现代程序设计第14周作业

谢奕飞 20377077

代码

WriteHead函数

```python
def WriteHead(head,dir_path):
    '''
    写入表头
    '''
    if not os.path.exists(dir_path):
        os.mkdir(dir_path)
    with open(dir_path+'/data.csv','w',encoding='utf8',newline='') as f:
        writer=csv.writer(f)
        writer.writerow(head)
```

GetPage函数

```python
def GetPage(url,headers):
    '''
    获取页数
    '''
    response=requests.get(url=url,headers=headers)
    soup=BeautifulSoup(response.text,'lxml')
    return int(soup.select('a[class="zpgi"]')[-1].get_text())
```

DownloadImage函数

```python
def DownloadImage(url,dir_path):
    '''
    下载图片
    '''
    if not os.path.exists(dir_path):
        os.mkdir(dir_path)
    name=url.split('/')[-1]
    img_path=dir_path+'/'+name
    try:
        urllib.request.urlretrieve(url,filename=img_path)
        urllib.request.urlcleanup()
    except:
        return 'error'
    return img_path
```

Producer函数

```
def Producer(q:Queue,url,headers):
    '''
    生产者
    '''
    response=requests.get(url=url,headers=headers)
    soup=BeautifulSoup(response.text,'lxml')
    soup_list=soup.select('a[class="tit f-thide s-fc0"]')
    href_list=[]
    for s in soup_list:
        href_list.append('https://music.163.com'+str(s['href']))
    q.put(href_list)
```

Consumer函数

```
def Consumer(q:Queue,headers,dir_path):
    '''
    消费者
    '''
    urls=q.get()
    if urls!=None:
        for url in urls:
            response=requests.get(url=url,headers=headers)
            soup=BeautifulSoup(response.text,'lxml')
            id=url.split('id=')[-1]#歌单id
            title=soup.select('.tit')[0].get_text()[1:]#歌单标题
            image_url=soup.select('img[class="j-img"]')[0]['data-src']#封面u
            image=DownloadImage(image_url,dir_path+'/images')#下载图片
            author_id=soup.select('a[class="s-fc7"]')[0]['href'].split('id='
            author=soup.select('a[class="s-fc7"]')[0].get_text()#作者名
            description=soup.select('p')[1].get_text()#简介
            count=soup.select('span[id="playlist-track-count"]')[0].get_text
            play=soup.select('strong[id="play-count"]')[0].get_text()#播放次
            add=soup.select('a[class="u-btni u-btni-fav"]')[0]['data-count'
            share=soup.select('a[class="u-btni u-btni-share"]')[0]['data-cou
            comment=soup.select('span[id="cnt_comment_count"]')[0].get_text
            #按行写入
            with open(dir_path+'/data.csv','a',encoding='utf8',newline='') a
                writer=csv.writer(f)
                writer.writerow([id,title,image,author_id,author,descriptio
```

CoroProducer函数

```
def CoroProducer(q:Queue,url):
    '''
    协程执行Producer
    '''
    tasks=[]
    headers={'User-Agent':'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWe
    N=GetPage(url,headers)#获取页数
    print('This website has {} pages to crawl'.format(N))
    #N=3
    urls=[]
    for i in range(N):
        urls.append(url[:-1]+str(i*35))
    for url in urls:
        task=gevent.spawn(Producer,q,url,headers)
        tasks.append(task)
    gevent.joinall(tasks)
    return N
```

