Programming Language HW2 Regular Expression with Python Report

資訊109 F74054067 柳孟芸

▷執行環境：

VS Code 作為Python 的開發環境

▷程式碼解說：

hw2.py

```python
import urllib.request
import re
import matplotlib.pyplot as plt
from matplotlib.ticker import FormatStrFormatter

author = "Ian+Goodfellow"
url = "https://arxiv.org/search/?query=" + author + "&searchtype=author"
content = urllib.request.urlopen(url)
html_str = content.read().decode('utf-8')
# number of the paper
num_pattern = 'title is-clearfix[\s\S]*?:'
num_result = re.findall(num_pattern,html_str)
for i in num_result:
    number = i.split("title is-clearfix\">\n    \n        Showing 1&ndash;50 of ")[1].split(" results for author:")[0].strip()
    start = int(int(number)/50)
    x = []
    y = []
    co_author = []
    times = []
    print("Input Author: [" + author + "]")
    for j in range(0,start+1) :
        url2="https://arxiv.org/search/?query="+author+"&searchtype=author&order=-announced_date_first&size=50&abstracts=show&start="+str(j*50)
        content2 = urllib.request.urlopen(url2)
        html_str2 = content2.read().decode('utf-8')
        # announced year
        year_pattern = 'originally announced[\s\S]*?</p>'
        year_result = re.findall(year_pattern,html_str2)
        for k in year_result:
            year = k.split("originally announced</span>")[1].split("\n      \n    </p>")[0].strip()
            x.append(year[-5:-1])
            X = list(set(x))
            X.sort()
        # co-author
        author_pattern = '<span class="search-hit">Authors:</span>[\s\S]*?</p>'
        author_result = re.findall(author_pattern,html_str2)
        for l in author_result:
            href = l.split("<span class=\"search-hit\">Authors:</span>\n      \n      ")[1].split("\n      \n    </p>")[0].strip()
            #獲取超連結<a>和</a>之間內容
            coauthor_pattern = '<a .*?>(.*?)</a>'
            coauthor_texts = re.findall(coauthor_pattern, href, re.S|re.M)
            for t in coauthor_texts:
                co_author.append(t.strip())
                Co_author = list(set(co_author))
                Co_author.sort()

# Question 2
for name in Co_author:
    times = co_author.count(name)
    if name == author.replace('+',' ') :  #去掉自己
        continue
    else:   # print the answer
        print('[' + name + ']: ' + str(times) + ' times' )

# Question 1
for m in X:
    counter = x.count(m)
    y.append(counter)
ax = plt.figure().gca()
ymajorFormatter = FormatStrFormatter('%d') #設置y軸標籤文本的格式
ax.yaxis.set_major_formatter(ymajorFormatter)
plt.ylabel('The number of papers been published')
plt.title("Input Author: [" + author + "]")
plt.bar(X,y)
plt.show()
```

先抓取input author的paper數，
之後用於更改url2規則，
以確保將所有搜尋到的result都整合

抓取所有paper的
originally announced year
並且存入x list裡，
X list則作為barplot中x-axis的label

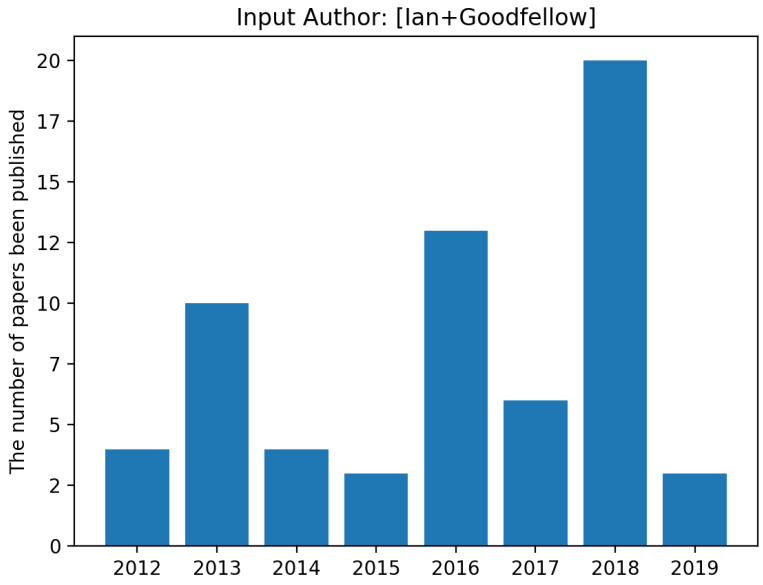
抓取所有paper的co-author
並且存入co_author list裡，
Co_author是be sorted according to alphabet合作作者名字的list

