

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

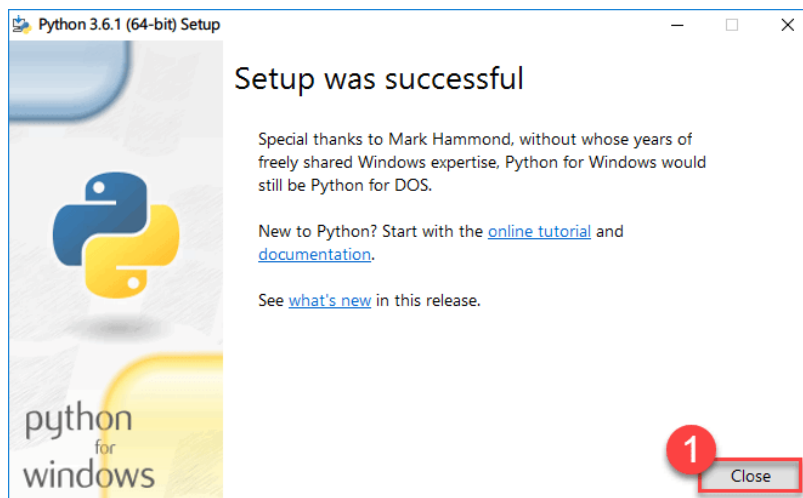
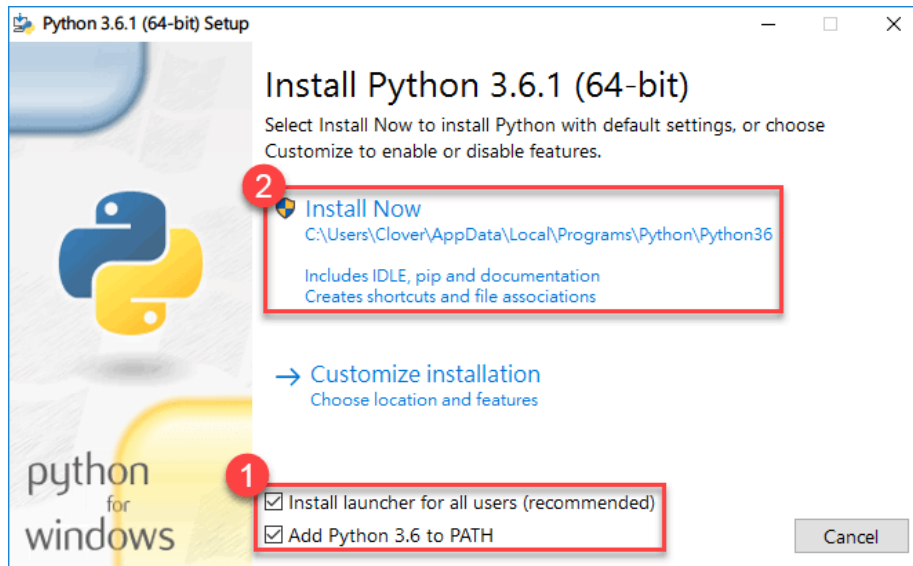
● Python 3

1. 依照作業系統的位元數下載 Python 3.6.1 （或更高版本）

- 32 位元：<https://www.python.org/ftp/python/3.6.1/python-3.6.1.exe>
- 64 位元：<https://www.python.org/ftp/python/3.6.1/python-3.6.1-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