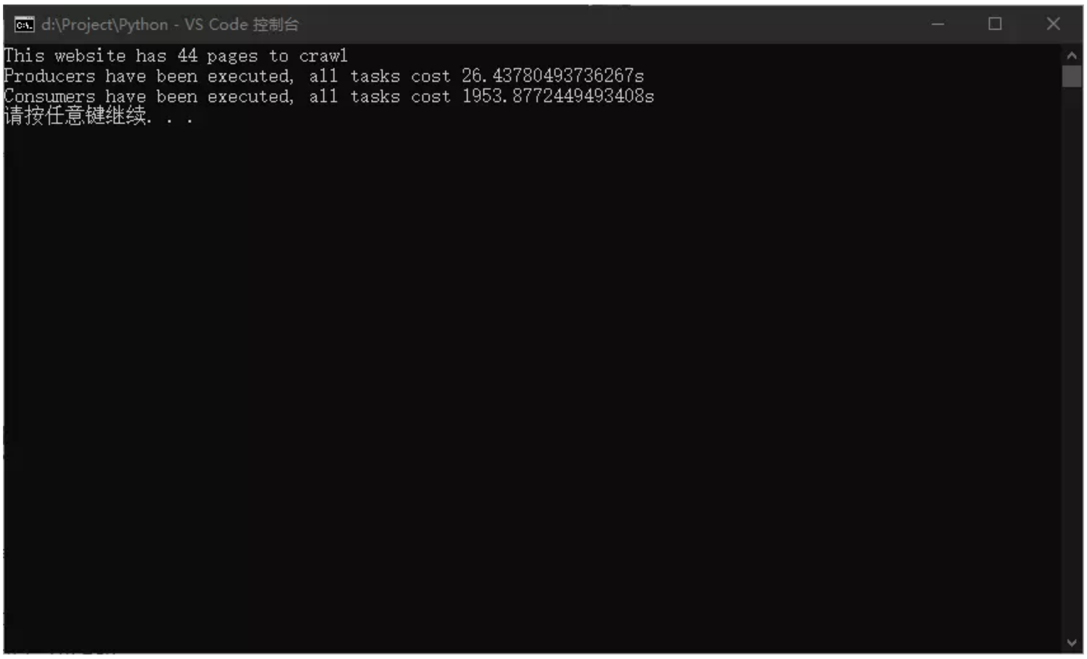
CoroConsumer函数

```
def CoroConsumer(q:Queue,N,dir_path):
    '''
    协程执行Consumer
    '''
    headers={'User-Agent':'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWe
    tasks=[]
    for i in range(N):
        task=gevent.spawn(Consumer,q,headers,dir_path)
        tasks.append(task)
    gevent.joinall(tasks)
```

main函数

```
if __name__=='__main__':
    dir_path='D:/Project/Python/week14/result'
    head=['id','title','image','author_id','author','description','count','p
    url='https://music.163.com/discover/playlist/?order=hot&cat=%E6%B0%91%E8
    q=Queue()
    t_start=time.time()
    N=CoroProducer(q,url)#协程执行Producer
    print('Producers have been executed, all tasks cost {}s'.format(time.tin
    WriteHead(head,dir_path)#写入表头
    CoroConsumer(q,N,dir_path)#协程执行Consumer
    print('Consumers have been executed, all tasks cost {}s'.format(time.tin
    playsound('D:\Project\Python\week14\over.mp3')
```