Count the number of papers written together of each co-author，
並且print出Question 2的答案

y list存取在某year中被announced幾次
並且作為barplot中y-axis的label，
最後用plt.bar(X,y)畫出Question 1的bar plot

▷執行結果：

◎Question 1



◎Question 2

```
liumengyunde-MacBook-Pro:hw2 newmileou$ python3 hw2.py
Input Author: [Ian+Goodfellow]
[Aaron Courville]: 10 times
[Abhibhav Garg]: 1 times
[Alan Yuille]: 1 times
[Alec Radford]: 1 times
[Aleksander Madry]: 1 times
[Alex Kurakin]: 1 times
[Alexander Belopolsky]: 1 times
[Alexander Matyasko]: 1 times
[Alexandre de Brébisson]: 1 times
[Alexey Kurakin]: 7 times
[Alireza Makhzani]: 1 times
[Amjad Almahairi]: 1 times
[Ananthram Swami]: 1 times
[Anatoly Belikov]: 1 times
[Andrew Harp]: 1 times
[Andrew M. Dai]: 3 times
[Andrew M. Saxe]: 1 times
[Andy Chu]: 1 times
[Andy Davis]: 1 times
[Anish Athalye]: 1 times
[Arnaud Bergeron]: 2 times
[Ashish Agarwal]: 1 times
[Augustus Odena]: 6 times
[Aurko Roy]: 2 times
[Avital Oliver]: 1 times
[Balaji Lakshminarayanan]: 1 times
[Been Kim]: 2 times
[Ben Hamner]: 1 times
[Bing Xu]: 1 times
[Brendan Frey]: 1 times
[Brian Cheung]: 1 times
[Catherine Olsson]: 4 times
[Chelsea Finn]: 1 times
[Chetan Ramaiah]: 1 times
[Chiyuan Zhang]: 1 times
[Chris Olah]: 1 times
[Christian Szegedy]: 2 times
[Christof Angermueller]: 1 times
[Christopher Olah]: 1 times
[Cihang Xie]: 2 times
[Colin Raffel]: 4 times
[Craig Citro]: 1 times
[Cristian Grozea]: 1 times
[Da Xiao]: 1 times
[Dan Boneh]: 2 times
[Dan Mane]: 1 times
[David Andersen]: 1 times
[David Berthelot]: 2 times
[David Thaler]: 1 times
[David Warde-Farley]: 5 times
[Derek Murray]: 1 times
[Dimitris Athanasakis]: 1 times
[Dimitris Metaxas]: 1 times
[Dimitris Tsipras]: 1 times
[Dong-Hyun Lee]: 1 times
[Dumitru Erhan]: 2 times
[Dzmitry Bahdanau]: 1 times
[Ekin D. Cubuk]: 1 times
[Eugene Brevdo]: 1 times
[Fangxiang Feng]: 1 times
[Fangzhou Liao]: 1 times
[Fartash Faghri]: 2 times
[Florian Tramèr]: 2 times
[Frédéric Bastien]: 3 times
[Gamaleldin F. Elsayed]: 2 times
[Garrison Cottrell]: 1 times
[Geoffrey Irving]: 1 times
[George E. Dahl]: 1 times
[Greg S. Corrado]: 1 times
[Guillaume Alain]: 1 times
[H. Brendan McMahan]: 2 times
[Han Zhang]: 1 times
[Harini Kannan]: 1 times
[Ian J. Goodfellow]: 15 times
[Ilya Mironov]: 2 times
[Ilya Sutskever]: 1 times
[Jacob Buckman]: 1 times
[James Bergstra]: 4 times
[Jan Chorowski]: 1 times
[Jascha Sohl-Dickstein]: 2 times
[Jean Pouget-Abadie]: 1 times
[Jeffrey Dean]: 1 times
[Jianyu Wang]: 1 times
[Jingjing Xie]: 1 times
[Joan Bruna]: 1 times
[John Park]: 1 times
[John Shawe-Taylor]: 1 times
[Jonas Rauber]: 2 times
[Jonathan Uesato]: 1 times
[Jonathon Shlens]: 3 times
[Josh Bleecher Snyder]: 1 times
[Josh Levenberg]: 1 times
[Julian Ibarz]: 1 times
[Julius Adebayo]: 2 times
[Jun Zhu]: 1 times
[Junjiajia Long]: 1 times
