_cross.bat.....	5
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
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。	5

一、環境安裝

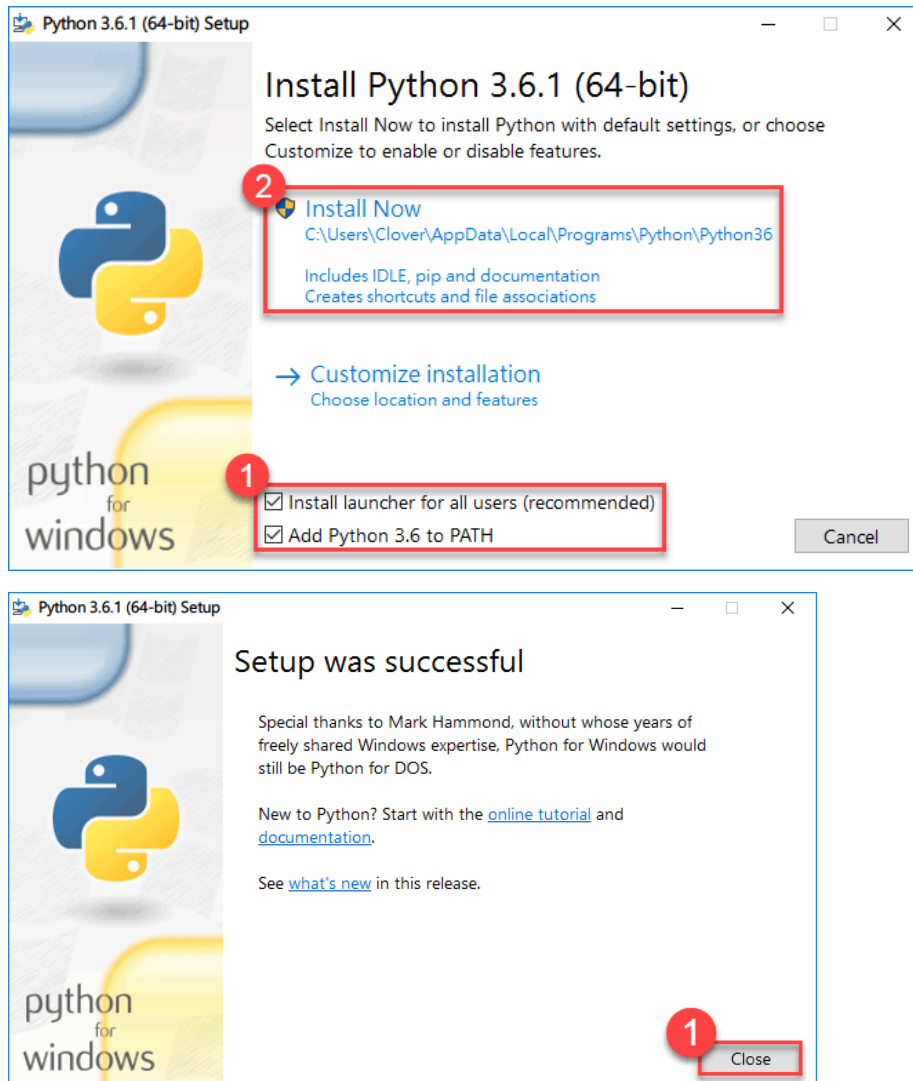
● Python 3

1. 依照作業系統的位元數下載 Python 3.6 （或更高版本）

- 32 位元：<https://www.python.org/ftp/python/3.6.2/python-3.6.2.exe>
- 64 位元：<https://www.python.org/ftp/python/3.6.2/python-3.6.2-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。

