程式語言 HW2_Regular Expression with Python 作業說明

資訊三乙 陳華嚴 F04056154

1. 執行環境：Windows 的 Anaconda 3.7 介面
2. 執行步驟：

   python re_python.py

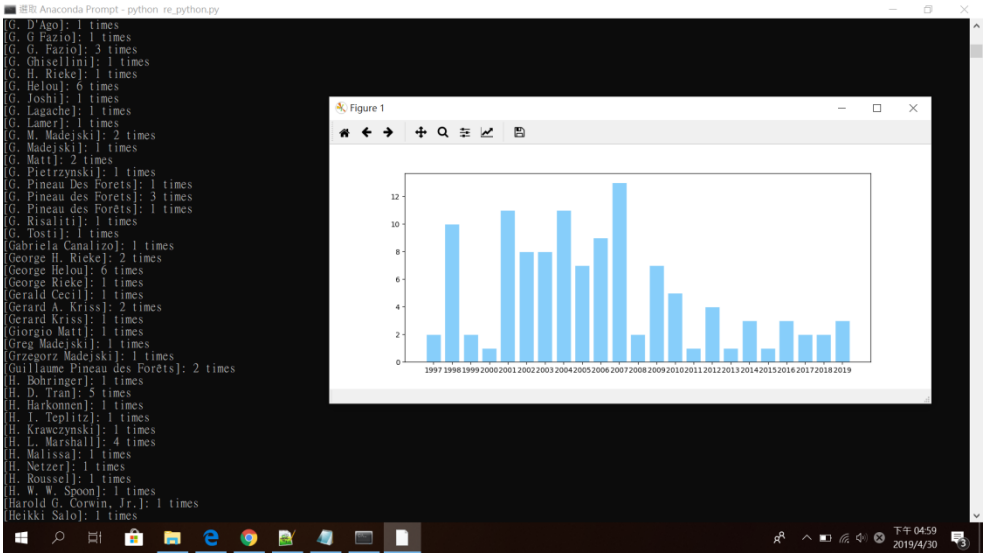
   輸入您要的作者名字

   舉例

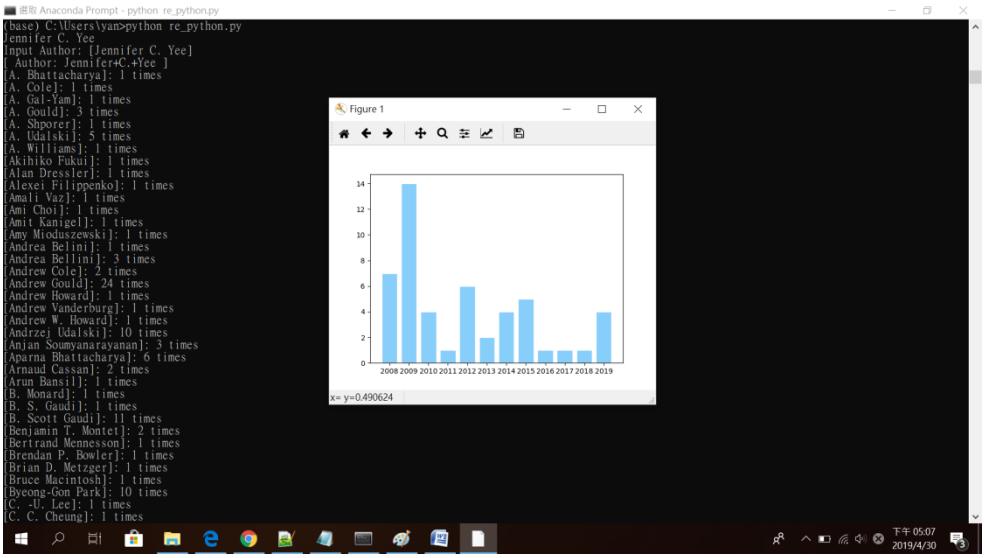
   ```
   python re_python.py
   Jennifer C. Yee
   ```

