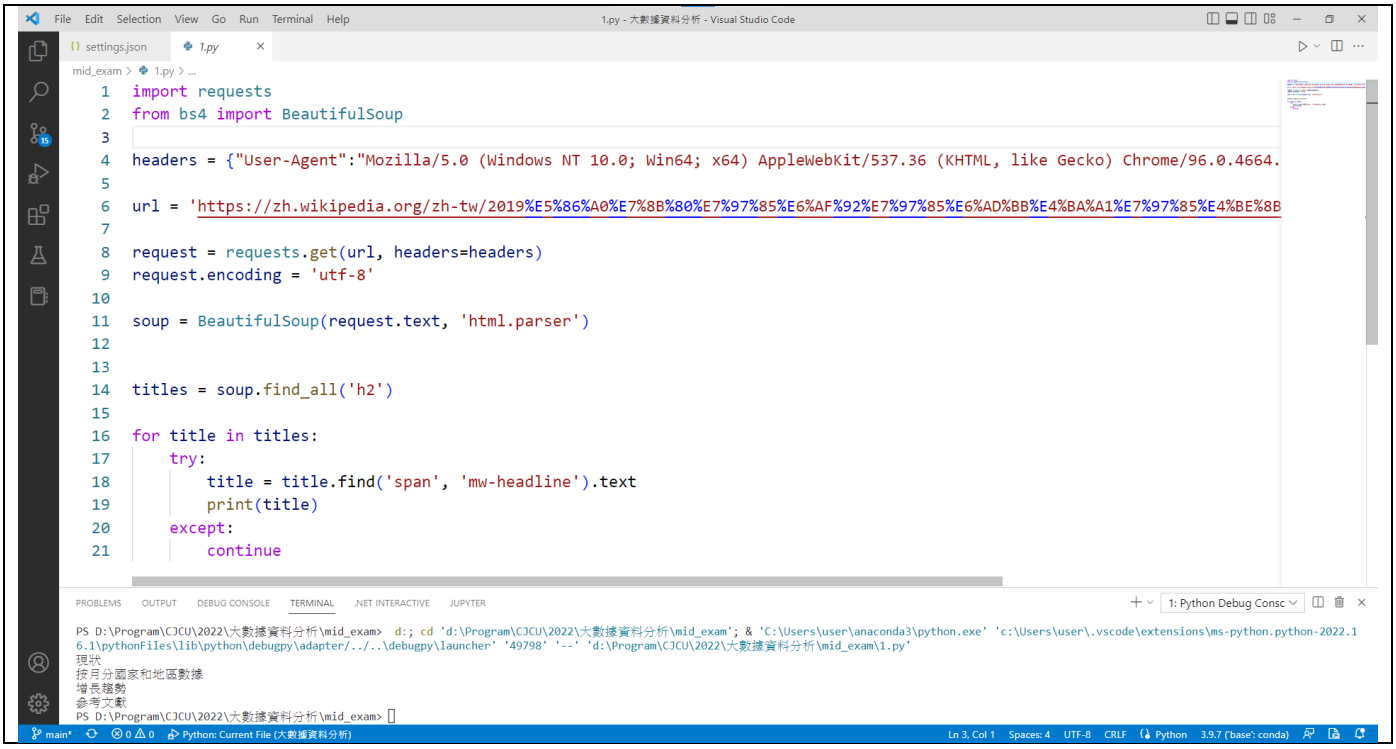
長榮大學 111 學年度第 1 學期 期中考試卷 Chang Jung Christian University Examination Sheet for Academic Year: 111 Semester: 1					
院系 College	資訊暨設計學院 資工系 3 年 B 班 College Department Year Class	姓名 Name	郭智榮	學號 Student No.	109B30612
科目名稱 Subject Name	大數據資料分析	教師簽章	黃琨義	評閱成績 Score	

1. 請撰寫程式於以下網頁中抓取標題現狀、按月分國家和地區數據、增長趨勢、參考文獻並逐行輸出於 Console 介面（25%）
https://zh.wikipedia.org/zh-tw/2019%E5%86%A0%E7%8B%80%E7%97%85%E6%AF%92%E7%97%85%E6%AD%BB%E4%BA%A1%E7%97%85%E4%BE%8B%E6%95%B8