2. 安裝 Python

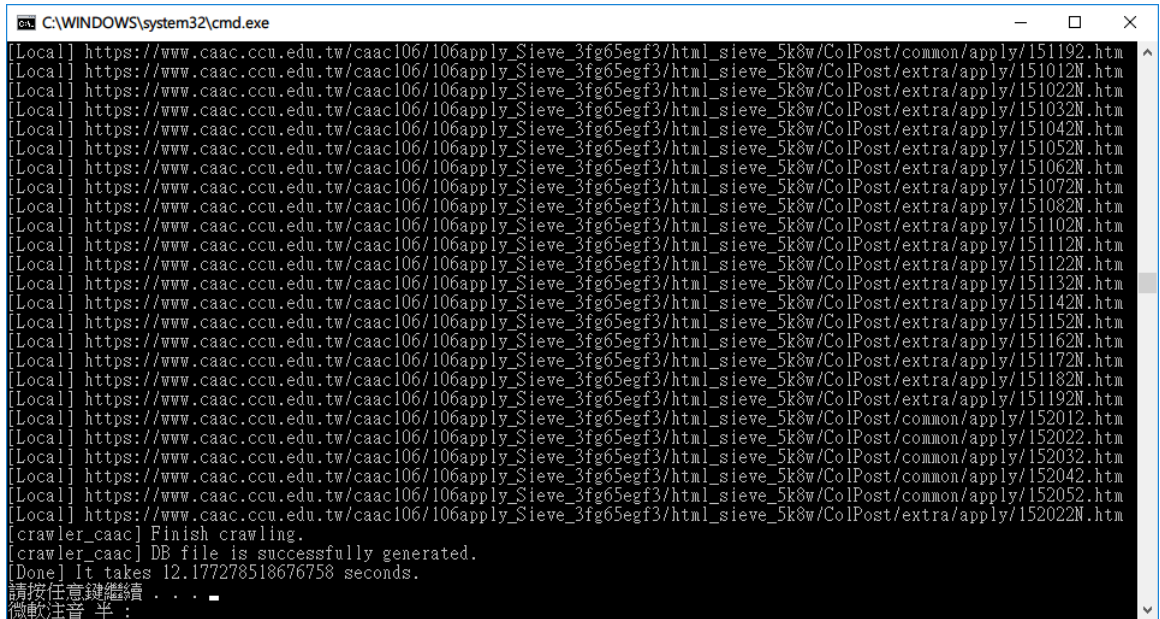


3. 雙擊 `install_requirements.bat` 安裝所需的 Python 套件

二、 第一階段-篩選結果

1. 抓取網站內容並建立本地資料庫

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



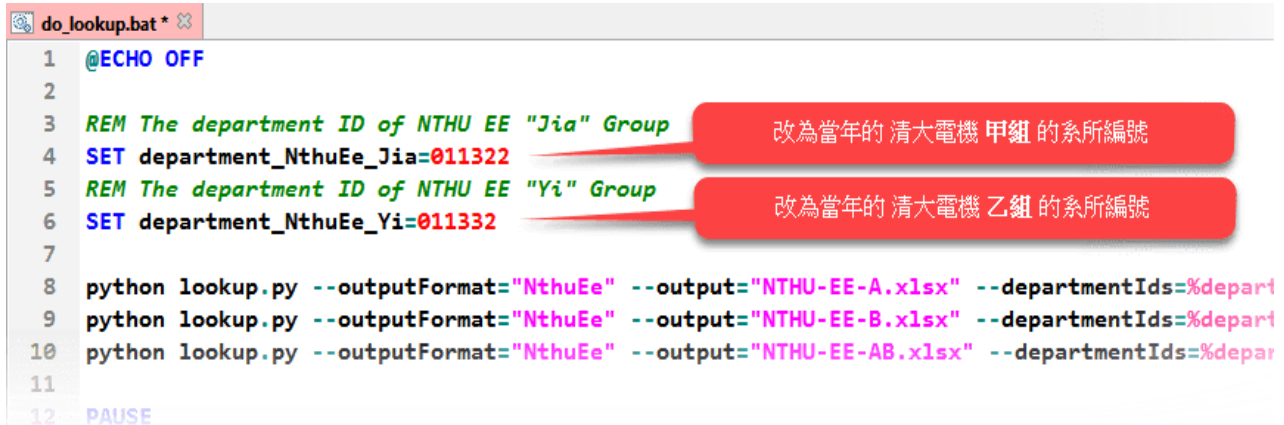
```
C:\WINDOWS\system32\cmd.exe
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/common/apply/151192.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151012N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151022N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151032N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151042N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151052N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151062N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151072N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151082N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151102N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151112N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151122N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151132N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151142N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151152N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151162N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151172N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151182N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151192N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/common/apply/152012.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/common/apply/152022.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/common/apply/152032.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/common/apply/152042.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/common/apply/152052.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/152022N.htm
[crawling_caac] Finish crawling.
[crawling_caac] DB file is successfully generated.
[Done] It takes 12.177278518676758 seconds.
請按任意鍵繼續...
```

此過程視網路環境的不同，可能耗費 20 至 40 分鐘。

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

2. 修改 do_lookup.bat 中的參數設定

- 對 do_lookup.bat 按 右鍵→編輯
- 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）



```
do_lookup.bat
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%
11
12 PAUSE
```

- 儲存並關閉檔案

3. 從資料庫取出資料

- 雙擊 do_lookup.bat

三、 第二階段-交叉查榜

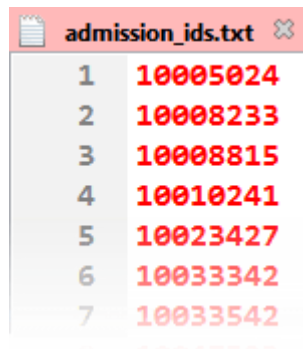
本階段的資料來源為：<http://freshman.tw>

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。

	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	10123821	██████	國立臺灣大學 機械工程學系	正14	國立臺灣大學 電機工程學系	正24	國立交通大學 電子工程學系(甲組)	備3	國立 電機
3	10131724	██████	國立臺灣大學 資訊工程學系	正57	國立交通大學 電機資訊學士班	正	國立交通大學 資訊工程學系資訊工程組	落	國立 資訊
4	10292915	██████	國立臺灣大學 哲學系	正7	國立臺灣大學 財務金融學系	正38	國立臺灣大學 資訊工程學系	正32	國立 電機
5	10142339	██████	國立臺灣大學 材料科學與工程學系	正11	國立臺灣大學 電機工程學系	正43	國立臺灣大學 資訊工程學系	落	國立 電機
6	10049438	██████	國立清華大學 電機工程學系甲組	正	國立清華大學 電機工程學系乙組	落			
7	10005024	██████	國立臺灣大學 機械工程學系	備75	國立臺灣大學 電機工程學系	落	國立交通大學 電子工程學系(甲組)	備4	國立 電機
8	10105503	██████	國立臺灣大學 物理學系	正13	國立臺灣大學 電機工程學系	落	國立交通大學 電子工程學系(甲組)	備15	國立 電子
9	10260518	██████	高雄醫學大學	備137	高雄醫學大學	落	中國醫藥大學	落	中國