3. 完成題目：全部
4. 執行結果：

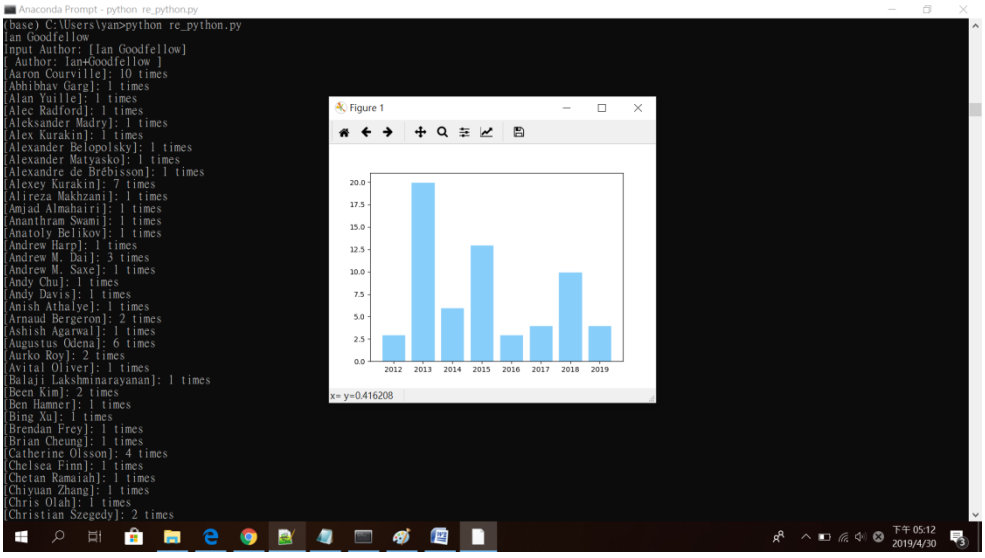
   【輸入 OGLE】(根據網站的搜尋紀錄總共有 116 筆)



【輸入 Jennifer C. Yee】(根據網站的搜尋紀錄總共有 50 筆)



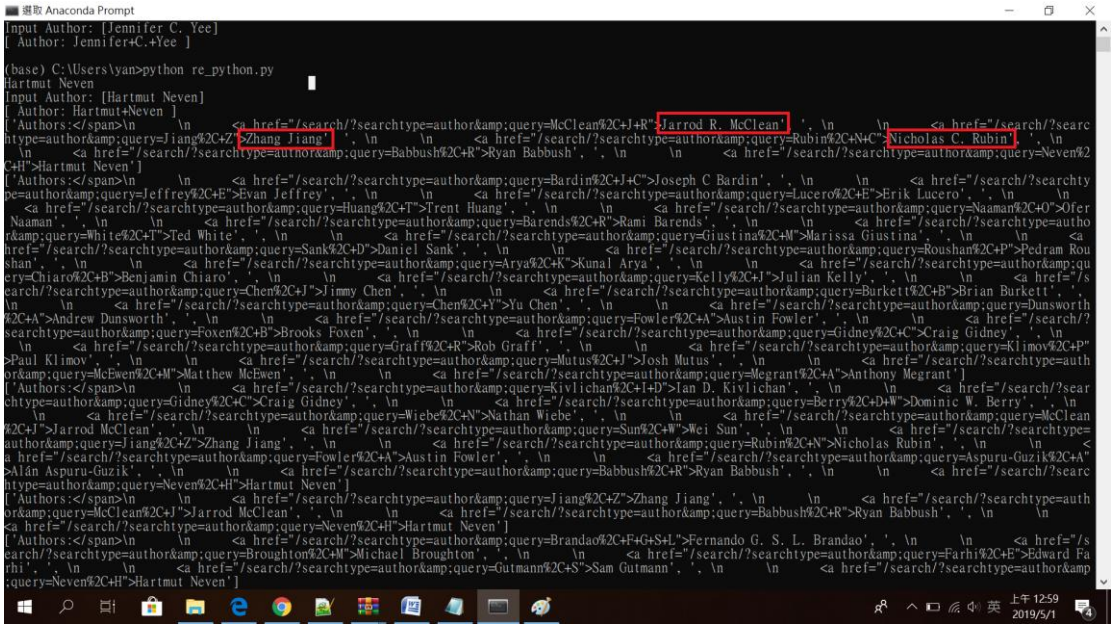
【輸入 Ian Goodfellow】(根據網站的搜尋紀錄總共有 63 筆)