[Justin Bayer]: 1 times
[Justin Gilmer]: 4 times
[Karen Hambardzumyan]: 1 times
[Kunal Talwar]: 3 times
[Kyunghyun Cho]: 1 times
[Li Zhang]: 2 times
[Lukasz Kaiser]: 1 times
[Lukasz Romaszko]: 1 times
[Luke Metz]: 1 times
[Maithra Raghu]: 1 times
[Manjunath Kudlur]: 1 times
[Marius Popescu]: 1 times
[Martin Wattenberg]: 1 times
[Martín Abadi]: 4 times
[Matthieu Devin]: 1 times
[Maxim Milakov]: 1 times
[Mehdi Mirza]: 5 times
[Michael Isard]: 1 times
[Michael Muelly]: 1 times
[Mihaela Rosca]: 1 times
[Ming Liang]: 1 times
[Moritz Hardt]: 1 times
[Motoki Abe]: 1 times
[Navdeep Jaitly]: 1 times
[Nicholas Carlini]: 4 times
[Nicolas Ballas]: 1 times
[Nicolas Bouchard]: 2 times
[Nicolas Boulanger-Lewandowski]: 1 times
[Nicolas Papernot]: 10 times
[Olivier Breuleux]: 1 times
[Oriol Vinyals]: 1 times
[Pascal Lamblin]: 2 times
[Patrick McDaniel]: 4 times
[Paul Barham]: 1 times
[Paul Christiano]: 2 times
[Paul Hendricks]: 1 times
[Pierre Luc Carrier]: 1 times
[Pierre-Luc Carrier]: 1 times
[Pieter Abbeel]: 1 times
[Radu Ionescu]: 1 times
[Rafal Jozefowicz]: 1 times
[Rajat Monga]: 1 times
[Rami Al-Rfou]: 1 times
[Razvan Pascanu]: 2 times
[Reuben Feinman]: 1 times
[Rob Fergus]: 1 times
[Ruifan Li]: 1 times
[Rujun Long]: 1 times
[Ryan P. Adams]: 1 times
[Ryan Sheatsley]: 1 times
[Sacha Arnoud]: 1 times
[Samaneh Azadi]: 1 times
[Samuel S. Schoenholz]: 1 times
[Samy Bengio]: 3 times
[Sandy Huang]: 1 times
[Sangxia Huang]: 1 times
[Sanjay Ghemawat]: 1 times
[Seiya Tokui]: 1 times
[Sergey Levine]: 1 times
[Shakir Mohamed]: 1 times
[Sherjil Ozair]: 1 times
[Sherry Moore]: 1 times
[Shreya Shankar]: 1 times
[Somesh Jha]: 1 times
[Stephan Zheng]: 1 times
[Surya Bhupatiraju]: 1 times
[Takeru Miyato]: 1 times
[Takuya Akiba]: 1 times
[The Theano Development Team]: 1 times
[Thomas Leung]: 1 times
[Tianqi Chen]: 1 times
[Tianyu Pang]: 1 times
[Tim Salimans]: 1 times
[Tom B. Brown]: 2 times
[Tom Brown]: 2 times
[Trevor Darrell]: 1 times
[Vahid Behzadan]: 1 times
[Valentin Bisson]: 1 times
[Vicki Cheung]: 1 times
[Vinay Shet]: 1 times
[Vincent Dumoulin]: 2 times
[Wieland Brendel]: 1 times
[Will Cukierski]: 1 times
[Willi Gierke]: 1 times
[William Fedus]: 2 times
[Wojciech Zaremba]: 2 times
[Xavier Bouthillier]: 1 times
[Xi Chen]: 1 times
[Xiaojie Wang]: 1 times
[Xiaolin Hu]: 1 times
[Yan Duan]: 1 times
[Yang Song]: 1 times
[Yangqing Jia]: 1 times
[Yao Qin]: 1 times
[Yao Zhao]: 1 times
[Yaroslav Bulatov]: 1 times
[Yash Sharma]: 1 times
[Yerkebulan Berdibekov]: 1 times
[Yi-Lin Juang]: 1 times
[Yichuan Tang]: 1 times
[Yingbo Zhou]: 1 times
[Yinpeng Dong]: 2 times
[Yoshua Bengio]: 12 times
[Yuzhe Zhao]: 1 times
[Z. Berkay Celik]: 1 times
[Zhi Li]: 1 times
[Zhifeng Chen]: 1 times
[Zhishuai Zhang]: 2 times
[Zhonglin Han]: 1 times
[Zhou Ren]: 1 times
[Úlfar Erlingsson]: 2 times
```