

第十一次作业

第一部分爬取 id

```
import requests
from bs4 import BeautifulSoup
import re
headers = {
           'Connection': 'close',
           'Referer': 'https://music.163.com/',
           'Host': 'music.163.com',
           "Accept-Encoding": "identity",
           "Accept": "text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/
           "user-agent": 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (
for i in range(0,1300,35):
    path = 'https://music.163.com'
    url = 'https://music.163.com/discover/playlist/?order=hot&cat=说唱&limit=35&offset=' + str
    response = requests.session().get(url=url, headers=headers)
   html = response.content.decode('utf-8')
   playlist_ids = []
    playlist_ids.extend(re.findall(r'playlist\?id=(\d+?)" class="msk"', response.text))
    #print(playlist_ids)
    #patt=re.compile('<article.*</article>',re.S)
    #print(html)
    soup = BeautifulSoup(html, 'html.parser')
    #test = soup.find('ul')
    #print(test)
    text = soup.find('ul',class_='m-cvrlst f-cb')
    #print(len(text))
    #print(text)
    id_list = []
    id_text = soup.select(".dec a")
    for i in range(len(id_text)):
       id_list.append(id_text[i]['href'])
   print(id_list)
    #print(title)
    #title_list = soup.select("#m-pl-container li")
    #print(title_list)
    break
```

```
from os import name
from bs4.element import CharsetMetaAttributeValue
import requests
from bs4 import BeautifulSoup
import urllib.request
save_path = 'info/'
img_path = 'cover/'
headers = {
           'Connection': 'close',
           'Referer': 'https://music.163.com/',
           'Host': 'music.163.com',
           "Accept-Encoding": "identity",
           "Accept": "text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/
           "user-agent": 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (
def download_img(path,id,src):
   with open(path+id+'.jpg','wb+') as f:
       response = urllib.request.urlopen(src)
       img = response.read()
       f.write(img)
def main():
   sheet = 0
   with open(save_path+str(sheet)+'说唱.csv','a+') as f:
       f.write('歌曲 ID'+','+'歌单标题'+','+'创建者昵称'+','+'创建者id'+','+'介绍'+','+'歌曲数量'+
   id_list = ['/playlist?id=6952442013', '/playlist?id=2695101341', '/playlist?id=6862602894'
   path = 'https://music.163.com'
   for i in range(len(id_list)):
       id = id_list[i][-10:]
       url = path+id_list[i]
       response = requests.session().get(url=url, headers=headers)
       html = response.content.decode('utf-8')
       soup = BeautifulSoup(html, 'html.parser')
       #print(soup)
       # 歌单的封面图片 (需把图片保存到本地) 、歌单标题、创建者 id、创建者昵称、介绍、歌曲数量、播放量、%
       # 图片下载
       img_info = soup.find('div',class_='cover u-cover u-cover-dj')
       img_url = img_info.find('img',class_='j-img')['data-src']
       img_id = id_list[i][-10:]
       download_img(img_path,img_id,img_url)
       #print(img_url)
       # 歌单标题
       title = soup.find('h2',class_ = 'f-ff2 f-brk').text
       #print(title)
       # 创建者昵称
       create_name = soup.find('span',class_='name').text.strip('\n')
       #print(create_name)
       # 创建者id
       create_id = soup.find('span',class_='name').select('a')[0]['href'][14:]
       #print(create_id)
       # 介绍
       describe = soup.find('p',class_='intr f-brk').text.replace('\n','\t').replace(',',',','
       #print(describe)
       # 歌曲数量
       song_num = soup.find('span','sub s-fc3').find('span').text
       #print(song_num)
       #播放量
       play_num = soup.find('strong','s-fc6').text
```

```
#print(play_num)

# 添加到播放列表次数
add_num = soup.find('a','u-btni u-btni-fav')['data-count']
#print(add_num)

# 分享次数
share_num = soup.find('a','u-btni u-btni-share')['data-count']
#print(share_num)

# 评论数
comment_num = soup.find('a','u-btni u-btni-cmmt').find('span').text
#print(comment_num)

with open(save_path+str(sheet)+'说唱.csv','a+') as f:
    f.write(id+','+title+','+create_name+','+create_id+','+str(describe)+','+song_num-main()
```

整理后的多线程程序

```
from multiprocessing import Queue
import requests
from bs4 import BeautifulSoup
import re