執行結果(請截圖貼上)：



程式碼(請將完整程式碼貼入下方表格，請貼文字，勿貼圖，否則不予計分)：

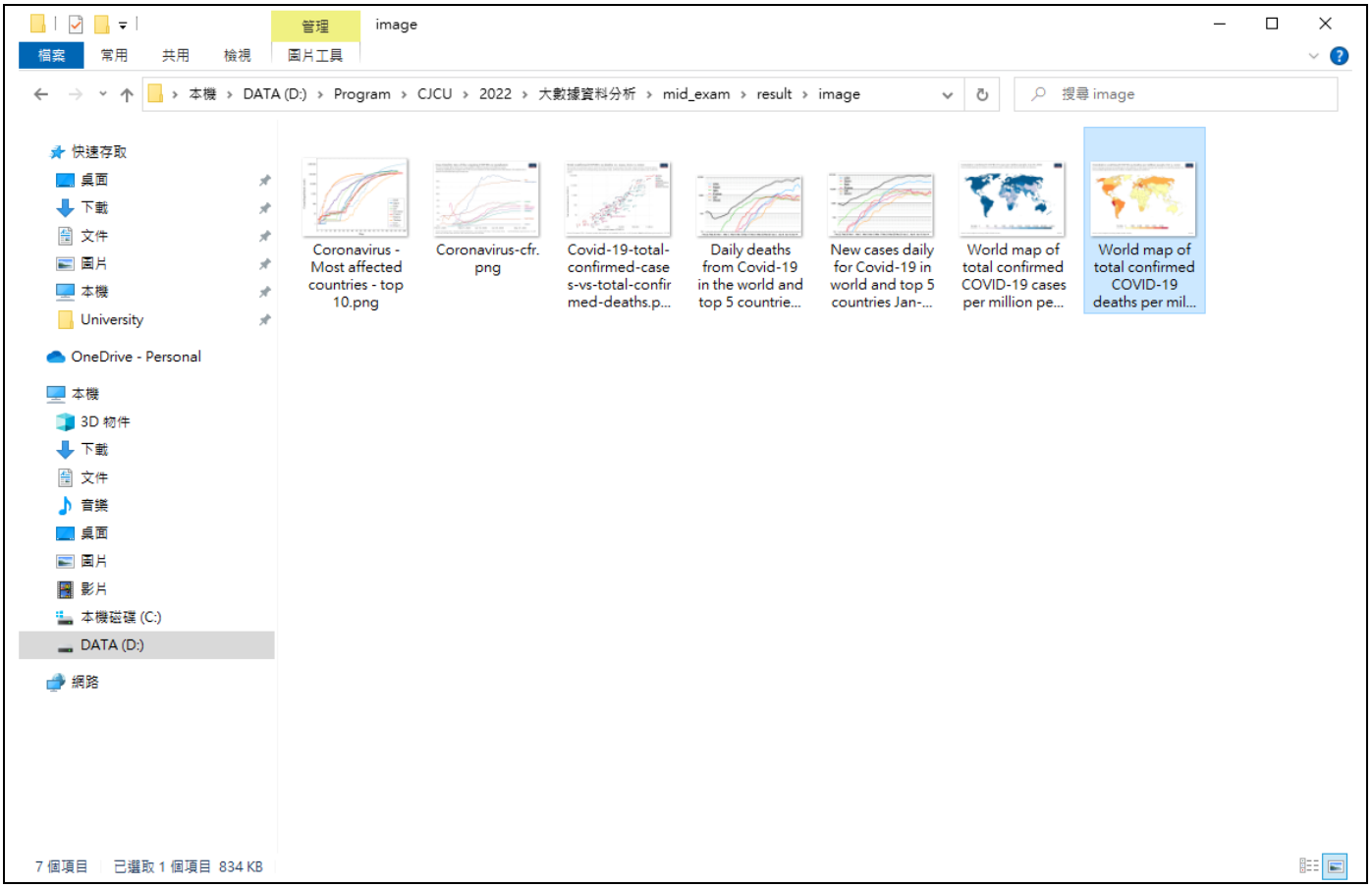
長榮大學 111 學年度第 1 學期 期中考試卷 Chang Jung Christian University Examination Sheet for Academic Year: 111 Semester: 1					
院系 College	資訊暨設計學院 資工系 3 年 B 班 College Department Year Class	姓名 Name	郭智榮	學號 Student No.	109B30612
科目名稱 Subject Name	大數據資料分析	教師簽章	黃琨義	評閱成績 Score	

```
import requests
from bs4 import BeautifulSoup

headers = {"User-Agent":"Mozilla/5.0 (Windows NT 10.0; Win64;
x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/96.0.4664.110 Safari/537.36 Edg/96.0.1054.57"}

url = 'https://zh.wikipedia.org/zh-
tw/2019%E5%86%A0%E7%8B%80%E7%97%85%E6%AF%92%E7%97%85%E6%AD%BB
%E4%BA%A1%E7%97%85%E4%BE%8B%E6%95%B8'

request = requests.get(url, headers=headers)
request.encoding = 'utf-8'

soup = BeautifulSoup(request.text, 'html.parser')


titles = soup.find_all('h2')

for title in titles:
    try:
        title = title.find('span', 'mw-headline').text
        print(title)
    except:
        continue
```