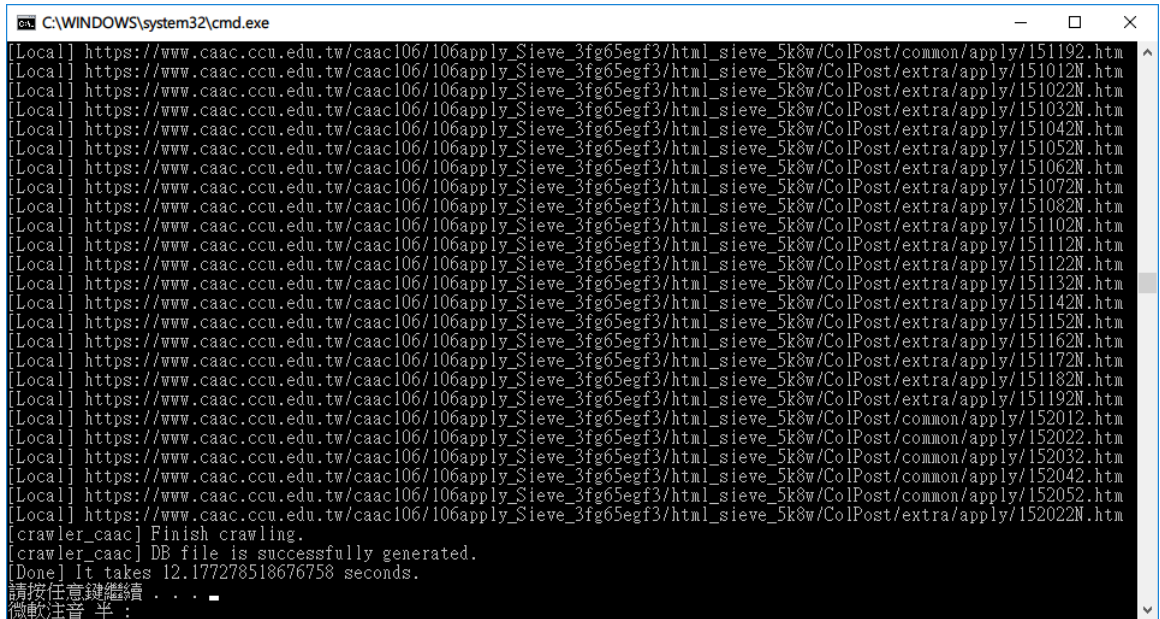


3. 雙擊 install_requirements.bat 安裝所需的 Python 套件

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



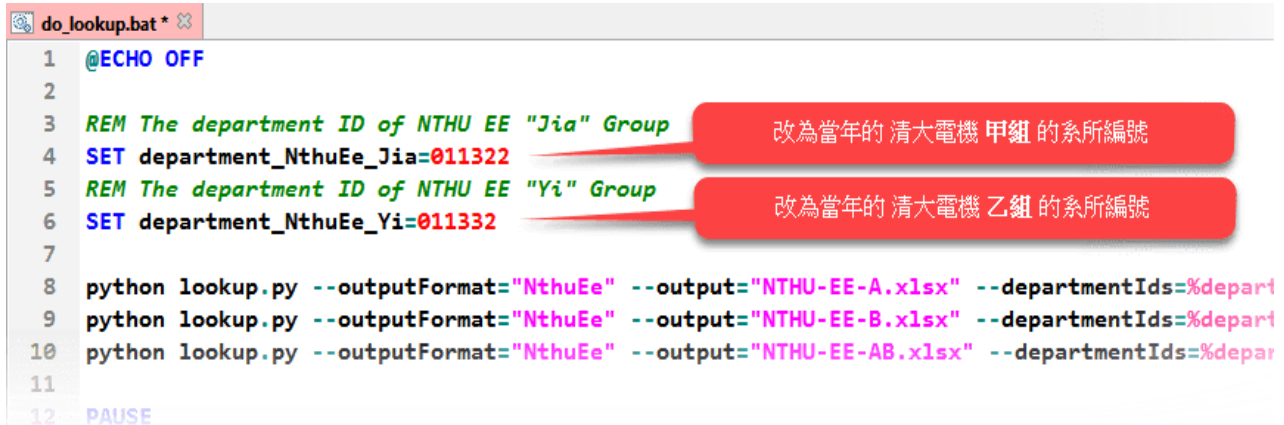
```
C:\WINDOWS\system32\cmd.exe
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/common/apply/151192.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/apply/151012N.htm
[Local] https://www.caac.ccu.edu.tw/caac106/106apply_Sieve_3fg65egf3/html_sieve_5k8w/ColPost/extra/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% %department_NthuEe_Yi%
11
12 PAUSE
```