from threading import Thread, currentThread
import queue
import urllib.request
import time
save_path = 'info/'
img_path = 'cover/'
produce_num = 0
consumer_num = 0
headers_1 = {
          'Connection': 'close',
          'Referer': 'https://music.163.com/',
          'Host': 'music.163.com',
          "Accept-Encoding": "identity",
          "Accept": "text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/
          "user-agent": 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (
headers_2 = {
          'Connection': 'close',
          'Referer': 'https://music.163.com/',
          'Host': 'music.163.com',
          "Accept-Encoding": "identity",
          "Accept": "text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/
          "user-agent": 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (
# 以爬取网易云歌单为例,练习多线程的使用。
def download_img(path,title,src):
   title = title.replace('/',' ')
   with open(path+title+'.jpg','wb+') as f:
       response = urllib.request.urlopen(src)
       img = response.read()
       f.write(img)
# 1. 获取一个分类下的所有歌单的 id。
# 观察 url 可以发现其页码规律: https://music.163.com/#/discover/playlist/?order=hot&cat=%E8%AF%B4%
# offset是本页开始的数据位置,第1页是 0,第2页是 35。分类参数是 utf-8编码后的汉字"说唱",使用 str.encode
class produce_id(Thread):
   def __init__(self,pages,q):
       super().__init__()
       self.pages = pages
```

```
self.q = q
       self.num = 0
    def run(self):
       print(f'the no{currentThread().name} process_id has started')
       for i in self.pages:
           time.sleep(1)
           path = 'https://music.163.com'
           url = 'https://music.163.com/discover/playlist/?order=hot&cat=说唱&limit=35&offset
           try:
               while True:
                   response = requests.session().get(url=url, headers=headers_1)
                   time.sleep(1)
                   if response != []:
                       break
           except:
               print('网络请求错误')
           html = response.content.decode('utf-8')
           playlist_ids = []
           playlist_ids.extend(re.findall(r'playlist\?id=(\d+?)" class="msk"', response.text))
           #print(playlist_ids)
           soup = BeautifulSoup(html, 'html.parser')
           #test = soup.find('ul')
           #print(test)
           text = soup.find('ul',class_='m-cvrlst f-cb')
           #print(len(text))
           #print(text)
           id_list = []
           id_text = soup.select(".dec a")
           for i in range(len(id_text)):
               id_list.append(id_text[i]['href'])
           print(id_list)
           #print(title)
           #title_list = soup.select("#m-pl-container li")
           #print(title_list)
           self.q.put(id_list)
# 2. 对每个 id, 获取歌单的详细信息, 至少包括:
# 歌单的封面图片 (需把图片保存到本地)、歌单标题、创建者 id、创建者昵称、介绍、歌曲数量、播放量、添加到播放
# 可以自行实现其他信息的获取。基本信息汇总到同一张表中,以 csv 文件保存。https://music.163.com/playlist
class consumer(Thread):
    def __init__(self,q):
       super().__init__()
       self.q = q
       self.num = 0
    def run(self):
       print(f'the no{currentThread().name} pconsumer has started')
       while True:
           id_list = self.q.get()
           print(id_list)
           if id_list == []:
               break
           path = 'https://music.163.com'
           for i in range(len(id_list)):
               time.sleep(1)
               id = id_list[i][-10:]
               url = path+id_list[i]
               try:
                   count = 1
                   while True:
                       response = requests.session().get(url=url, headers=headers_1)
                       html = response.content.decode('utf-8')
                       soup = BeautifulSoup(html, 'html.parser')
                       time.sleep(count)
                       if soup.find('p',class_='intr f-brk') != None:
                           break
```

```
html = response.content.decode('utf-8')
                       soup = BeautifulSoup(html, 'html.parser')
                      time.sleep(count)
                       if soup.find('p',class_='intr f-brk') != None:
                       count += 1
