

20377199 赵芮箐 第15周作业

作业内容： 本周要求在第14周的网络爬虫基础上设计用于存储爬取内容的数据库。由于爬取内容包括文本这样的非结构化数据，且数据间没有复杂的关系结构，为了操作简便，采用结构（PostgreSQL）或非结构化数据库（mongodb）对数据进行处理。具体要求如下：

1. 安装所选择数据库，并启动数据库，设用户以及密码，供后续存储文档使用
2. 实现python对数据库的连接
3. 创建数据表（集合） voa，在已有的爬虫系统中补充相关功能，将爬取到的数据存入数据库，要求包括如下字段：ID, title, text, author, date, mp3_path（对mp3文件，在表中只存储路径即可），related_articles。其中ID可在url中获取，title, author, date, related_articles等内容可在数据页面内找到。
4. （附加）编写相关python函数，实现基于数据库的简单查询功能(可以加入其他搜索字段，例如时间，作者等，对应通用的网站搜索功能)。

总体思路

- 又又又很摸鱼的一周哈哈哈，这周真的要求 is 真的很少
- 主要是在14周的基础上把信息写入文件改成了写入数据库
- 实现了简单的查询功能

数据库连接：

```
if __name__=='__main__':
    async_time_start = time.time()

    p_jobs = []
    p = pool.Pool(20)
    for i in range(0, 1300, 35):
        url = 'https://music.163.com/discover/playlist/?cat=流行
&order=hot&limit=35&offset=' + str(i)
        p_jobs.append(p.spawn(producer, url))
    gevent.joinall(p_jobs)

    # 话说我把游标当参数传给consumer，就不需要异步连接了啊
    with psycopg2.connect(dsn) as conn:
        with conn.cursor() as cur:
            c_jobs=[p.spawn(consumer, id, cur) for id in id_list]
            gevent.joinall(c_jobs)
        conn.commit()

    print("总耗时: %.4f" %(time.time()-async_time_start))
```

爬虫部分：

```
id_list = []
dsn = "dbname=zrq user=postgres password=20020909ZRQ"
headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/63.0.3239.132 Safari/537.36'
}

def producer(url):
    response = requests.get(url=url, headers=headers)
    html = response.text
    soup = BeautifulSoup(html, 'html.parser')
    ids = soup.select('.dec a') # 获取包含歌单详情页网址的标签
```

```

id_list.extend(ids)

def consumer(id, cur):
    url = 'https://music.163.com/' + id['href'] # 获取歌单详情页地址
    response = requests.get(url=url, headers=headers)
    html = response.text
    soup = BeautifulSoup(html, 'html.parser')
    # 获取ID
    ID = id['href'].split('=')[-1]
    # 获取歌单标题
    title = soup.find('h2', "f-ff2 f-brk").text
    # 获取歌单介绍
    try:
        text = soup.find('p', id="album-desc-more").text.replace('\n', '')
    except:
        text = '无'
    # 获取创建者昵称
    author = soup.find('a', "s-fc7").text
    # 获取创建日期
    date = soup.find('span', 'time s-fc4').text[:10]
    #print(ID,title,text,author,date)

    # 将详情页信息写入数据库中
    cur.execute("INSERT INTO music VALUES (%s, %s, %s, %s, %s)",
        (ID,title,text,author,date))

```

PS: 其实相较于上周就改了最后一句话...

PPS: 我着实没懂 mp3_path 和 related_articles 是啥

数据查询：

```

import datetime
import psycopg2
import psycopg2.extras

dsn = "dbname=zrq user=postgres password=20020909ZRQ"

def select(keyword, mode, sheet='music'):
    if mode == 'title':
        command = f"select * from {sheet} where {mode} like '%{keyword}%"
    else:
        command = f"Select * from {sheet} where {mode} = '{keyword}%"
    with psycopg2.connect(dsn) as conn:
        with conn.cursor(cursor_factory=psycopg2.extras.DictCursor) as dict_cur:
            dict_cur.execute(command)
            res = dict_cur.fetchall()
            music = [r['title'] for r in res]
    return music

print('-----按ID-----')
for music in select('4891546330', 'ID'): print(music)
print('-----按title-----')
for music in select('夏天', 'title'): print(music)
print('-----按author-----')

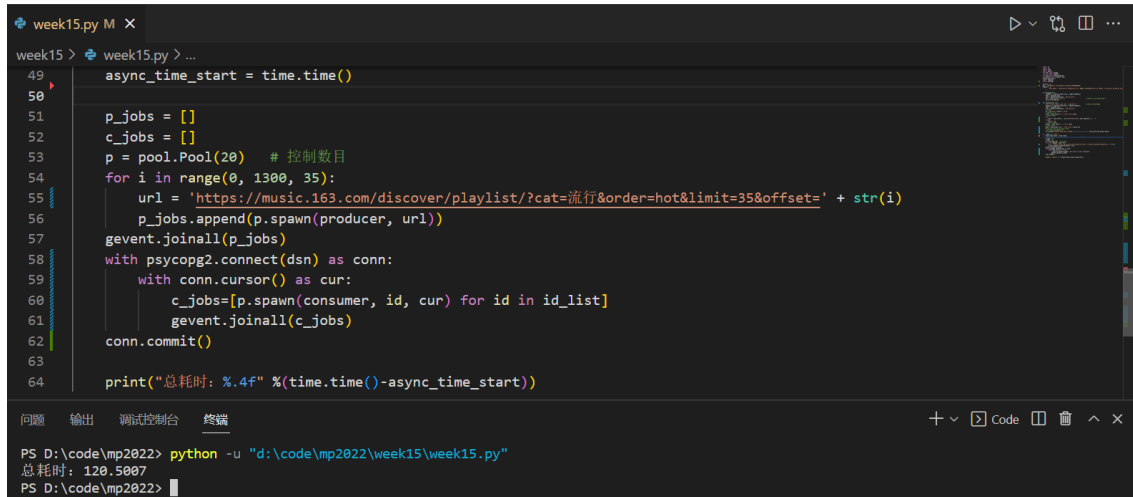
```

```
for music in select('网易云音乐', 'author'): print(music)
print('-----按date-----')
for music in select(datetime.date(2020,5,20), 'date'): print(music)
```

