เรื่อง Google Scholar

จัดทำโดย

นาย ณัฐพัชร อนุโรจน์

6209650016

เสนอ

อาจารย์ วสิศ ลิ้มประเสริฐ

รายงานนี้เป็นส่วนหนึ่งของวิชา DSI200

ภาคเรียนที่ 2 ปีการศึกษา 2563

มหาวิทยาลัยธรรมศาสตร์(ศูนย์รังสิต)

Google Scholar Assignment

1) อธิบาย source code และ Out put

1.1) Google Scholar author

ขั้นตอนแรก import library ต่างๆที่จำเป็นต้องใช่และกำหนดค่า webdriver ไปที่ chrome

```
import pandas as pd
from selenium import webdriver
from selenium.webdriver.common.by import By
driver = webdriver.Chrome('C:\\Users\\mark\\Desktop\\dsi200\\chromedriver.exe')
```

ขั้นตอนที่สอง เปิด google scholar และส่งคำว่า Thammasat University

```
driver.get('https://scholar.google.com/')
```

```
#frist page
search_box = driver.find_element_by_css_selector('#gs_hdr_tsi')
search_box.send_keys("Thammasat University\n")
```

ขั้นตอนที่สาม สร้าง Data frame

```
df = pd.DataFrame(
    {
        'user_id',
        'author',
        'affliation'
    }
)
```

ขั้นตอนที่สี่ให้วนลูปแต่ละคนเพื่อเอาข้อมูลมาใส่ใน Data frame

```
for i in range(30):
    for i in driver.find_elements(By.CSS_SELECTOR, "div.gs_ai_t"):
        a = i.find_element_by_css_selector('a')
        author = a.text
        user_id = a.get_attribute('href').split('=')[-1]
        affliation = i.find_element_by_css_selector('div.gs_ai_aff').text
        #print(user_id, author, affliation)
        df = df.append(
            {
                'user_id':user_id,
                'author':author,
                'affliation':affliation
            }, ignore_index=True
        )

    next_page = driver.find_element_by_css_selector('#gsc_authors_bottom_pag > div > button.gs_btnPR.gs
    next_page.click()
```

Out put ที่ได้

|:

	user_id	author	affliation
1	Re819VUAAAAJ	sandhya Babel	Professor, Sirindhorn International Institute ...
2	97Zz_TIAAAAJ	Sombat Muengtaweepongsa	Thammasat University
3	u-vl_aIAAAAJ	Wutiphol Sintunavarat	Department of Mathematics and Statistics, Facu...
4	9bUei6wAAAAJ	Bunyarit Uyyanonvara	Associated Professor, SIIT, Thammasat University
5	UOeuXvQAAAAJ	Chanathip Namprempre	Assistant Professor, Thammasat University
...	...	...	...
294	NLshr9gAAAAJ	Ines Ayu Handayani	Master's Student, SIIT, Thammasat University
295	oALsv34AAAAJ	Kritsasith Warin	Lecturer, Faculty of Dentistry, Thammasat Univ...
296	kJaNuNMAAAAJ	Dulyaphab Chaturongkul	Lecturer in Political Science, Thammasat Unive...
297	7XmRxVUAAAAJ	Punnavich Khowrurk	Department of Computer Science,Thammasat Unive...
298	vvP6YwkAAAAJ	Adipon Euajarusphan	Lecturer , Thammasat University

298 rows × 3 columns

1.2) Google Scholar paper

ขั้นตอนแรก import library ต่างๆที่จำเป็นต้องใช่และกำหนดค่า webdriver ไปที่ chrome

```
import pandas as pd
import re
import time
import requests as req
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import StaleElementReferenceException
from bs4 import BeautifulSoup
```

```
driver = webdriver.Chrome('C:\\Users\\mark\\Desktop\\dsi200\\chromedriver.exe')
```

ขั้นตอนที่สอง เปิด google scholar สร้าง Data frame และส่งคำว่า Thammasat University

```
driver.get('https://scholar.google.com/')
#pd.set_option('display.max_columns', None)
#pd.set_option('display.max_rows', None)
df = pd.DataFrame()
#driver.set_page_load_timeout(30)
```

```
search_box = driver.find_element_by_css_selector('#gs_hdr_tsi')
search_box.send_keys("Thammasat University\n")
```