長榮大學 111 學年度第 1 學期 期中考試卷 Chang Jung Christian University Examination Sheet for Academic Year: 111 Semester: 1					
院系 College	資訊暨設計學院 資工系 3 年 B 班 College Department Year Class	姓名 Name	郭智榮	學號 Student No.	109B30612
科目名稱 Subject Name	大數據資料分析	教師簽章	黃琨義	評閱成績 Score	

2. 請撰寫程式於以下網頁中抓取如圖表區塊中所有圖片(需找到最原始圖片)並下載於電腦中，請略過 svg 圖 (25%)
https://zh.wikipedia.org/zh-tw/2019%E5%86%A0%E7%8B%80%E7%97%85%E6%AF%92%E7%97%85%E5%85%A8%E7%90%83%E5%90%84%E5%9C%B0%E7%96%AB%E6%83%85

執行結果(請截圖貼上)：



程式碼(請將完整程式碼貼入下方表格，請貼文字，勿貼圖，否則不予計分)：

長榮大學 111 學年度第 1 學期 期中考試卷 Chang Jung Christian University Examination Sheet for Academic Year: 111 Semester: 1					
院系 College	資訊暨設計學院 資工系 3 年 B 班 College Department Year Class	姓名 Name	郭智榮	學號 Student No.	109B30612
科目名稱 Subject Name	大數據資料分析	教師簽章	黃琨義	評閱成績 Score	

```
from click import style
import requests, csv
from bs4 import BeautifulSoup

headers = {"User-Agent":"Mozilla/5.0 (Windows NT 10.0; Win64;
x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/96.0.4664.110 Safari/537.36 Edg/96.0.1054.57"}

url =
'https://zh.wikipedia.org/wiki/2019%E5%86%A0%E7%8B%80%E7%97%8
5%E6%AF%92%E7%97%85%E5%85%A8%E7%90%83%E5%90%84%E5%9C%B0%E7%96
%AB%E6%83%85'

request = requests.get(url, headers=headers)
request.encoding = 'utf-8'

soup = BeautifulSoup(request.text, 'html.parser')

img_div = soup.find_all('div', style='margin:0px auto;')

for i, div in enumerate(img_div):
    link = div.find('a')['href']
    if (link.split('.')[-1] != 'svg'):
        url = f'https://zh.wikipedia.org/{link}'

        request = requests.get(url, headers=headers)
        request.encoding = 'utf-8'
        soup = BeautifulSoup(request.text, 'html.parser')

        file_name = link.split(':')[-1].replace('_', ' ')

        image_link = soup.find('a', 'internal')['href']

        image = requests.get(f'http:{image_link}')
```

長榮大學 111 學年度第 1 學期 期中考試卷 Chang Jung Christian University Examination Sheet for Academic Year: 111 Semester: 1					
院系 College	資訊暨設計學院 資工系 3 年 B 班 College Department Year Class	姓名 Name	郭智榮	學號 Student No.	109B30612
科目名稱 Subject Name	大數據資料分析	教師簽章	黃琨義	評閱成績 Score	

```
print(f'https://zh.wikipedia.org/{link}')
print(f'http:{image_link}')

with open("result/image/" + file_name, "wb") as file:
    file.write(image.content)
```

