获取电视剧《大秦帝国之裂变》九页评论代码

import csv  
import random  
import time  
import requests  
from bs4 import BeautifulSoup  
  
user\_agents = [  
 'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/83.0.4103.61 Safari/537.36'  
 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_14\_3) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/12.0.3 Safari/605.1.15',  
 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_14\_3) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/83.0.4103.116 Safari/537.36',  
 "Mozilla/5.0 (Windows NT 10.0; WOW64; rv:38.0) Gecko/20100101 Firefox/38.0",  
 "Mozilla/5.0 (Windows NT 10.0; WOW64; Trident/7.0; .NET4.0C; .NET4.0E; .NET CLR 2.0.50727; .NET CLR 3.0.30729; .NET CLR 3.5.30729; InfoPath.3; rv:11.0) like Gecko",  
 "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0)",  
 "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; Trident/4.0)",  
 "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0)",  
 "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)",  
 "Mozilla/5.0 (Macintosh; Intel Mac OS X 10.6; rv:2.0.1) Gecko/20100101 Firefox/4.0.1",  
 "Mozilla/5.0 (Windows NT 6.1; rv:2.0.1) Gecko/20100101 Firefox/4.0.1",  
 "Opera/9.80 (Macintosh; Intel Mac OS X 10.6.8; U; en) Presto/2.8.131 Version/11.11"  
]  
  
  
# 获取html  
def get\_resource(url, params=None, flag='html'):  
 headers = {  
 'Host': 'movie.douban.com',  
 'User-Agent': random.choice(user\_agents),  
 'accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,\*/\*;q=0.8,application/signed-exchange;v=b3;q=0.9'  
 }  
 response = requests.get(url=url, params=params, headers=headers)  
 if response.status\_code == 200:  
 # 判断flag  
 if flag == 'html':  
 return response.text  
 elif flag == 'media':  
 return response.content  
 else:  
 print('获取资源有误！')  
  
  
def parse\_html\_other(resource):  
 comment\_list = []  
 soup = BeautifulSoup(resource, 'lxml')  
 comments = soup.select\_one('#comments')  
 comment\_items = comments.select('.comment-item')  
 for item in comment\_items:  
 comment\_info = item.select\_one('.comment h3 .comment-info')  
 username = comment\_info.find('a').text  
 comment\_text = item.find('span', attrs={'class': 'short'}).text  
 comment\_time = comment\_info.select\_one('.comment-time ').text  
 comment = [username, comment\_text, comment\_time]  
 comment[2] = comment[2].strip()  
 comment\_list.append(comment)  
  
 return comment\_list  
  
  
# 保存  
def save\_data(comment\_list):  
 with open('data/douban.csv', mode='a', newline='', encoding='utf-8') as fw:  
 writer = csv.writer(fw)  
 # 遍历评论列表  
 writer.writerows(comment\_list)  
  
 print('保存完毕！')  
  
  
if \_\_name\_\_ == '\_\_main\_\_':  
 url = 'https://movie.douban.com/subject/3114220/comments'  
 params = {'status': 'P', 'limit': 20, 'sort': 'new\_score'}  
 for i in range(9):  
 n = i \* 20  
 params['start'] = n  
 resource = get\_resource(url=url, params=params)  
 comment\_list = parse\_html\_other(resource)  
 save\_data(comment\_list)  
 time.sleep(random.randint(4, 6))

获取评论词云图代码

import csv  
import jieba  
import numpy  
from PIL import Image  
from wordcloud import WordCloud  
  
# 获取评论字符串  
comments = ''  
with open('data/douban.csv', 'r', encoding='utf-8') as fr:  
 reader = csv.reader(fr)  
 for line in reader:  
 comments += line[1]  
result = jieba.cut(comments)  
result = list(result)  
text = ' '.join(result)  
image = numpy.array(Image.open('pictures/sheep.jpg'))  
wcloud = WordCloud(font\_path='simsun.ttc', mask=image)  
wcloud.generate(text)  
wcloud.to\_file('pictures/result\_picture.png')

（3）结果截图



②

（1）搜索电视剧名称爬取电视剧评论和评论用户所在城市，并且可视化每集播放量代码

# 仅针对腾讯视频此平台具有的电视资源（不包括电影以及其他平台）  
# 7.16 7：48  
# 提前进程  
import csv  
import random  
import re  
import os  
import requests  
import time  
from bs4 import BeautifulSoup  
  
user\_agents = [  
 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_14\_3) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/12.0.3 Safari/605.1.15',  
 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_14\_3) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/83.0.4103.116 Safari/537.36',  
 "Mozilla/5.0 (Windows NT 10.0; WOW64; rv:38.0) Gecko/20100101 Firefox/38.0",  
 "Mozilla/5.0 (Windows NT 10.0; WOW64; Trident/7.0; .NET4.0C; .NET4.0E; .NET CLR 2.0.50727; .NET CLR 3.0.30729; .NET CLR 3.5.30729; InfoPath.3; rv:11.0) like Gecko",  
 "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0)",  
 "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; Trident/4.0)",  
 "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0)",  
 "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)",  
 "Mozilla/5.0 (Macintosh; Intel Mac OS X 10.6; rv:2.0.1) Gecko/20100101 Firefox/4.0.1",  
 "Mozilla/5.0 (Windows NT 6.1; rv:2.0.1) Gecko/20100101 Firefox/4.0.1",  
 "Opera/9.80 (Macintosh; Intel Mac OS X 10.6.8; U; en) Presto/2.8.131 Version/11.11",  
]  
  