PS: 通用网站搜索结果貌似就是歌单名, 所以这里只show歌单名

结果展示

- 运行时间:

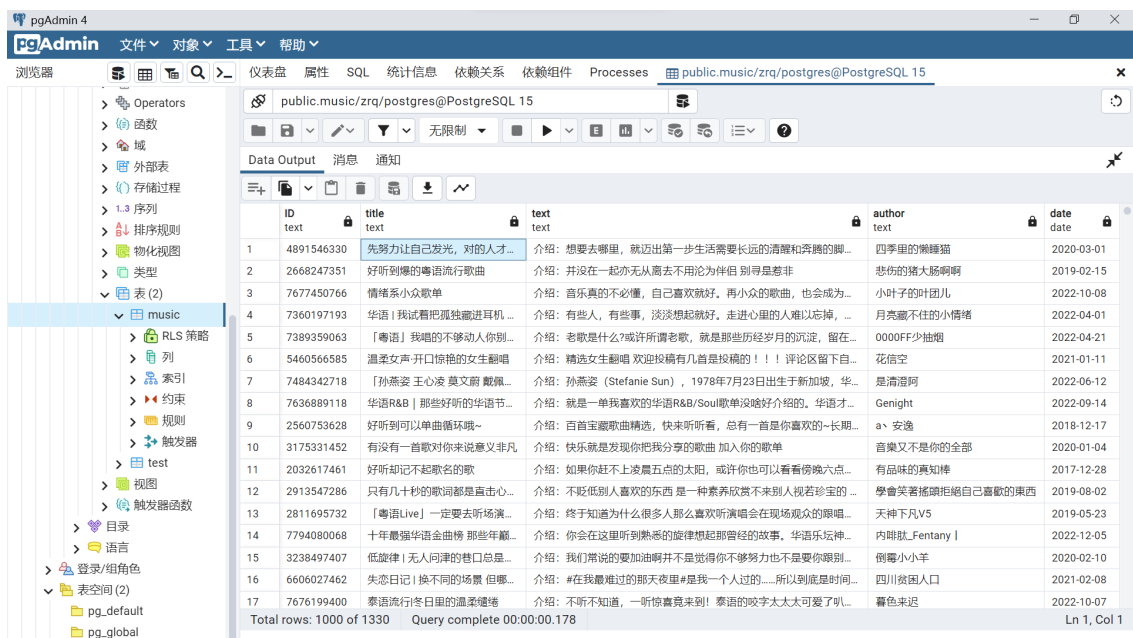


```
week15.py M X
week15 > week15.py > ...
49  async_time_start = time.time()
50
51  p_jobs = []
52  c_jobs = []
53  p = pool.Pool(20) # 控制数目
54  for i in range(0, 1300, 35):
55      url = 'https://music.163.com/discover/playlist/?cat=流行&order=hot&limit=35&offset=' + str(i)
56      p_jobs.append(p.spawn(producer, url))
57  gevent.joinall(p_jobs)
58  with psycopg2.connect(dsn) as conn:
59      with conn.cursor() as cur:
60          c_jobs=[p.spawn(consumer, id, cur) for id in id_list]
61          gevent.joinall(c_jobs)
62  conn.commit()
63
64  print("总耗时: %.4f" %(time.time()-async_time_start))

问题 输出 调试控制台 终端
PS D:\code\mp2022> python -u "d:\code\mp2022\week15\week15.py"
总耗时: 120.5007
PS D:\code\mp2022>
```