               except:
                   print(soup.find('p',class_='intr f-brk'))
                   print('网络超时')
               #html = response.content.decode('utf-8')
               #soup = BeautifulSoup(html, 'html.parser')
               #print(soup)
               # 歌单的封面图片 (需把图片保存到本地)、歌单标题、创建者 id、创建者昵称、介绍、歌曲数量、
               # 歌单标题
               title = soup.find('h2',class_ = 'f-ff2 f-brk').text
               #print(title)
               # 图片下载
               img_info = soup.find('div',class_='cover u-cover u-cover-dj')
               img_url = img_info.find('img',class_='j-img')['data-src']
               img_id = id_list[i][-10:]
               download_img(img_path,title,img_url)
               #print(img_url)
               # 创建者昵称
               create_name = soup.find('span',class_='name').text.strip('\n')
               #print(create_name)
               # 创建者id
               create_id = soup.find('span',class_='name').select('a')[0]['href'][14:]
               #print(create_id)
               # 介绍
               describe = soup.find('p',class_='intr f-brk').text.replace('\n','\t').replace(
               #print(describe)
               # 歌曲数量
               song_num = soup.find('span','sub s-fc3').find('span').text
               #print(song_num)
               #播放量
               play_num = soup.find('strong','s-fc6').text
               #print(play_num)
               #添加到播放列表次数
               add_num = soup.find('a','u-btni u-btni-fav')['data-count']
               #print(add_num)
               # 分享次数
               share_num = soup.find('a', 'u-btni u-btni-share')['data-count']
               #print(share_num)
               # 评论数
               comment_num = soup.find('a','u-btni u-btni-cmmt').find('span').text
               #print(comment_num)
               with open(save_path+'说唱.csv','a+') as f:
                   f.write(id+','+title+','+create_name+','+create_id+','+str(describe)+','+s
# 3. 要求使用生产者-消费者模式实现,要求1作为生产者,每次请求后产生新的任务交给消费者,消费者执行要求2。
if __name__ == '__main__':
   q = queue.Queue()
    producer_list = []
   consumer_list = []
   with open(save_path+'说唱.csv','a+') as f:
       f.write('歌曲 ID'+','+'歌单标题'+','+'创建者昵称'+','+'创建者id'+','+'介绍'+','+'歌曲数量'+
```

response = requests.session().get(url=url,headers=headers_2)

```
sheet = list(range(0, 1300, 35))
   p_t = produce_id(sheet,q)
   p_t.start()
   p_t.join()
   thread_num = 38
   '''num = len(sheet)//thread_num
   for i in range(thread_num-1):
       p_t = produce_id(sheet[num*i:num*(i+1)],q)
       producer_list.append(p_t)
   p_t = produce_id(sheet[num*(thread_num-1):],q)
   producer_list.append(p_t)'''
   for i in range(thread_num):
       c_t = consumer(q)
       consumer_list.append(c_t)
   '''for p in producer_list:
       p.start()
   for p in producer_list:
       p.join()'''
   for c in consumer_list:
       c.start()
   #p_t.join()
   for c in consumer_list:
       c.join()
   '''produce_thread = produce_id([0],q)
   produce_thread.start()
   produce_thread.join()
   consumer_thread = consumer(q)
   consumer_thread.start()
   consumer_thread.join()'''
# 4. (附加) 爬虫程序往往需要稳定运行较长的时间,因此如果你的程序突然中 断或异常 (比如网络或被封) ,如何能够
# 保留断点处的信息,包括断点处的文件保存路径,以及接下来要爬的网页地址,和其他在程序中需要保留的所有变量,
# 5. (附加) 爬虫程序往往需要比较友好的状态输出,因此可否专门有一个线程 动态地进行输出更新,来显示当前的状态
# 比如程序连续运行的时长,要完成的 总页面数,其中有多少已被爬取,已收集的文件占用了多少空间,大概还需要多少
                                                                         JavaScript ~
```

结果展示

```
tap『压马路』.jpg
     ·嘻哈风格 | 乘风破浪 万物皆可说唱..jpg
     NAP BEAT ◇精选 | 说唱伴奏.jpg
     > CHILL BEAT ◆高质量 | 说唱伴奏.jpg
     BEAT> 倚栏听风雨 无关风与月.jpg
     BEAT>情歌·制作人的浪漫.jpg
FREE INSTRUMENT"免费伴奏|DAWN MUSIC出品.jpg
    "好听到爆炸的说唱".jpg
"rap 讲情话 悸动的感觉.".jpg
     你真的甘心错过吗?".jpg
     你能容纳世界 难道容不下我".jpg
     '别唱了宝贝,我好了"情歌说唱.jpg
                                                                                                                                                                       Dec 15, 2021 at 3:39 PM
