**大學甄試入學工具箱**

|  |  |
| --- | --- |
| 作者 | [jfcherng@gmail.com](mailto:jfcherng@gmail.com) |
| 原始碼 | <https://github.com/jfcherng/CAAC-Toolkit> |

目錄

[一、 環境安裝 2](#_Toc510050745)

[1. 依照作業系統的位元數下載 Python 3.6 2](#_Toc510050746)

[2. 安裝 Python 2](#_Toc510050747)

[3. 雙擊 install\_requirements.bat 安裝所需的 Python套件 2](#_Toc510050748)

[二、 第一階段-篩選結果 3](#_Toc510050749)

[1. 修改 do\_ crawl.bat 中的參數設定 3](#_Toc510050750)

[2. 抓取網站內容並建立本地資料庫 3](#_Toc510050751)

[3. 修改 do\_lookup.bat 中的參數設定 4](#_Toc510050752)

[4. 從資料庫取出資料 4](#_Toc510050753)

[三、 第二階段-交叉查榜 6](#_Toc510050754)

[1. 先執行過 **第一階段-篩選結果** 6](#_Toc510050755)

[2. 將要查詢的准考證號碼寫入 admission\_ids.txt 中 6](#_Toc510050756)

[3. 雙擊 do\_cross.bat 6](#_Toc510050757)

[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](#_Toc510050758)

[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](#_Toc510050759)

1. 環境安裝

* Python 3

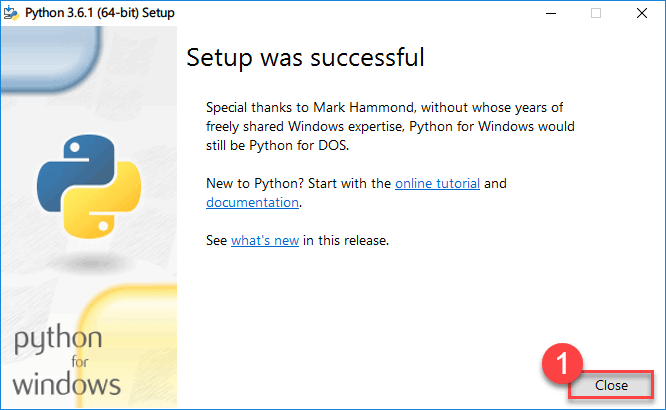
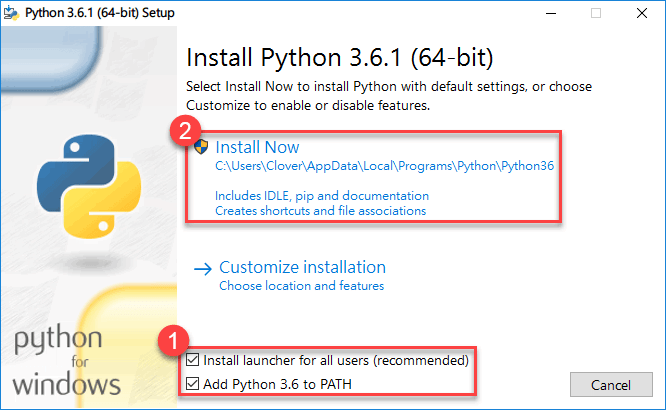
1. 依照作業系統的位元數下載 Python 3.6

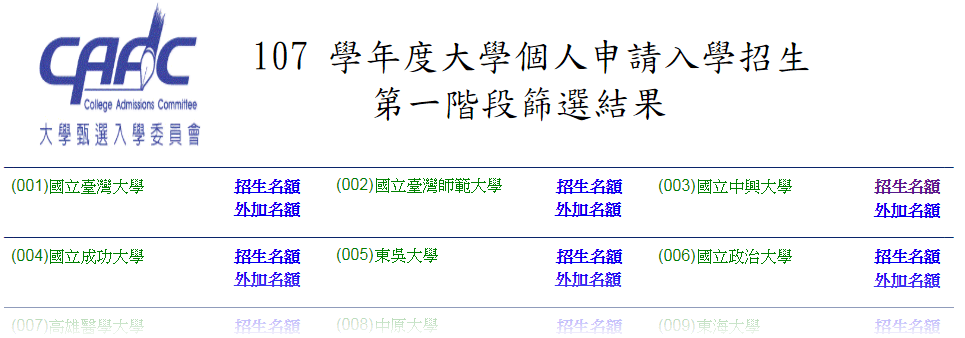
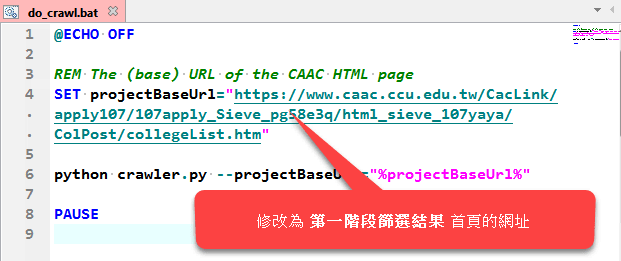
※備註：其他版本的 Python 3 或許也可以，但我沒做過測試。