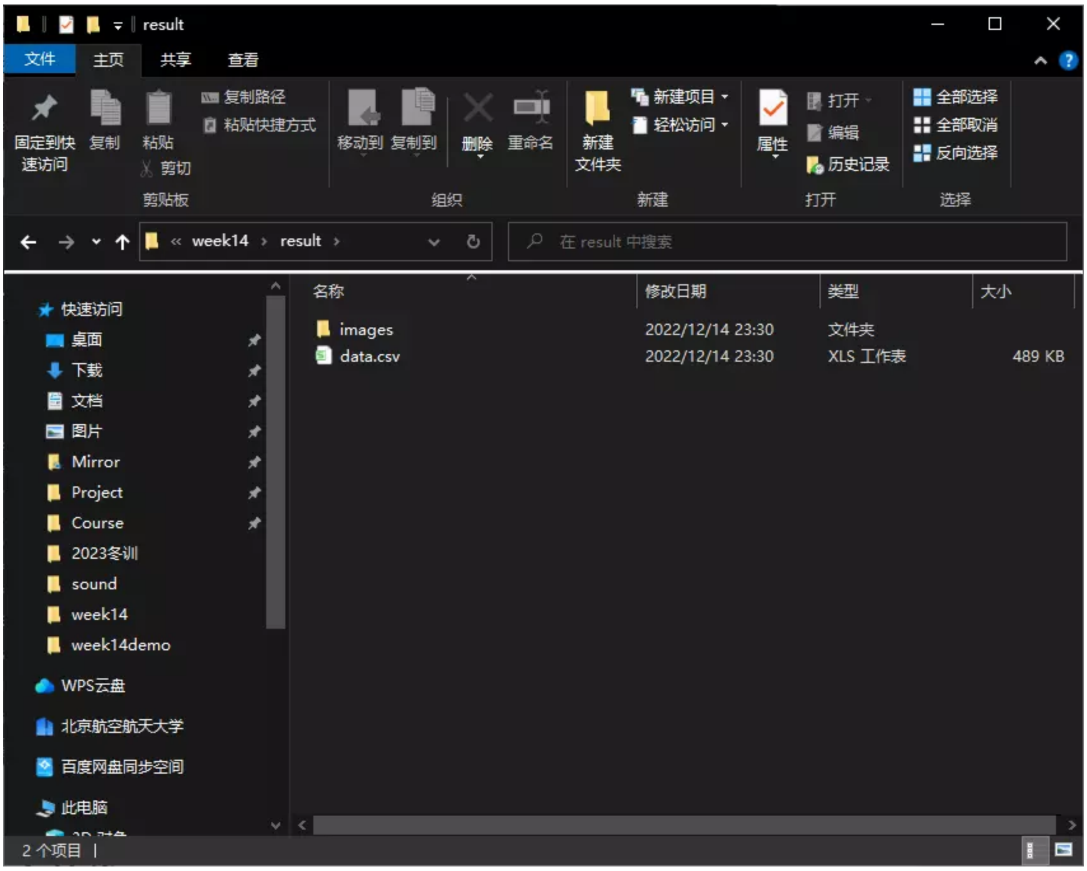
运行结果



```
This website has 44 pages to crawl
Producers have been executed, all tasks cost 26.43780493736267s
Consumers have been executed, all tasks cost 1953.8772449493408s
请按任意键继续. . .
```