ขั้นตอนที่สี่ สร้างลูปสำหรับไว้ใช้เก็บค่า link ของแต่ละ users

```
wait = WebDriverWait(driver, 5)
users = []
line = 3
for person in range(30):
    request = req.get(driver.current_url)
    parser = BeautifulSoup(request.content, 'html.parser')
    gsc = parser.find_all('div', {'class': 'gsc_1usr'})
    for i in gsc:
        #print("1")
        per = i.find('a')['href']
        sp = str(per).split('/citations?hl=th&user=')[1]
        print(sp)
        users.append(sp)
    driver.find_element_by_xpath('//*[@id="gsc_authors_bottom_pag"]/div/button[2]').click()
```

ขั้นตอนที่ห้า สร้างลูปเข้าไปที่ลิงค์แต่ละ users และให้กดปุ่มแสดงบทความจนกดไม่ได้

```
]: for i in users:
       url = 'https://scholar.google.com/citations?hl=th&user=' + i
       driver.get(url)
       t = 0
       while(t == 0) :
           try:
               wait.until(EC.element_to_be_clickable((By.ID, "gsc_bpf_more")))
           except Exception:
               #print("ex")
               break
           finally :
               element = driver.find_element_by_id("gsc_bpf_more")
               b = element.is_enabled()
               #print(b)
               element.click()
```

ขั้นตอนที่ห้า ให้วนลูปคลิกเข้าไปเก็บข้อมูลในแต่ละบทความ

```
for article in driver.find_elements(By.CLASS_NAME, 'gsc_a_tr'):
    time.sleep(2)
    h = article.find_element_by_class_name('gsc_a_at')
    h.click()
   # wait = WebDriverWait(driver, 2)
    while (1):
        try:
            wait.until(EC.invisibility_of_element_located((By.ID, "gsc_vcd_table")))
        except:
            print("Error time out 1")
            continue
        finally:
            break
    while (1):
        try:
            wait.until(EC.visibility_of_element_located((By.ID, "gsc_vcd_table")))
        except:
            print("Error time out 2")
            continue
        finally:
            break
    header = driver.find_element_by_id('gsc_vcd_title')
    try:
        title = header.find_element_by_xpath('//*[@id="gsc_vcd_title"]/a').text
        table = driver.find_element_by_id('gsc_vcd_table')
        all_children_by_gsl = table.find_elements_by_class_name('gs_scl')
        for paragraph in all_children_by_gsl:
            sen = paragraph.find_element_by_class_name('gsc_vcd_value')
            fila = paragraph.find_element_by_class_name('gsc_vcd_field')

            fil = fila.text
            if fil == "ผู้เขียน":
                authors = sen.text
                print(authors)
            if fil == "วันที่เผยแพร่":
                date = sen.text
                print(date)
```

```
            if fil == "คำอธิบาย":
                descr = sen.text
                print(descr)
            if fil == "การอ้างอิงทั้งหมด":
                c = a.get_attribute("style")
                cite = sen.find_element_by_tag_name('a').text
                ci = cite[7:]
    except:
        title = driver.find_element_by_id('gsc_vcd_title').text
        table = driver.find_element_by_id('gsc_vcd_table')
        all_children_by_gsl = table.find_elements_by_class_name('gs_scl')
        for paragraph in all_children_by_gsl:
            sen = paragraph.find_element_by_class_name('gsc_vcd_value')
            fila = paragraph.find_element_by_class_name('gsc_vcd_field')

            fil = fila.text
            if fil == "ผู้เขียน":
                authors = sen.text
                print(authors)
            if fil == "วันที่เผยแพร่":
                date = sen.text
                print(date)
            if fil == "คำอธิบาย":
                descr = sen.text
                print(descr)
            if fil == "การอ้างอิงทั้งหมด":
                #c = a.get_attribute("style")
                cite = sen.find_element_by_tag_name('a').text
                ci = cite[7:]
                print(ci)
    df = df.append(
        {
            'title':title,
            'authors':authors,
            'publication_date':date,
            'description':descr,
            'cite_by':ci
        }, ignore_index=True
```

Out put ที่ได้

Out[16]:

