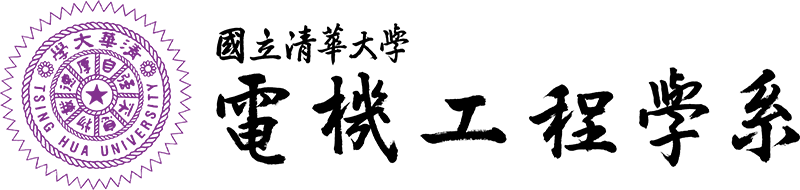
**大學甄試入學工具箱**



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

目錄

[一、 環境安裝 2](#_Toc41628052)

[1. 依照作業系統的位元數下載 Python 3.6或更高版本（低版本未測試） 2](#_Toc41628053)

[2. 安裝 Python 2](#_Toc41628054)

[3. 雙擊 install\_requirements.bat 安裝所需的 Python套件 2](#_Toc41628055)

[二、 第一階段-篩選結果（甄選委員會） 3](#_Toc41628056)

[1. 修改 do\_crawl.bat 中的參數設定 3](#_Toc41628057)

[2. 抓取網站內容並建立本地資料庫（只需進行一次） 4](#_Toc41628058)

[3. 修改 do\_lookup.bat 中的參數設定 4](#_Toc41628059)

[4. 從資料庫取出資料 5](#_Toc41628060)

[三、 第二階段-交叉查榜 6](#_Toc41628061)

[1. 初次使用時，需要先安裝 tesseract (OCR 工具)，一台電腦上只需要裝一次。 6](#_Toc41628062)

[2. 將要查詢的校系代碼寫入 department\_ids.txt 中，然後儲存並關閉檔案 6](#_Toc41628063)

[3. 雙擊 do\_cross.bat 6](#_Toc41628064)

[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](#_Toc41628065)

[5. result\_xxx.xlsx 即為結果 6](#_Toc41628066)

[四、 第二階段-分發結果（甄選委員會） 7](#_Toc41628067)

[1. 修改 do\_crawl.bat 中的參數設定 7](#_Toc41628068)

[2. 抓取網站內容並建立本地資料庫（只需進行一次） 8](#_Toc41628069)

[3. 將要查詢的准考證號寫入 admission\_ids.txt 中，然後儲存並關閉檔案 8](#_Toc41628070)

[4. 從資料庫取出資料 8](#_Toc41628071)

1. 環境安裝

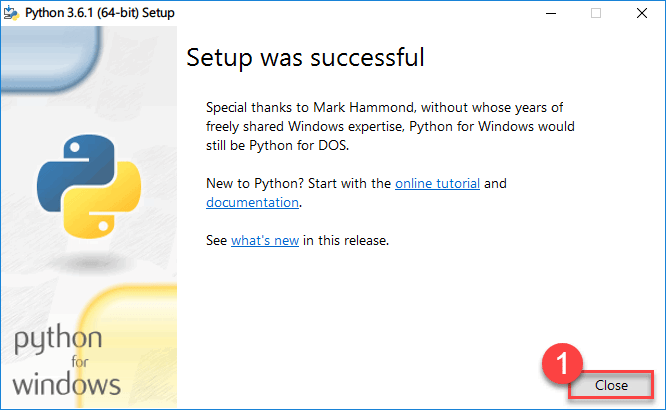
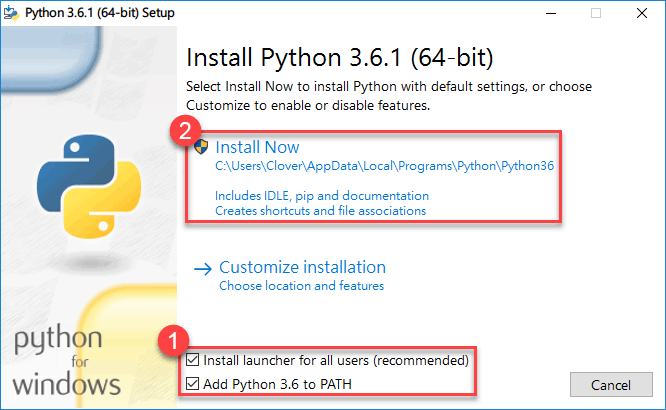
* Python 3