長榮大學 111 學年度第 1 學期 期中考試卷 Chang Jung Christian University Examination Sheet for Academic Year: 111 Semester: 1					
院系 College	資訊暨設計學院 資工系 3 年 B 班 College Department Year Class	姓名 Name	郭智榮	學號 Student No.	109B30612
科目名稱 Subject Name	大數據資料分析	教師簽章	黃琨義	評閱成績 Score	

3. 請撰寫程式於以下網頁中抓取如表之區塊中所有文字資料

(a) 將資料以 csv 檔案格式儲存 (10%)

(b) 計算死亡人數 0、死亡人數超過 3000 人以上、死亡人數超過 5000 人以上、死亡人數超過 10000 人以上分別為多少個月份，並逐行輸出於 Console 介面 (15%)

https://zh.wikipedia.org/zh-tw/2019%E5%86%A0%E7%8B%80%E7%97%85%E6%AF%92%E7%97%85%E6%AD%BB%E4%BA%A1%E7%97%85%E4%BE%8B%E6%95%B8

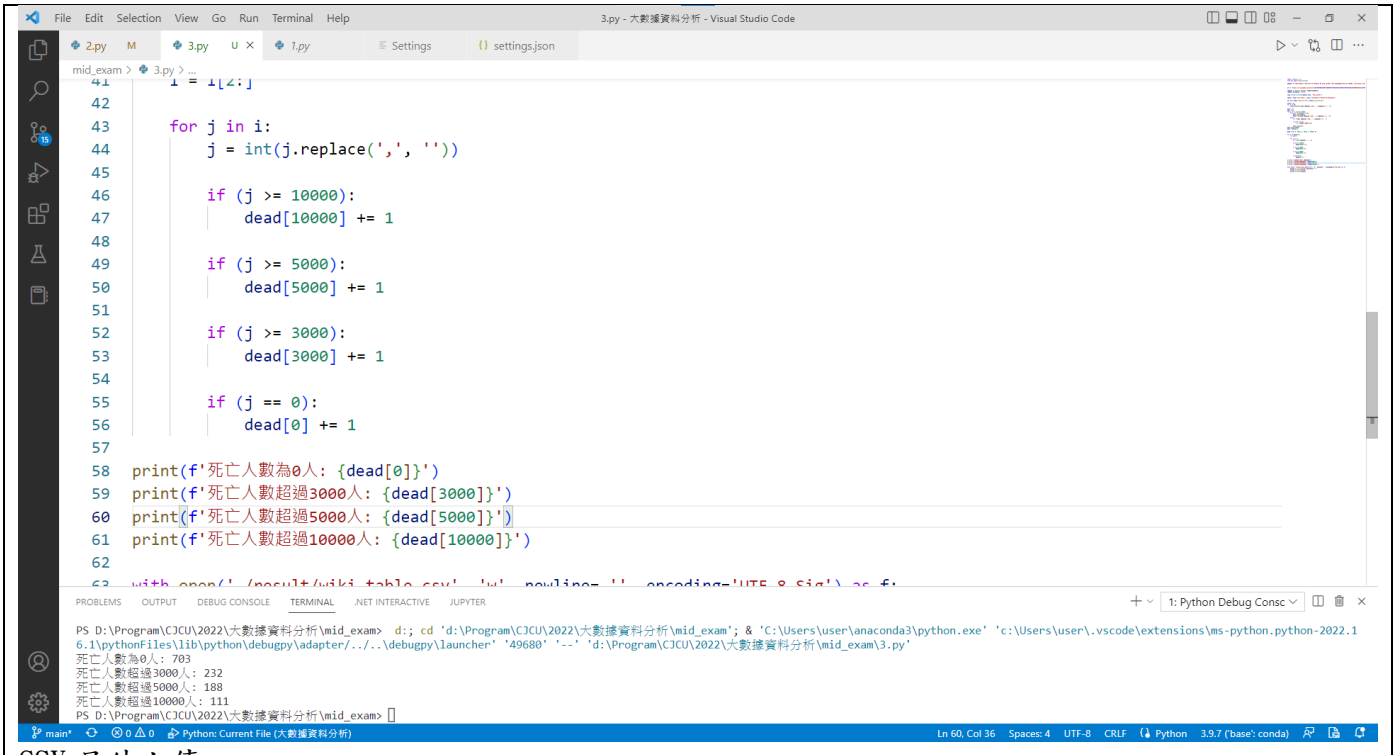
按月分國家和地區數據 [編輯]