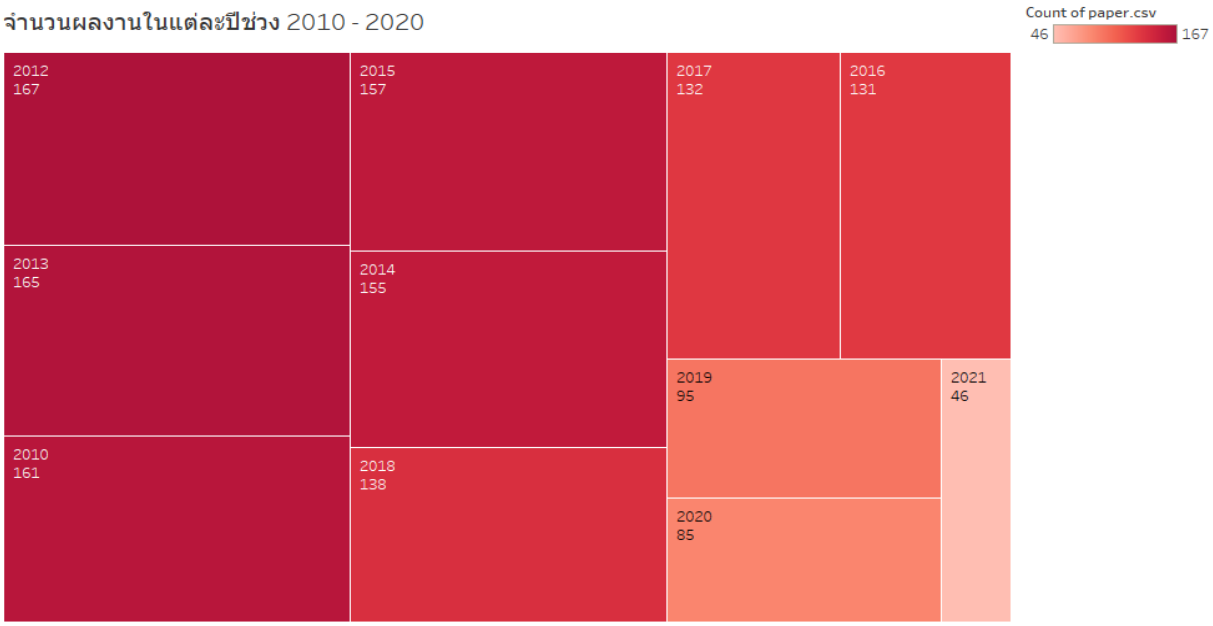
	title	authors	publication_date	description	cite_by
0	Low-cost adsorbents for heavy metals uptake fr...	Sandhya Babel, Tonni Agustiono Kurniawan	2003/2/28	In this article, the technical feasibility of ...	4013
1	Physico–chemical treatment techniques for wast...	Tonni Agustiono Kurniawan, Gilbert YS Chan, Wa...	2006/5/1	This article reviews the technical applicabili...	1973
2	Cr (VI) removal from synthetic wastewater usin...	Sandhya Babel, Tonni Agustiono Kurniawan	2004/2/1	In this study, the technical feasibility of co...	948
3	Comparisons of low-cost adsorbents for treatin...	Tonni Agustiono Kurniawan, Gilbert YS Chan, Wa...	2006/8/1	In this article, the removal performance and c...	804
4	Equilibrium and kinetic modelling of biosorpti...	Zümriye Aksu, Sevilay Tezer	2000/12/1	The biosorption of Remazol Black B, a vinyl su...	584
5	Heavy metal removal from contaminated sludge f...	Sandhya Babel, Dominica del Mundo Dacera	2006/1/1	In recent years, various methods for heavy met...	340
6	Microfiltration membrane fouling and cake beha...	Sandhya Babel, Satoshi Takizawa	2010/10/15	This study was carried out to investigate the ...	169
7	A matrix in life cycle perspective for selecti...	UG Yasantha Abeysundara, Sandhya Babel, Shabbi...	2009/5/1	This paper presents a matrix to select sustain...	152

2.) Data Visualization

2.1) จำนวนผลงานในแต่ละปีช่วง 2010 – 2020

จากกราฟที่เห็นด้านล่างนั้นจะเป็นการแสดงข้อมูลเกี่ยวกับจำนวนผลงานในแต่ละปีในช่างปี 2010 – 2020 โดยจัดทำ show ออกมาในรูปแบบ tree map เพื่อที่จะนำมาเปรียบเทียบในแต่ละปี