跟上一次写入文件耗时差不多

- 数据库文件



ID	title	text	author	date
1	4891546330	先努力让自己发光, 对的人才...	介绍: 想要去哪里, 就迈出第一步生活需要长远的清醒和奔跑的脚...	四季里的懒猫
2	2668247351	好听到爆的粤语流行歌曲	介绍: 并没在一起亦无从离去不用沦为伴侣 别寻是是非	悲伤的猪大肠啊
3	7677450766	情绪系小众歌单	介绍: 音乐真的不必懂, 自己喜欢就好。再小众的歌曲, 也会成为...	小叶子的叶团儿
4	7360197193	华语 我试着把孤独藏进耳机...	介绍: 有些人, 有些事, 淡淡想起就好。走进心里的人难以忘掉, ...	月亮藏不住的小情绪
5	7389359063	【粤语】我唱的不够动人你别...	介绍: 老歌是什么?或许所谓老歌, 就是那些历经岁月的沉淀, 留在...	0000FF少抽烟
6	5460566585	温柔女声 开口惊艳的女生翻唱	介绍: 精选女生翻唱 欢迎投稿有几首是投稿的!!! 评论区留下自...	花信空
7	7484342718	【孙燕姿 王心凌 莫文蔚 戴佩...	介绍: 孙燕姿 (Stefanie Sun), 1978年7月23日出生于新加坡, 华...	是清泪阿
8	7636889118	华语R&B 那些好听的华语节...	介绍: 就是一单我喜欢的华语R&B/Soul歌单没啥好介绍的。华语才...	Genight
9	2560753628	好听到可以单曲循环哦~	介绍: 百首宝藏歌曲精选, 快来听听看, 总有一首是你喜欢的~长期...	a、安逸
10	3175331452	有没有一首歌对你来说意义非凡	介绍: 快乐就是发现你把我分享的歌曲 加入你的歌单	音乐又不是你的全部
11	2032617461	好听却记不起歌名的歌	介绍: 如果你赶不上凌晨五点的太阳, 或许你也可以看看傍晚六点...	有味道的真知棒
12	2913547286	只有几十秒的歌词都是直击心...	介绍: 不贬低别人喜欢的东西 是一种素养欣赏不来别人视若珍宝的 ...	學會笑著搖頭拒絕自己喜歡的東西
13	2811695732	【粤语Live】一定要去听场演...	介绍: 终于知道为什么很多人那么喜欢听演唱会会在现场观众的限唱...	天神下凡V5
14	7794080068	十年最强华语金曲榜 那些年最...	介绍: 你会在这里听到熟悉的旋律想起那曾经的故事。华语乐坛神...	内附就_Fantasy
15	3238497407	低旋律 无人问津的巷口总是...	介绍: 我们常说的要加油啊并不是觉得你不努力也不是要你跟别...	倒霉小小羊
16	6606027462	失恋日记 换不同的场景 但哪...	介绍: #在我最难过的那天夜里#是我一个人过的.....所以到底底是时间...	四川贫困人口
17	7676199400	粤语流行冬日里的温柔情绪	介绍: 不听不知道, 一听惊喜竟来到! 粤语的咬字太太太可爱了叭...	暮色来迟

Total rows: 1000 of 1330 Query complete 00:00:00.178 Ln 1, Col 1

- 数据查询

```

-----按ID-----
先努力让自己发光，对的人才会迎着光而来
-----按title-----
毕业季 | 青春怀旧纪念手册 那年夏天
希望你最后的成绩能点亮整个夏天
平井 大 | "夏天本夏"的海洋系创作歌手
“十七八岁的夏天我现在还能记得”
夏天的日子 最炙热而浪漫
华语流行 | 夏天 大海 汽水 音乐 和你
整个夏天，想带你环游世界
18年的夏天，除了有点热，其他都很好
-----按author-----
歌手毛不易 | 谢谢你唱，我们会一直听
灵魂歌手Adele回归 | 空灵嗓音诠释纯粹深情
温柔岁月 | 若梦想有方向，平凡的日子也泛着光
WayV | 星光不问赶路人，时光不负有心人 ✨
罗大佑宝藏曲库 | 我们会永远需要罗大佑
滚石唱片 | 青涩岁月里的五月天
网易云音乐X爱贝克思 | 打卡JPOP的时代记忆
平井 大 | "夏天本夏"的海洋系创作歌手
滨崎 步 | 永远闪耀的亚洲天后
大塚愛 | 日本乐坛萌甜型创作才女
2018上半年最热新歌TOP50
歌手华晨宇 | 歌王诞生，热血开唱！
张杰金曲 | 你的青春一定听过这些歌
以音乐为桥 | 综艺《时光音乐会》音频收录
2021上半年度最热新歌TOP50
歌手周深 | 破茧成蝶，全新风格动人心魄
-----按date-----
「宝藏热歌」网易云不容错过的热歌精选
翻唱 | 我为你翻山越岭 却无心看风景

```

代码：

https://github.com/rachhhing/mp2022_python/blob/master/week15

Ref:

- pyscopg2: <https://www.psycopg.org/docs/>
- sql语法: <https://www.runoob.com/sql/sql-syntax.html>

程设课程完结撒花 ✨,°.☆(￣▽￣)/\$.°★。

感谢辛苦授课的赵老师 and 辛苦批改作业和答疑解惑的两位助教姐姐(づ￣ 3￣)づ