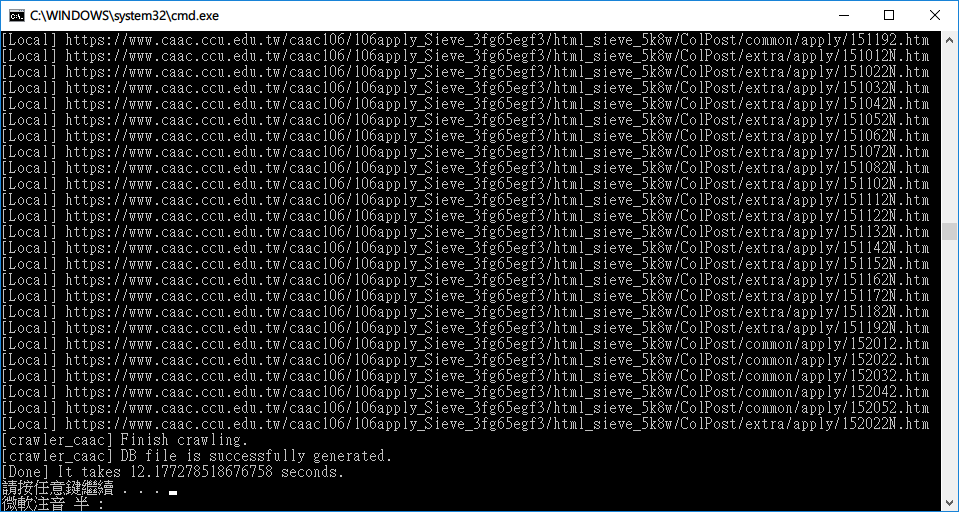
* 32位元：<https://www.python.org/ftp/python/3.6.5/python-3.6.5.exe>
* 64位元：<https://www.python.org/ftp/python/3.6.5/python-3.6.5-amd64.exe>

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。

1. 安裝 Python

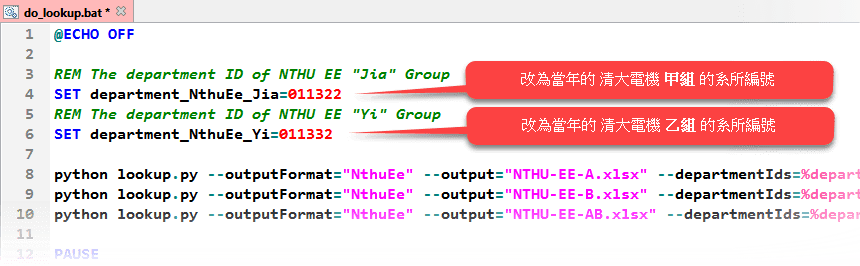


1. 雙擊 install\_requirements.bat 安裝所需的 Python套件
2. 第一階段-篩選結果
   1. 修改 do\_ crawl.bat 中的參數設定
      1. 對 do\_ crawl.bat 按 右鍵→編輯
      2. 打開網頁瀏覽器，找到當年度 **第一階段篩選結果** 的網頁
      3. 複製該網頁的網址，貼上到 do\_ crawl.bat 中的 **projectBaseUrl**
      4. 儲存並關閉檔案
   2. 抓取網站內容並建立本地資料庫
      1. 雙擊 do\_crawl.bat
      2. 等待抓取當年度的網站內容，直到看見 **[Done] It takes XXX seconds.** 後關閉

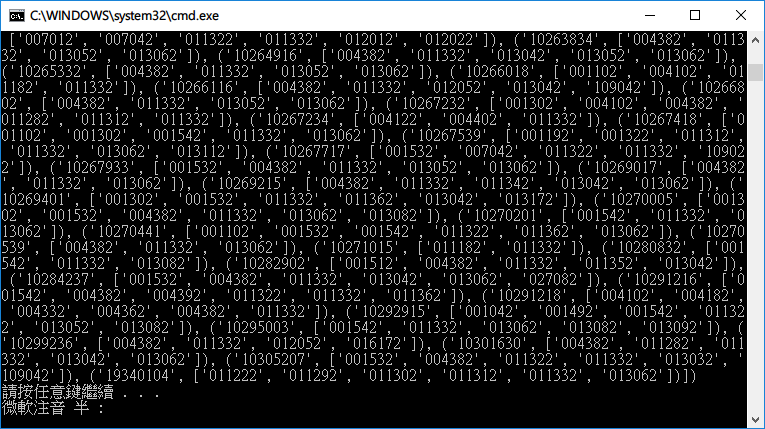


此過程視網路環境的不同，可能耗費5至40分鐘。

* + 1. 如果想要**完全**重新抓取網站內容，可以先刪除 **crawler\_XXX** 資料夾
  1. 修改 do\_lookup.bat 中的參數設定
     1. 對 do\_lookup.bat 按 右鍵→編輯
     2. 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）

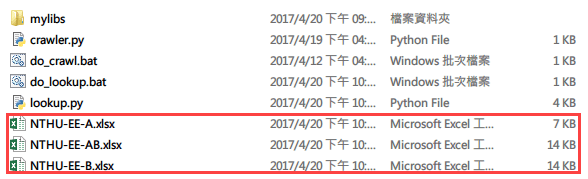


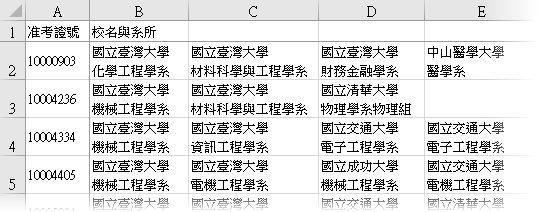
* + 1. 儲存並關閉檔案
  1. 從資料庫取出資料
     1. 雙擊 do\_lookup.bat



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。

* + 1. 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





1. 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

1. 將要查詢的准考證號碼寫入 admission\_ids.txt 中

|  |  |
| --- | --- |
| C:\Users\Clover\Desktop\asdasd.png | 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。 |

1. 雙擊 do\_cross.bat
2. 等待抓取當年度的網站內容，直到看見 **[Done] It takes XXX seconds.** 後關閉
3. 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。

