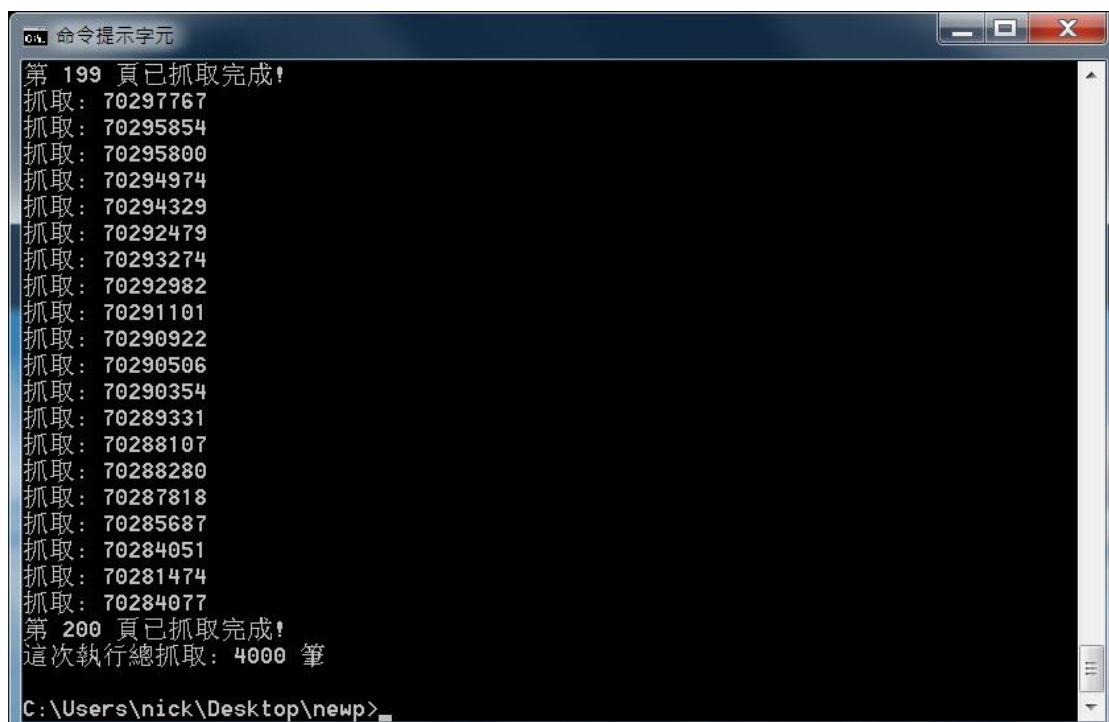


## 爬蟲程式說明

2018/01/29 姜博文

- 使用 python (3.5.2) + [Beautiful Soup](#) (bs4)
- 抓取 ask120 心理健康科的**抑鬱問答**
- 爬蟲程式會**輸出 2 個檔案 askyiyu.csv 和 doctors.csv**
- askyiyu.csv 記錄每一篇提問  
‘問題編號’，‘問題連結’，‘提問人性別’，‘年齡’，‘標題’，‘內容’，‘回覆數(醫生)’，‘醫生們的回覆內容’
- doctors.csv 記錄每一篇提問有哪些醫生回覆以及醫生個人資訊 link  
'問題編號','問題連結','醫生','醫生個人資訊 link'
- 程式每次執行最多可以抓取 200 頁共 4000 筆 (建議凌晨時段再大量抓取)，目前預設是第 1 到第 5 頁(可調整)。

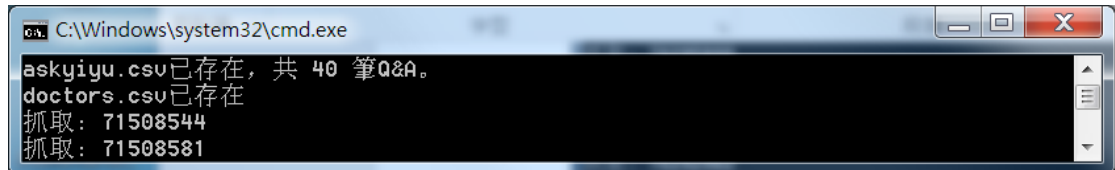


```
命令提示字元
第 199 頁已抓取完成!
抓取: 70297767
抓取: 70295854
抓取: 70295800
抓取: 70294974
抓取: 70294329
抓取: 70292479
抓取: 70293274
抓取: 70292982
抓取: 70291101
抓取: 70290922
抓取: 70290506
抓取: 70290354
抓取: 70289331
抓取: 70288107
抓取: 70288280
抓取: 70287818
抓取: 70285687
抓取: 70284051
抓取: 70281474
抓取: 70284077
第 200 頁已抓取完成!
這次執行總抓取: 4000 筆
C:\Users\nick\Desktop\newp>
```

- 可以判斷是否有新提問出現，造成重複抓取問題

已存在: 71497776

- 可以重複執行累積資料



```
C:\Windows\system32\cmd.exe
askyiyu.csv已存在, 共 40 筆Q&A。
doctors.csv已存在
抓取: 71508544
抓取: 71508581
```

- 每篇提問若有新的醫生回覆也會更新

已存在,有新回覆: 71497199  
抓取: 71497199

- 可以判斷舊資料已存在的提問且沒有新的醫生回覆

已存在: 71497776