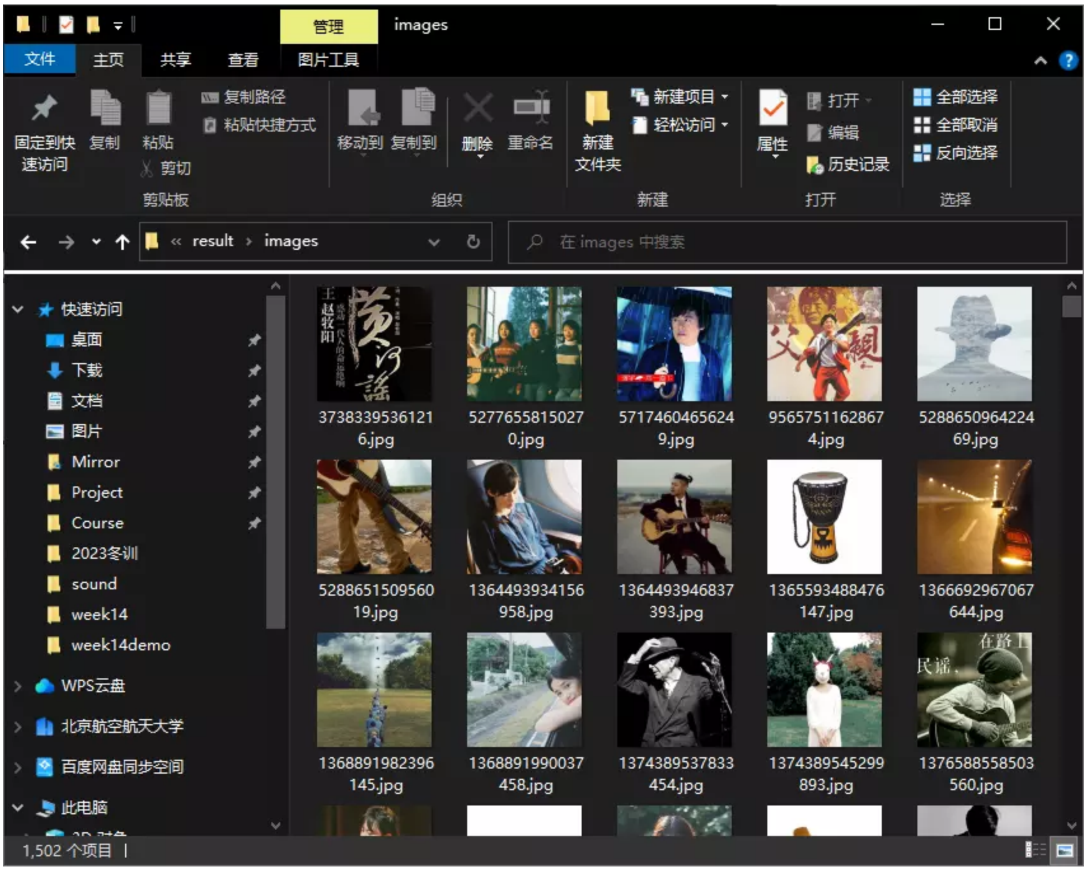
分别用时26.43s和1953.87s

results目录

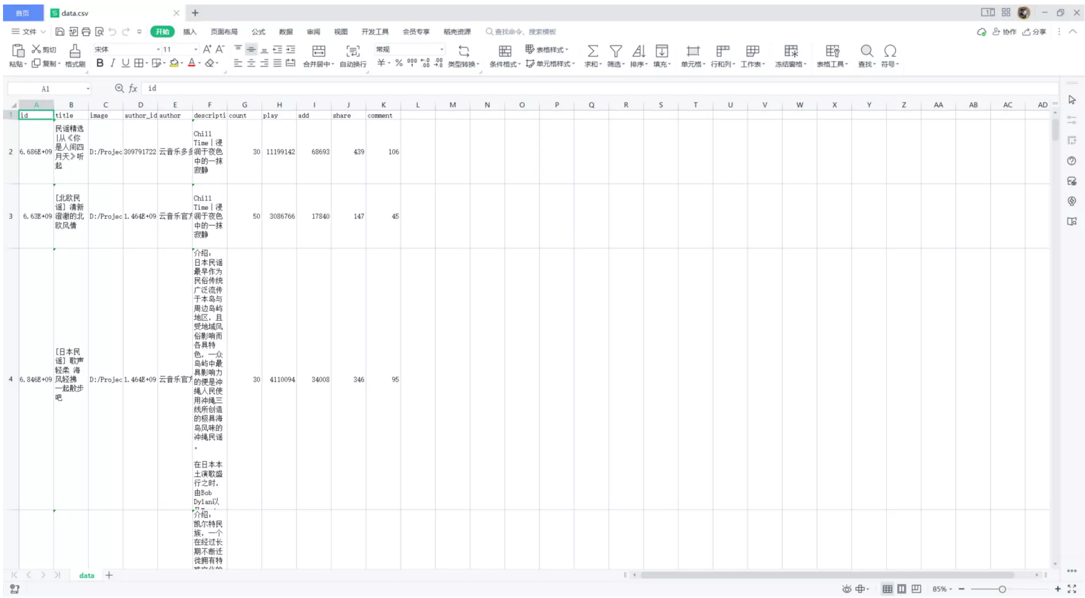


名称	修改日期	类型	大小
images	2022/12/14 23:30	文件夹	
data.csv	2022/12/14 23:30	XLS 工作表	489 KB