- 儲存並關閉檔案

3. 從資料庫取出資料

- 雙擊 do_lookup.bat

```

C:\WINDOWS\system32\cmd.exe
['007012', '007042', '011322', '011332', '012012', '012022'], ('10263834', ['004382', '0113
32', '013052', '013062']), ('10264916', ['004382', '011332', '013042', '013052', '013062']),
('10265332', ['004382', '011332', '013052', '013062']), ('10266018', ['001102', '004102', '01
1182', '011332']), ('10266116', ['004382', '011332', '012052', '013042', '013052', '013062']), ('102668
02', ['004382', '011332', '013052', '013062']), ('10267232', ['001302', '004102', '004382',
011282', '011312', '011332']), ('10267234', ['004122', '004402', '011332']), ('10267418', ['0
01102', '001302', '001542', '011332', '013062']), ('10267539', ['001192', '001322', '011312',
'011332', '013062']), ('10267717', ['001532', '007042', '011322', '011332', '013062']), ('10267933', ['001532', '004382', '011332', '013052', '013062']), ('10269017', ['004382',
'011332', '013062']), ('10269215', ['004382', '011332', '011342', '013042', '013062']), ('1
0269401', ['001302', '001532', '011332', '011362', '013042', '013172']), ('10270005', ['0013
02', '001532', '004382', '011332', '013062', '013082']), ('10270201', ['001542', '011332', '0
13062']), ('10270441', ['001102', '001532', '001542', '011322', '011362', '013062']), ('10270
539', ['004382', '011332', '013062']), ('10271015', ['011182', '011332']), ('10280832', ['001
542', '011332', '013082']), ('10282902', ['001512', '004382', '011332', '011352', '013042']),
('10284237', ['001532', '004382', '011332', '013042', '013062', '027082']), ('10291216', ['0
01542', '004382', '004392', '011322', '011332', '011362']), ('10291218', ['004102', '004182',
'004332', '004362', '004382', '011332']), ('10292915', ['001042', '001492', '001542', '01132
2', '013052', '013082']), ('10295003', ['001542', '011332', '013062', '013082', '013092']), ('
10299236', ['004382', '011332', '012052', '016172']), ('10301630', ['004382', '011282', '011
332', '013042', '013062']), ('10305207', ['001532', '004382', '011322', '011332', '013032', '
09042']), ('19340104', ['011222', '011292', '011302', '011312', '011332', '013062']))
請按任意鍵繼續 . . .
微軟注音 半：

```

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

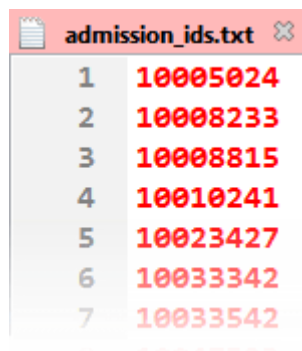
本階段的資料來源為：<http://freshman.tw>

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	
1	准考證號	校系	榜單狀態					
2	10005024	國立臺灣大學 機械工程學系	未放榜	國立臺灣大學 電機工程學系	未放榜	國立交通大學 電子工程學系	備4	國立清 電機工
3	10008233	國立臺灣大學 電機工程學系	未放榜	國立臺灣大學 資訊工程學系	未放榜	國立交通大學 電機工程學系	落	國立交 資訊工
4	10008815	國立臺灣大學 機械工程學系	未放榜	國立臺灣大學 電機工程學系	未放榜	國立交通大學 電機資訊學士班	落	國立交 電子工
5	10010241	國立臺灣大學 材料科學與工程學系	未放榜	國立臺灣大學 電機工程學系	未放榜	國立臺灣大學 資訊工程學系	未放榜	國立交 電子工
6	10023427	國立臺灣大學 資訊工程學系	未放榜	國立交通大學 資訊工程學系資電工程組	正	國立清華大學 電機工程學系乙組	正	國立清 電機工
7	10033342	國立臺灣大學 材料科學與工程學系	未放榜	國立臺灣大學 財務金融學系	未放榜	國立臺灣大學 資訊工程學系	未放榜	國立交 電機工
8	10033542	國立臺灣大學 資訊工程學系	未放榜	國立交通大學 電機資訊學士班	備26	國立交通大學 電子工程學系	正	國立交 電機工
9	10047523	國立臺灣大學 經濟學系	未放榜	國立臺灣大學 機械工程學系	未放榜	國立臺灣大學 資訊工程學系	未放榜	國立交 電子工
	10048006	國立臺灣大學	未放榜	國立臺灣大學	未放榜	國立交通大學	備2	國立交