2020年 [編輯]

嚴重特殊傳染性肺炎死亡數（2020年1月11日和2020年每月1日）[2]

日期	首例死亡日期	1月11日	2月1日	3月1日	4月1日	5月1日	6月1日	7月1日	8月1日	9月1日	10月1日	11月1日	12月1日
全球		1	259	2,977	40,598	224,172	371,166	508,055	675,060	848,445	1,010,639	1,192,911	1,465,144
死亡數翻倍天數		6	4	16	8	18	37	56	70	80	94	101	110
國家和地區		1	1	8	125	175	185	186	192	191	193	193	193
美國	2020年3月3日	0	0	0	2,850	55,337	102,640	126,573	151,265	182,162	204,642	228,185	264,808
巴西	2020年3月19日	0	0	0	159	5,466	28,834	58,314	91,263	120,828	142,921	159,477	172,833
印度	2020年3月13日	0	0	0	38	1,147	5,394	17,400	36,511	65,288	98,678	122,111	137,621

執行結果(請截圖貼上)：



CSV 另外上傳

程式碼(請將完整程式碼貼入下方表格，請貼文字，勿貼圖，否則不予計分)：

長榮大學 111 學年度第 1 學期 期中考試卷 Chang Jung Christian University Examination Sheet for Academic Year: 111 Semester: 1					
院系 College	資訊暨設計學院 資工系 3 年 B 班 College Department Year Class	姓名 Name	郭智榮	學號 Student No.	109B30612
科目名稱 Subject Name	大數據資料分析	教師簽章	黃琨義	評閱成績 Score	

```
import requests, csv
from bs4 import BeautifulSoup

headers = {"User-Agent":"Mozilla/5.0 (Windows NT 10.0; Win64;
x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/96.0.4664.110 Safari/537.36 Edg/96.0.1054.57"}


url = 'https://zh.wikipedia.org/zh-
tw/2019%E5%86%A0%E7%8B%80%E7%97%85%E6%AF%92%E7%97%85%E6%AD%BB
%E4%BA%A1%E7%97%85%E4%BE%8B%E6%95%B8'

request = requests.get(url, headers=headers)
request.encoding = 'utf-8'

soup = BeautifulSoup(request.text, 'html.parser')

table = soup.find('table', class_='wikitable sortable mw-
datatable')

th, td = table.find_all('th'), table.find_all('td')

title = []
for t in th:
    title.append(t.text.replace('\xa0', '').replace('\n',
''))

main = []
this = []
for i, t in enumerate(td):
    if ((i % len(title)) == 0):
        main.append(this)
        this = [t.text.replace('\xa0', '').replace('\n', '')]
    else:
```

長榮大學 111 學年度第 1 學期 期中考試卷 Chang Jung Christian University Examination Sheet for Academic Year: 111 Semester: 1					
院系 College	資訊暨設計學院 資工系 3 年 B 班 College Department Year Class	姓名 Name	郭智榮	學號 Student No.	109B30612
科目名稱 Subject Name	大數據資料分析	教師簽章	黃琨義	評閱成績 Score	

```
        t = t.text.replace('\xa0', '').replace('\n', '')

        if ('年' in t):
            t = t.split('0000')[-1]

        this.append(t)
main.append(this)
main = main[1:]

dead = {0: 0, 3000: 0, 5000: 0, 10000: 0}

for i in main[3:]:
    i = i[2:]

    for j in i:
        j = int(j.replace(',', ''))

        if (j >= 10000):
            dead[10000] += 1

        if (j >= 5000):
            dead[5000] += 1

        if (j >= 3000):
            dead[3000] += 1

        if (j == 0):
            dead[0] += 1

print(f'死亡人數為 0 人: {dead[0]}')

print(f'死亡人數超過 3000 人: {dead[3000]}')
```