# 获取html  
def get\_resource(url, params=None, flag='html'):  
 headers = {  
 # 'Host': 'coral.qq.com',  
 'User-Agent': random.choice(user\_agents)  
 }  
 # 使用requests发出请求  
 response = requests.get(url=url, params=params, headers=headers)  
 # 判断response的状态码  
 if response.status\_code == 200:  
 # 判断flag  
 if flag == 'html':  
 return response.text  
 elif flag == 'media':  
 return response.content  
 else:  
 print('response.status\_code:',response.status\_code)  
 print('获取资源有误！')  
  
# 搜寻影视剧并获取电视剧总集数，同时得到每一集的网址  
def search\_television():  
 print('\*' \* 70)  
 print("\*" \* 12, '腾讯视频电视剧的评论下载和分析', '\*' \* 12)  
 print('\*' \* 70)  
 # https://v.qq.com/x/search/?q=大秦帝国之裂变  
 name = input('输入电视剧名称: ')  
 params = {'q': name}  
 url = 'https://v.qq.com/x/search/'  
 resource = get\_resource(url=url, params=params)  
 if resource:  
 soup = BeautifulSoup(resource, 'lxml')  
 tv\_url = soup.find('a', {'class': 'figure result\_figure'}).get('href')  
 # print('tv\_url',tv\_url)  
 tv\_resource = get\_resource(url=tv\_url)  
 if(tv\_resource):  
 episode\_list = []  
 tv\_soup = BeautifulSoup(tv\_resource, 'lxml')  
 episode\_items = tv\_soup.find\_all('span', {'\_stat':'series:numbtn'})  
 for item in episode\_items:  
 # episode\_url ---> 某集地址  
 episode\_url = item.find('a').get('href')  
 # episode\_id ---> 某集集数  
 episode\_id = item.find('a').text  
 episode\_id = episode\_id.replace('\n', '')  
 episode\_id = episode\_id.replace(' ', '')  
 episode = [episode\_id, episode\_url]  
 episode\_list.append(episode)  
 return name, episode\_list  
 else:  
 print('请注意是否搜索的是电视资源！')  
 else:  
 print('获取资源有误！')  
  
# 获取不同集数的评论网页网址并保存相应内容  
def select\_episode():  
 name, episode\_list = search\_television()  
 comnum\_list = []  
 columns = []  
 for episode in episode\_list:  
  
 episode\_resource = get\_resource(url=episode[1])  
 path = name + '/' + episode[0]  
 # comment\_id  
 patid = '"comment\_id":"(.\*?)",'  
 comment\_id = re.compile(patid).findall(episode\_resource)[0]  
  
 # https://coral.qq.com/4003145426  
 # 原地址 https://coral.qq.com/article/4003145426/comment/v2?callback=\_article4003145426commentv2&orinum=10&oriorder=o&pageflag=1&cursor=0  
 # 简化后 https://coral.qq.com/article/4003145426/comment/v2?callback=\_article4003145426commentv2&oriorder=o&cursor=0  
 url = 'https://coral.qq.com/article/{}/comment/v2'.format(comment\_id)  
 callback = '\_article{}commentv2'.format(comment\_id)  
 # 初始化为0  
 cursor = 0  
 params = {'callback': callback, 'oriorder': 'o', 'cursor': cursor}  
 comment\_resource = get\_resource(url, params)  
 patnum = '"commentnum":"(.\*?)",'  
  
 comment\_num = re.compile(patnum).findall(comment\_resource)[0]  
 comnum\_list.append(comment\_num)  
 columns.append(episode[0])  
 # print('comment\_num:', comment\_num)  
 # 每集至多保存100条评论  
 city\_list1 = []  
 for i in range(0, 10):  
 data = comment\_resource  
 # 获取下一个Cursor  
 pat\_next = '"last":"(.\*?)",'  
 nextcursor = re.compile(pat\_next).findall(data)[0]  
  
 # 抓取评论信息  
 pat\_com = '"content":"(.\*?)",'  
 comdata = re.compile(pat\_com).findall(data)  
 with open(path+'.txt', mode='a', encoding='utf-8') as fw:  
 fw.write(str(comdata))  
 # 抓取用户地址  
 pat\_region = '"region":"(中国:.\*?)",'  
 region\_data = re.compile(pat\_region).findall(data)  
  
 for item in region\_data:  
 item = item.split(':')[2]  
 if(len(item)):  
 city\_list1.append(item)  
  
  
 # 时间间隔  
 # time.sleep(random.randint(1, 3))  
  
 # 更新url  
 params = {'callback': callback, 'oriorder': 'o', 'cursor': nextcursor}  
 comment\_resource = get\_resource(url, params)  
 with open(name + '/'+ episode[0]+ 'city.txt', mode='w', encoding='utf-8') as fw:  
 fw.write(str(city\_list1))  
 print("\*" \* 12, '第' + episode[0] + '集评论保存完毕', '\*' \* 12)  
 print("\*" \* 12, name + '的评论保存完毕', '\*' \* 12)  
 return name, comnum\_list, columns  
  
if \_\_name\_\_ == '\_\_main\_\_':  
 name, comnum\_list, columns= select\_episode()  
 from pyecharts import Line  
 line = Line("折线图", name+"每集评论数")  
 # is\_label\_show是设置上方数据是否显示  
 line.add("评论数", columns, comnum\_list, is\_label\_show=True)  
 line.render(name+'/'+ 'render1.html')

（2）结果截图