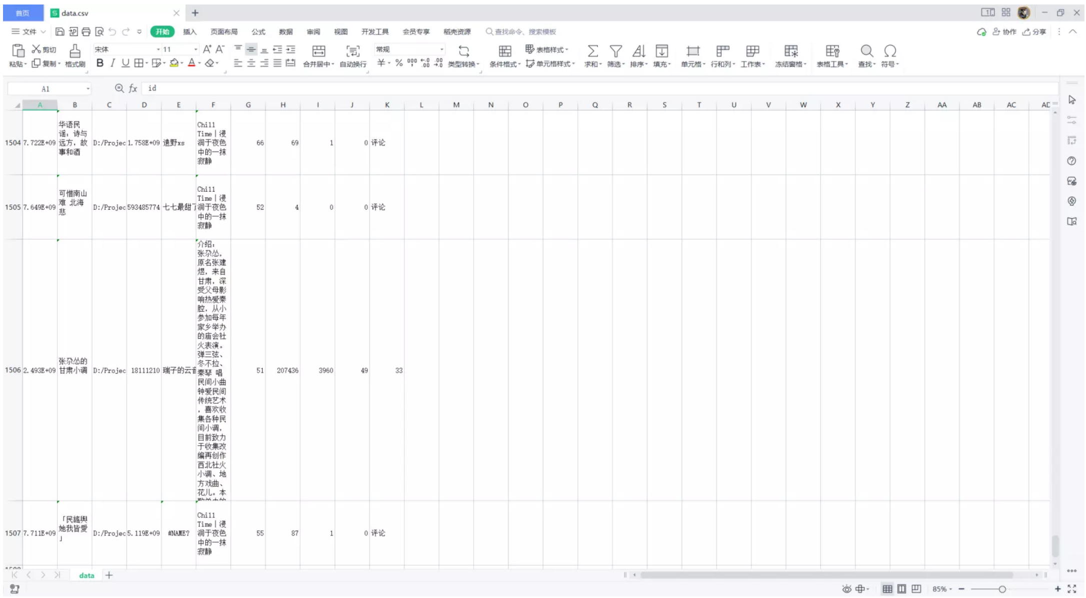
images目录



data.csv



	A	B	C	D	E	F	G	H	I	J	K
1	id	title	image	author_id	author	descripti	count	play	add	share	comment
2	6.686E+09	民谣精选 |从《你是人间四月天》听起	D:/Projec	309791722	云音乐多多	Chill Time | 浸润于夜色中的一抹寂静	30	11199142	68693	439	106
3	6.63E+09	[北欧民谣] 清新澄澈的北欧风情	D:/Projec	1.464E+09	云音乐官方	Chill Time | 浸润于夜色中的一抹寂静	50	3086766	17840	147	45
4	6.846E+09	[日本民谣] 歌声轻柔 海风轻拂 一起散步吧	D:/Projec	1.464E+09	云音乐官方	介绍： 日本民谣最早作为民俗传统广泛流传于本岛与周边岛屿地区，且受地域风俗影响而各具特色，一众岛屿中最具影响力的便是冲绳人民使用冲绳三线所创造的极具海岛风味的冲绳民谣。 在日本本土演歌盛行之时，由Bob Dylan以	30	4110094	34008	346	95
						介绍： 凯尔特民族，一个在经过长期不断迁徙拥有特					



	A	B	C	D	E	F	G	H	I	J	K
1504	7.722E+09	华语民谣：诗与远方，故事和酒	D:/Projec	1.758E+09	遠野xs	Chill Time丨浸润于夜色中的一抹寂静	66	69	1	0	评论
1505	7.649E+09	可惜南山难 北海悲	D:/Projec	593485774	七七最甜了	Chill Time丨浸润于夜色中的一抹寂静	52	4	0	0	评论
1506	2.493E+09	张尕怂的甘肃小调	D:/Projec	18111210	瑞子的云音	介绍：张尕怂，原名张建煜，来自甘肃，深受父母影响热爱秦腔，从小参加每年家乡举办的庙会社火表演。弹三弦、冬不拉、秦琴 唱民间小曲钟爱民间传统艺术，喜欢收集各种民间小调，目前致力于收集改编再创作西北社火小调、地方戏曲、花儿。本	51	207436	3960	49	33
1507	7.711E+09	「民謠與她我皆愛」	D:/Projec	5.119E+09	#NAME?	Chill Time丨浸润于夜色中的一抹寂静	55	87	1	0	评论