長榮大學 111 學年度第 1 學期 期中考試卷 Chang Jung Christian University Examination Sheet for Academic Year: 111 Semester: 1					
院系 College	資訊暨設計學院 資工系 3 年 B 班 College Department Year Class	姓名 Name	郭智榮	學號 Student No.	109B30612
科目名稱 Subject Name	大數據資料分析	教師簽章	黃琨義	評閱成績 Score	

```
print(f'死亡人數超過 5000 人: {dead[5000]}')

print(f'死亡人數超過 10000 人: {dead[10000]}')


with open('./result/wiki_table.csv', 'w', newline= '',
encoding='UTF-8-Sig') as f:
    writer = csv.writer(f, delimiter=',')
    writer.writerow(title)
    writer.writerows(main)
```

長榮大學 111 學年度第 1 學期 期中考試卷 Chang Jung Christian University Examination Sheet for Academic Year: 111 Semester: 1					
院系 College	資訊暨設計學院 資工系 3 年 B 班 College Department Year Class	姓名 Name	郭智榮	學號 Student No.	109B30612
科目名稱 Subject Name	大數據資料分析	教師簽章	黃琨義	評閱成績 Score	

4. 請撰寫程式於以下網頁中抓取新聞中心-媒體報導中所有日期與新聞(包含 1~40 頁)

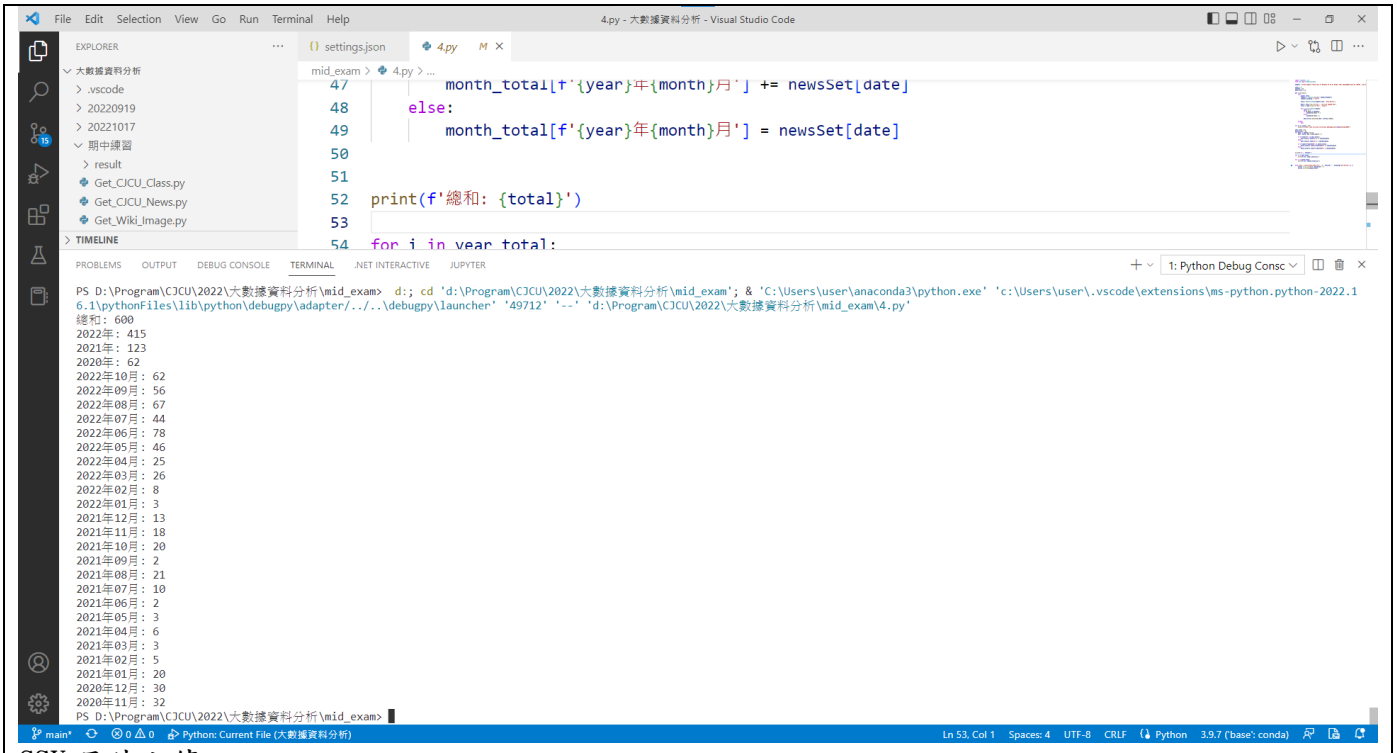
(a) 將資料以 csv 檔案格式儲存 (5%)