                                                                                                                                                                                                      123 KB JPEG image
₹ "我的世界只有你".jpg 增掉进陷阱也不影响我摇起来"trap说唱.jpg
■ "教堂的白鸽不会亲吻乌鸦";pg
図 "此刻我只想听一些安静的歌"失眠歌单.jpg
图 "甜味 旋律 Rap情歌";pg
                                                                                                                                                                                                       46 KB JPEG image
133 KB JPEG image
                                                                                                                                                                                                       90 KB JPEG image
75 KB JPEG image
── "这就是甜到让你发麻的说唱吗?".jpg
图 "个性"说唱伴奏.jpg
② (Rap说唱)跟着节奏狂奔 开车时听的神仙歌.jpg
                                                                                                                                                                                                       68 KB JPEG image
☑ [Popping 练习曲] 机械律动 节奏强袭.jpg
❷ [RAP]中文励志说唱-为梦想努力奋斗.jpg
                                                                                                                                                                                                       65 KB JPEG image
129 KB JPEG image
                                                                                                                                                                      Dec 15, 2021 at 3:39 PM
                                                                                                                                                                                                       53 KB JPEG image
57 KB JPEG image
 🔣 [Z世代说唱] 在节奏中感受Z世代的嘻哈力量.jpg
[一周欧美说唱] 枪姐联手 Push拉开新专序幕,果汁遗作重现旋律荣光.jpg
                                                                                                                                                                      Dec 15, 2021 at 3:39 PM
   [中文旋律说唱] 旋律与节奏撞出轻松惬意的时光.jpg
☑ [宝藏歌单]国产精选Rap.jpg
☑ [打游戏听说唱] 说唱律动是你最佳游戏伴侣.jpg
                                                                                                                                                                                                       86 KB JPEG image
                                                                                                                                                                      Dec 15, 2021 at 3:39 PM
                                                                                                                                                                                                      880 KB JPEG image
56 KB JPEG image
 [图 [拒绝瞌睡] 跟着节拍抖腿的嘻哈音乐.ipg
                                                                                                                                                                      Dec 15, 2021 at 3:39 PM
图 [爵士说唱] 一场浪漫的化学反应 jpg 医 [远离沉闷与懒惰]说唱就该跟着摇 jpg
    [爵士说唱]一场浪漫的化学反应.jpg
圖 (雷击顿唱片行) 让你扭动身体的异域节奏,jpg
图 《NBA2K20》 手机版歌单,jpg
图 《少年说唱企划》 注:
                                                                                                                                                                                                       94 KB JPEG image
79 KB JPEG image
                                                                                                                                                                      Dec 15, 2021 at 3:39 PM
                                                                                                                                                                                                       186 KB JPEG image
82 KB JPEG image
                                                                                                                                                                      Dec 15, 2021 at 3:39 PM
    《说唱新世代》学员精选作品集.jpg
《说唱新世代》第一季成员battle战炸曲收录.jpg
                                                                                                                                                                                                       64 KB JPEG image