1. 依照作業系統的位元數下載 Python 3.6或更高版本（低版本未測試）

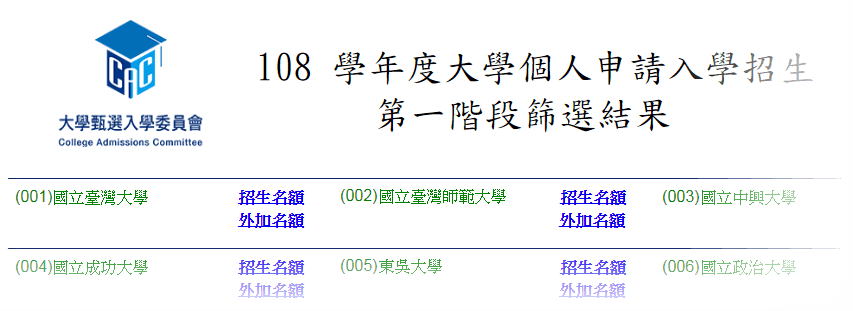
* 32位元：<https://www.python.org/ftp/python/3.8.3/python-3.8.3.exe>
* 64位元：<https://www.python.org/ftp/python/3.8.3/python-3.8.3-amd64.exe>

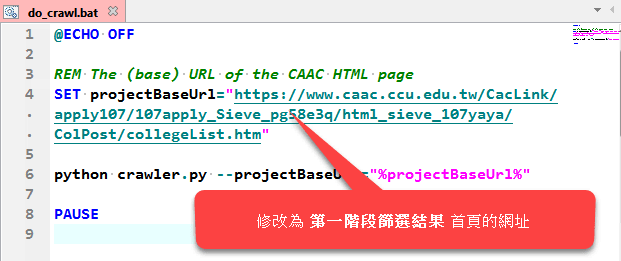
※備註：現代的電腦應該已經大多為 **64位元** 的作業系統了。

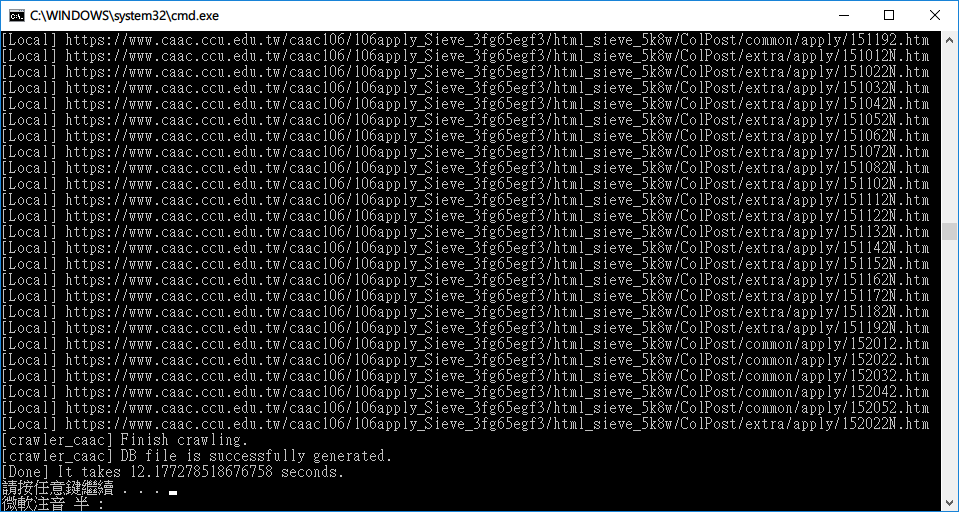
1. 安裝 Python



1. 雙擊 install\_requirements.bat 安裝所需的 Python套件
2. 第一階段-篩選結果（甄選委員會）
   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

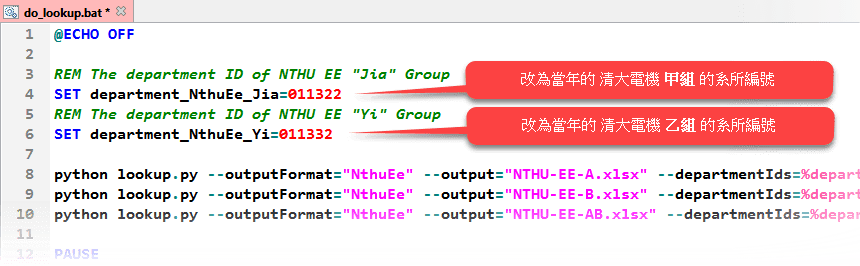


* + 1. 複製該網頁的網址，貼上到 do\_ crawl.bat 中的 **projectBaseUrl**
    2. 儲存並關閉檔案
  1. 抓取網站內容並建立本地資料庫（只需進行一次）
     1. 雙擊 do\_crawl.bat
     2. 等待抓取當年度的網站內容，直到看見 **[Done] It takes XXX seconds.** 後關閉