附录-完整代码

```python
import requests,time,csv,os,urllib.request,gevent
from bs4 import BeautifulSoup
from queue import Queue
from threading import Thread
from playsound import playsound

def WriteHead(head,dir_path):
    '''
    写入表头
    '''
    if not os.path.exists(dir_path):
        os.mkdir(dir_path)
    with open(dir_path+'/data.csv','w',encoding='utf8',newline='') as f:
        writer=csv.writer(f)
        writer.writerow(head)

def GetPage(url,headers):
    '''
    获取页数
    '''
    response=requests.get(url=url,headers=headers)
    soup=BeautifulSoup(response.text,'lxml')
    return int(soup.select('a[class="zpgi"]')[-1].get_text())

def DownloadImage(url,dir_path):
    '''
    下载图片
    '''
    if not os.path.exists(dir_path):
        os.mkdir(dir_path)
    name=url.split('/')[-1]
    img_path=dir_path+'/'+name
    try:
        urllib.request.urlretrieve(url,filename=img_path)
        urllib.request.urlcleanup()
    except:
        return 'error'
    return img_path

def Producer(q:Queue,url,headers):
    '''
    生产者
    '''
    response=requests.get(url=url,headers=headers)
    soup=BeautifulSoup(response.text,'lxml')
    soup_list=soup.select('a[class="tit f-thide s-fc0"]')
    href_list=[]
    for s in soup_list:
        href_list.append('https://music.163.com'+str(s['href']))
    q.put(href_list)

def Consumer(q:Queue,headers,dir_path):
    '''
    消费者
    '''
    urls=q.get()
    if urls!=None:
        for url in urls:
            response=requests.get(url=url,headers=headers)
            soup=BeautifulSoup(response.text,'lxml')
            id=url.split('id=')[-1]#歌单id
            title=soup.select('.tit')[0].get_text()[1:]#歌单标题
            image_url=soup.select('img[class="j-img"]')[0]['data-src']#封面url
            image=DownloadImage(image_url,dir_path+'/images')#下载图片
            author_id=soup.select('a[class="s-fc7"]')[0]['href'].split('id=')[-1]#作者id
            author=soup.select('a[class="s-fc7"]')[0].get_text()#作者名
            description=soup.select('p')[1].get_text()#简介
            count=soup.select('span[id="playlist-track-count"]')[0].get_text()#歌曲数
            play=soup.select('strong[id="play-count"]')[0].get_text()#播放次数
            add=soup.select('a[class="u-btni u-btni-fav"]')[0]['data-count']#收藏数
            share=soup.select('a[class="u-btni u-btni-share"]')[0]['data-count']#分享数
            comment=soup.select('span[id="cnt_comment_count"]')[0].get_text()#评论数
            #按行写入
            with open(dir_path+'/data.csv','a',encoding='utf8',newline='') as f:
                writer=csv.writer(f)
                writer.writerow([id,title,image,author_id,author,description,count,play,add,share,comment])

def CoroProducer(q:Queue,url):
    '''
    协程执行Producer
    '''
    tasks=[]
    headers={'User-Agent':'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36 Edg/107.0.1418.35'}
    N=GetPage(url,headers)#获取页数
    print('This website has {} pages to crawl'.format(N))
    #N=3
    urls=[]
    for i in range(N):
        urls.append(url[:-1]+str(i*35))
    for url in urls:
        task=gevent.spawn(Producer,q,url,headers)
        tasks.append(task)
    gevent.joinall(tasks)
    return N

def CoroConsumer(q:Queue,N,dir_path):
    '''
    协程执行Consumer
    '''
    headers={'User-Agent':'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36 Edg/107.0.1418.35'}
    tasks=[]
    for i in range(N):
        task=gevent.spawn(Consumer,q,headers,dir_path)
        tasks.append(task)
    gevent.joinall(tasks)

if __name__=='__main__':
    dir_path='D:/Project/Python/week14/result'
    head=['id','title','image','author_id','author','description','count','play','add','share','comment']
    url='https://music.163.com/discover/playlist/?order=hot&cat=%E6%B0%91%E8%B0%A3&limit=35&offset=0'
    q=Queue()
    t_start=time.time()
    N=CoroProducer(q,url)#协程执行Producer
    print('Producers have been executed, all tasks cost {}s'.format(time.time()-t_start))
    WriteHead(head,dir_path)#写入表头
    CoroConsumer(q,N,dir_path)#协程执行Consumer
    print('Consumers have been executed, all tasks cost {}s'.format(time.time()-t_start))
    playsound('D:\Project\Python\week14\over.mp3')
```