176 KB JPEG image
                                                                                                                                                                      Dec 15, 2021 at 3:39 PM
    《高等rapper4》| 比赛 Beat BGM 音源大收录.jpg
《黑怕女孩》选手精选作品集.jpg
                                                                                                                                                                                                       89 KB JPEG image
                                                                                                                                                                                                       72 KB JPEG image
99 KB JPEG image
     「Jazz Rap 」你会跳舞吗,我不会.jpg
     「Rap song 」hey 要 点 首 说 唱 吗.jpg
                                                                                                                                                                      Dec 15, 2021 at 3:39 PM
     「社恐必备 」社交牛逼症养成走路生风系列.jpg
                                                                                                                                                                                                       86 KB JPEG image
     「Hip-hop研究所」潮酷出街 | 开启你的 Swag.jpg
                                                                                                                                                                      Dec 15, 2021 at 3:39 PM
     「中国风说唱伴奏精选集」Rap Beat.jpg
    「中文说唱」因你而在的梦.jpg
```

歌曲ID 歌单标题	创建者昵称	创建者id 介绍		歌曲数量	播放量	添加到播放	分享次数	评论数	
4.922E+09 Hiphop说唱宝藏库	Hiphop说唱宝藏库	474348309 介绍	: 只	349	2667731	37229	261	61	
7.106E+09 "如何融化 你的冰川"	L07o	540211076 介绍	: 我	52	10469	267	233	2	
4.996E+09 免费 beat 情歌专用	SEA · JI	2.115E+09 介绍	: 免	860	160825	2338	43	5	
3.12E+09 国产hiphop百听不厌	Double_Mini	421421612 介绍	: (接	452	21008536	125948	4332	462	
6.832E+09 粉红法拉利贝贝	Dokeiii	1.724E+09 介绍			100916	203	4	评论	
2.991E+09 别见 别想 别喜欢	TWENTY-Y	126913104 介绍	: 你!	184	739118	10895	790	448	
5.05E+09 中国顶尖rap歌单	MorniN9_Star	1.39E+09 介绍	: 绝	69	468105	5754	20	18	
5.469E+09 BoomShake杭州总店 HIPHOP热单	-Jackson-X-	386332059 介绍	: 杭	96	26983	542	32	3	
3.195E+09 《NBA2K20》 手机版歌单	TottenhamHotspurs钟爱我刺	1.75E+09 介绍	: NE	23	59071	1063	12	36	
5.306E+09 旋律说唱 余音绕梁,三日不绝	香蕉之城	415271837 介绍	: 这	25	5668	2353	0	1	
5.365E+09 免费drill beat 伴奏,拿走做皇家driller	DRILLMUSIC	2.139E+09 介绍			49894	699	8	5	
6.84E+09 再狠的rapper在孤独de夜晚也会emo	苏方一杰	335703712 介绍	: 孤	45	521210	9128	2961	14	
7.075E+09	220)7607 1.803E+09 介绍			116	2	0	评论	
5.07E+09 齁甜的情歌说唱歌单啊	illusory铁塔猫	1.674E+09 介绍	: 甜;	203	1285808	1189	14	2	
6.583E+09 《黑怕女孩》选手精选作品集	偷偷的叹了口气	1.591E+09 介绍	: «	100	350996	3626	48	67	
5.368E+09 【说唱】 为了抒发来自内心深处的斥责怒吼	UKnlo	2.088E+09 介绍	: [1	232	380102	3986	237	617	
6.823E+09 说唱 忘掉不开心,做愉快的梦	少女的喜欢_	3.277E+09 介绍	: 大	51	654598	719	3	3 评论	
6.724E+09 Emotional Rap:慢情绪节奏艺术	G-DDRRAAGGOONN	453482654 介绍	: em	18	41003	312	5	评论	
5.379E+09 说唱新世代第一季	UKnlo	2.088E+09 介绍	: 说	47	30452	721	1	评论	
5.039E+09 抒情说唱伴奏 Chill beat	拯救wakeup	476114092 介绍	: 风	256	138926	3100	122	53	
6.961E+09 『驾驶』露营/ 买菜/ 炸街副驾专属Hiphop	Ceaoee	1.368E+09 介绍	: 新	52	4992	98	1	1	
5.097E+09 404 RAPPER HNBMC 唯有喜欢	大概是个好人a	1.513E+09 介绍	: HN	121	3694	86	1	6	
2.581E+09 催眠麦克风DRB/TDD全曲收入	yaamatokai	564646179 介绍	: DR	101	628679	17187	150	82	
6.751E+09 浪漫Rap 开口心动的说唱情话	少女的喜欢_	3.277E+09 介绍	: 最	86	853165	1476	12	4	
7.063E+09 HipHop Club常用	Big_Shun	71983390 介绍	: 不	126	229	7	0	评论	
6.682E+09 Lil Nas X (李娜叉) 无重复收录	子半天胜	431018448 介绍	: 李	29	106674	1136	6	17	
6.885E+09 【情歌rap】想和你陷入浪漫的无限循环	少女的喜欢_	3.277E+09 介绍	: 这	70	225135	851	5	评论	
6.854E+09 甜味说唱 微醺hiphop	饭醉小熊	3.311E+09 介绍	: Skr	97	2703	139	1	评论	
4.859E+09 酷盖说唱【蹦迪必备】	MONEY-HONEY-	1.999E+09 介绍	: 节	105	245455	2312	37	28	
6.654E+09 Beattle beat 上等好货 freestyle 伴奏	一个EDA	374150187 介绍	: Ва	34	14833	2407	1	评论	
6.611E+09 国产hiphop从这里开始	Zxy在向下扎根	443546368 介绍	: 说!	50	274737	7434	2780	7	
6.817E+09 FullHouse满堂2021	FullHouse满堂	4.933E+09 介绍	: Ful	133	5257	231	14	32	
6.986E+09 唯有一腔诗意浓的歌单	唯有一腔诗意浓	334480710 介绍	. 40	13	4618	9	0	评论	