5. 程式碼說明：

特別註明有引入的函式庫，使用 math 是因為我要使用 ceil 這個函式(要換頁的時候會用到)、sys 是因為要用到類似 scanf 功能的 readline()。我分兩個 case 來寫：一個是輸入長度大於二的（以空白分隔），另一個是只有單一文字的 case，會這麼寫的原因是觀察到該網址搜尋人名的方式，若有空白則會用+號連接。之後就要考慮是否有換頁，我用了兩組人名觀察換頁的模式，（在此的換頁考量都是以一頁有 50 筆資料）Hartmut Neven 會有 38 筆資料，也就是說只會有 1 頁；Jie Ma 有 160 筆資料，也就是說 160/50=3....X（有餘數），共會有 4 頁，而其中的差異是，有換頁的原始碼就會跑出 pagination-list，所以我是根據這個判斷的，而我在程式碼中所寫的 nextPage_array 只是為了判斷這個陣列會不會有東東（如果有就代表確實有換頁的數字跑出來）所以程式就可以於此分流，分成「會換頁」以及「只有單頁」的考量。（即 Line:37~75 和 Line:76~131）

會換頁以及只有單頁的共通點就是找到共同的模式去 parse，我是找到 Authors:到
的區間去 parse 作者名稱，再用這個區間去分割</p>之前的東西（都用作為結尾分割）



見以上圖，要得那紅框裡的文字，則需要再次分割，而這次分割的方式是找">之後的東西，再用 strip 去整理多餘的空白（如果有的話）。想特別講我在這邊所遇到的困難：我發現爬蟲因為是爬網頁原始碼，所以像有些特殊字不會是我們一般所見的長相，比如說<'>單引號會長成'的形式，會發現這個是由於我在到處測試的時候看到有個叫做 Peter O'Malley 的人，結果竟然顯示成 Peter O'Malley，所以我開始想要怎麼改掉他，我很明確的想法就是直接用 replace，但是一直失敗！大概隔了好幾個小時才發現「啊…會不會是要用 assign…」才終於正確地顯示出單引號。我覺得 Python 真的是一個很妙的語言，其實這份作業是我人生第二次打 Python，所以很多語法我都不甚熟悉，像 append 就不需要用 assign，所以我對於 replace 有莫名的誤解，想說應該直接用了就會直接取代了呀；再來我又發現一個我寫作業上面的謬誤，在測試的時候我發現有人名中會有<:>冒號，所以我才多加了人名不是冒號的判斷。接下來的年份的 parse 也與上面同理，在這裡想特別說的是感謝同學教我 dictionary，要不然我一直在苦思「array 的人名和數值之間的關係」，在寫作業的時候不斷地覺得 Python 真是太逆天的語言了，dictionary 竟然有鍵值的功能！也因此可以很方便地直接使用 for 迴圈，直接輸入人名及年份進 dictionary 並且計數，再配合 matplot 畫出圖來。我想特別說一下我處理翻頁的方式，我是直接計算最上面的數值（如下圖的紅框處）：

Showing 1–38 of 38 results for author: Hartmut Neven

我觀察到有 50 筆資料就會多一個分頁，所以計算方式需要搭配上高斯，以上面來說，運算式為(38+1)/50 取上高斯，所以結果為 1；若今天有 50 筆資料，

則為(50+1)/50 取上高斯，結果為 2。值得注意的是這種情形：

因為由 Regular Expression 再切完的東西是字串，所以 5,561 的那個逗點也會被當成字串存下來，所以我用 split(',')的方式去把它切成一塊一塊的部分（所以上述例子）會變成['5','561']，再用 join 的方式組合成一個完整的數值（所以會變成'5561'），當然在要使用的時候轉型成 int 就可以了，因為 Python 真的是好好用且好逆天的語言！然後我觀察到這個網站換頁的方式就是直接載網址末尾加上&start=50、&start=100、&start=150…以此類推，其實沒有換頁時在網址末尾加上&start=0 也是可以連接的，於是我才寫出了 append(url+"&start="+ str(i*50))這行，為得就是計算出所有的網址模樣。

（不過要特別註明，如果資料量太大，比如說像這樣五千多頁資料，程式還是會跑完，但是就顯現不出結果了…也依循過助教的建議檢查過配置了，應該是都沒有問題，上網查了 dictionary 及 list 是有最大值的，但是我在這裡不是很確定如果有五千多頁資料會不會到達那個最大值。但是如果幾頁以下應該是都可以呈現出正確結果的！）

最後要說明的是，一開始定義的 removekey(這其實是我上網找的函式寫法)，功用是為了拔除自己在 dictionary 的名字，因為到時候在 coauthor 都是不計算且不列表呈現的。