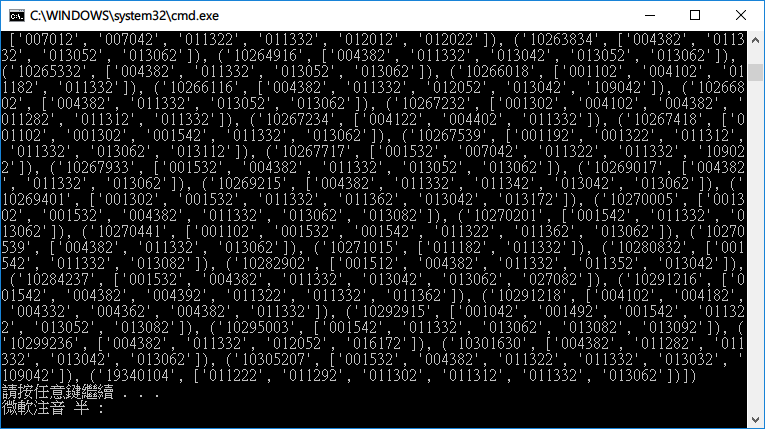


此過程視網路環境的不同，可能耗費2至5分鐘。

* + 1. 如果想要**完全**重新抓取網站內容，可以先刪除 **data/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）

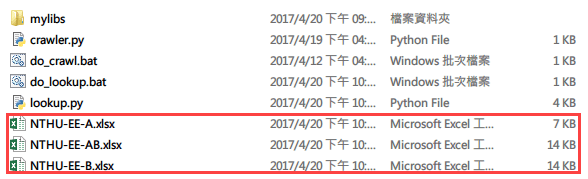


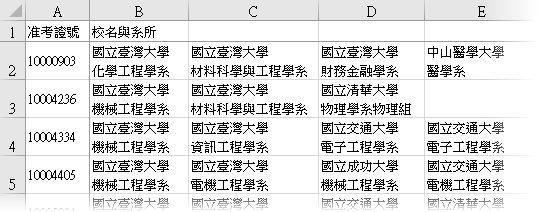
* + 1. 儲存並關閉檔案
  1. 從資料庫取出資料
     1. 雙擊 do\_lookup.bat



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 .xlsx 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：大學交叉查榜** [**https://www.com.tw/cross/**](https://www.com.tw/cross/)

1. 初次使用時，需要先安裝 [tesseract](https://github.com/UB-Mannheim/tesseract/wiki) (OCR 工具)，一台電腦上只需要裝一次。

* 32位元：[tesseract-ocr-w32-setup-v4.1.0.20190314.exe](https://digi.bib.uni-mannheim.de/tesseract/tesseract-ocr-w32-setup-v4.1.0.20190314.exe)
* 64位元：[tesseract-ocr-w64-setup-v4.1.0.20190314.exe](https://digi.bib.uni-mannheim.de/tesseract/tesseract-ocr-w64-setup-v4.1.0.20190314.exe)

※備註：現代的電腦應該已經大多為 **64位元** 的作業系統了。

安裝時只需要一直按下一步即可。

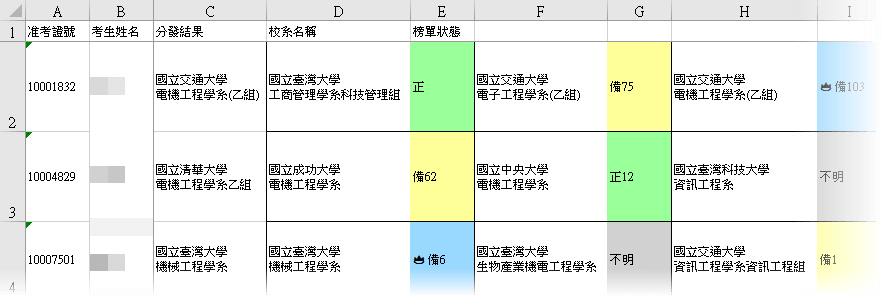
1. 將要查詢的校系代碼寫入 department\_ids.txt 中，然後儲存並關閉檔案