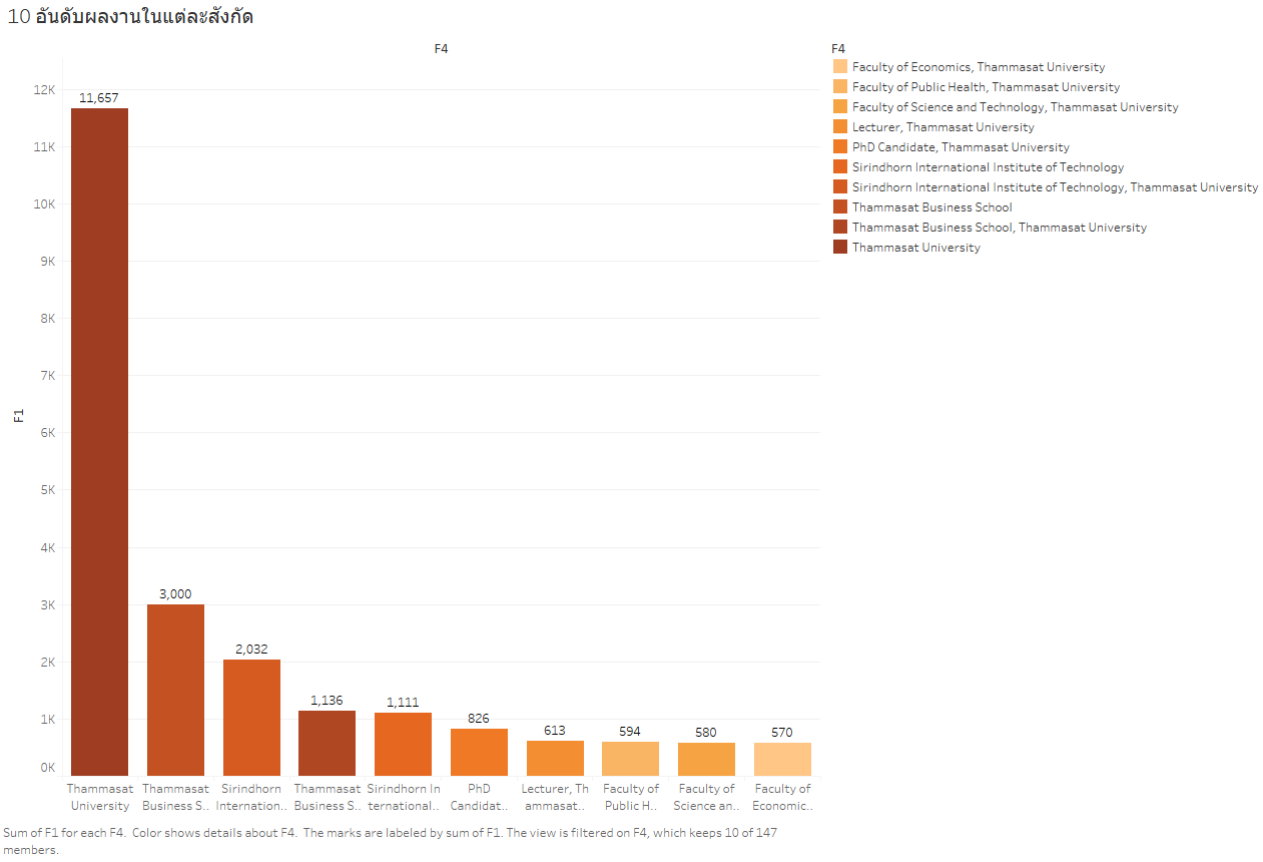
จากกราฟจะเห็นได้ว่าในปี 2012 นั้นเป็นช่วงเวลาที่มีจำนวนผลงานเยอะที่สุดถึง 167 ผลงาน ลองลงมา ก็จะเป็นในปี 2013 ที่ทำไป 165 ผลงาน ซึ่งในบางผลงานนั้นไม่ได้มีการบอกปีเอาไว้หรือบอกแต่ไม่ชัดเจนก็เลยเอาส่วนที่บอกข้อมูลแค่บอกเป็นปีมาอย่างเดียวไม่ได้เอาข้อมูลที่บอกเป็นแบบปีเดือนวัน



2.2) 10 อันดับจำนวนผลงานในแต่ละสังกัด

จากกราฟที่เห็นด้านล่างนั้นจะเป็นการแสดงข้อมูลเกี่ยวกับ10 อันดับจำนวนผลงานในแต่ละสังกัด โดยจัดทำ show ออกมาในรูปแบบของกราฟแท่งเพื่อที่จะนำมาเปรียบเทียบในแต่ละสังกัด เพื่อที่จะสามารถดูได้ง่าย

จากกราฟจะเห็นได้ว่า Thammasat University มีผลงานถึง 11657 ผลงานลองลงมาจะเป็น Thammasat Business School ซึ่งมีผลงานถึง 3000 ผลงาน จะได้ว่าในบางผลงานนั้นไม่ได้มีการบอกสังกัดเอาไว้หรือ.ในหนึ่ง users อาจมีได้หลายสังกัด



ใ