(b) 計算所有新聞數量，並輸出於 Console 介面 (10%)

(c) 計算每年每月新聞數量，並逐行輸出於 Console 介面 (10%)

https://www.cjcu.edu.tw/tw/news.php?id=NEWS

執行結果(請截圖貼上)：



CSV 另外上傳

程式碼(請將完整程式碼貼入下方表格，請貼文字，勿貼圖，否則不予計分)：

長榮大學 111 學年度第 1 學期 期中考試卷 Chang Jung Christian University Examination Sheet for Academic Year: 111 Semester: 1					
院系 College	資訊暨設計學院 資工系 3 年 B 班 College Department Year Class	姓名 Name	郭智榮	學號 Student No.	109B30612
科目名稱 Subject Name	大數據資料分析	教師簽章	黃琨義	評閱成績 Score	

```
import requests, csv
from bs4 import BeautifulSoup

headers = {"User-Agent":"Mozilla/5.0 (Windows NT 10.0; Win64;
x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/96.0.4664.110 Safari/537.36 Edg/96.0.1054.57"}

total = 0
newsSet = {}
date_title = []

def spider(url):
    try:
        global total
        request = requests.get(url, headers=headers)
        request.encoding = 'utf-8'

        soup = BeautifulSoup(request.text, 'html.parser')

        date = soup.find_all('div', 'col-sm-3 padded-tb5')
        title = soup.find_all('div', 'topic')

        for i, d in enumerate(date):
            total += 1
            if (d.text in newsSet):
                newsSet[d.text] += 1
            else:
                newsSet[d.text] = 1

            date_title.append([d.text, title[i].text])

    except:
        pass
```

長榮大學 111 學年度第 1 學期 期中考試卷 Chang Jung Christian University Examination Sheet for Academic Year: 111 Semester: 1					
院系 College	資訊暨設計學院 資工系 3 年 B 班 College Department Year Class	姓名 Name	郭智榮	學號 Student No.	109B30612
科目名稱 Subject Name	大數據資料分析	教師簽章	黃琨義	評閱成績 Score	

```
for p in range(1, 41):
    spider(f'https://www.cjcu.edu.tw/tw/news.php?page_want={p
}&newstype=NEWS')

year_total = {}
month_total = {}
for date in newsSet.keys():
    year, month, day = date.split('/')

    if (f'{year}年' in year_total):

        year_total[f'{year}年'] += newsSet[date]
    else:
        year_total[f'{year}年'] = newsSet[date]


    if (f'{year}年{month}月' in month_total):

        month_total[f'{year}年{month}月'] += newsSet[date]
    else:
        month_total[f'{year}年{month}月'] = newsSet[date]



print(f'總和: {total}')

for i in year_total:
    print(f'{i}: {year_total[i]}')

for i in month_total:
    print(f'{i}: {month_total[i]}')
```

長榮大學 111 學年度第 1 學期 期中考試卷 Chang Jung Christian University Examination Sheet for Academic Year: 111 Semester: 1					
院系 College	資訊暨設計學院 資工系 3 年 B 班 College Department Year Class	姓名 Name	郭智榮	學號 Student No.	109B30612
科目名稱 Subject Name	大數據資料分析	教師簽章	黃琨義	評閱成績 Score	

```
with open('./result/news_table.csv', 'w', newline= '',
encoding='UTF-8-Sig') as f:
    writer = csv.writer(f, delimiter=',')
    writer.writerows(date_title)
```