|  |  |
| --- | --- |
|  | 如左圖，一行寫一個校系代碼。  圖中範例為108年清大電機甲、乙組的校系代碼。 |

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

其內容以准考證號碼排序，若一名考生出現在多個校系代碼的結果中，在輸出的檔案裡只會出現一次。

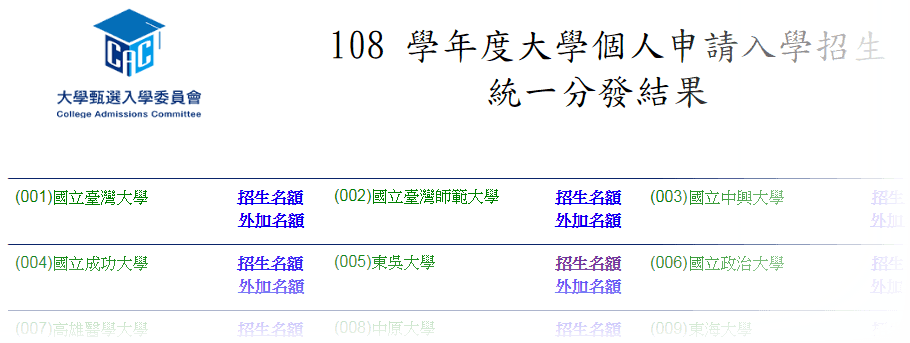
※備註：由於學校放榜時可能會做個資保護，因此網路上的資料可能不齊全，某些人的放榜狀態會是「不明」（可能是落榜） 。

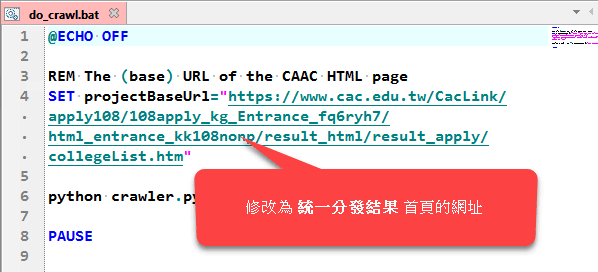


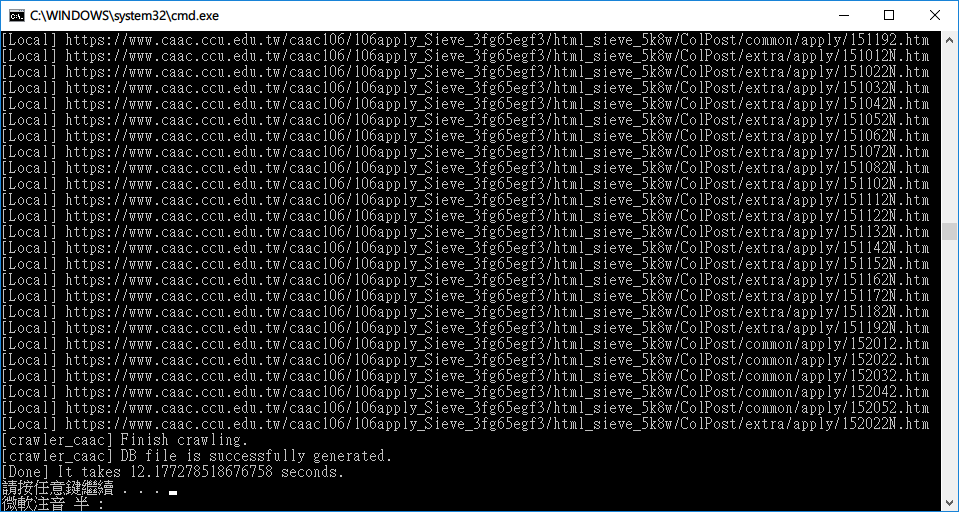
1. 第二階段-分發結果（甄選委員會）

**※備註：通常，只有當 第二階段-交叉查榜 無法使用時（像是來源網站已經倒閉） 才需要使用此方式。此方式為從「甄選委員會」的官方網站中抓取資料，因此僅能查詢「統一分發結果」。**

1. 修改 do\_crawl.bat 中的參數設定
2. 對 do\_crawl.bat 按 右鍵→編輯
3. 打開網頁瀏覽器，找到當年度 **統一分發結果** 的網頁



1. 複製該網頁的網址，貼上到 do\_ crawl.bat 中的 **projectBaseUrl**
2. 儲存並關閉檔案
3. 抓取網站內容並建立本地資料庫（只需進行一次）
4. 雙擊 do\_crawl.bat
5. 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2至5分鐘。

1. 如果想要**完全**重新抓取網站內容，可以先刪除 **data/crawler\_XXX** 資料夾
2. 將要查詢的准考證號寫入 admission\_ids.txt 中，然後儲存並關閉檔案

|  |  |
| --- | --- |
|  | 如左圖，一行寫一個准考證號。  圖中範例為108年清大電機甲、乙組考生的准考證號。 |

1. 從資料庫取出資料
2. 雙擊 do\_lookup.bat
3. result\_xxx.xlsx 即為結果

其內容以准考證號碼排序，若一名考生出現在多個校系代碼的結果中，在輸出的檔案裡只會出現一次。

