

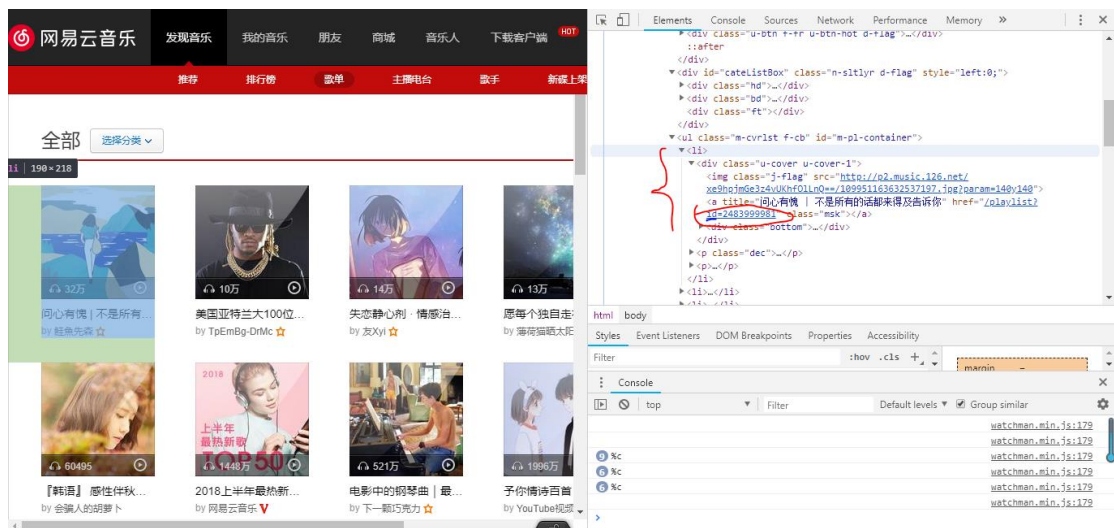
爬虫新手：网易云歌单爬取

1. 本次爬取主要使用 BeautifulSoup 库，Selenium 库，正则表达式以及 traceback 库

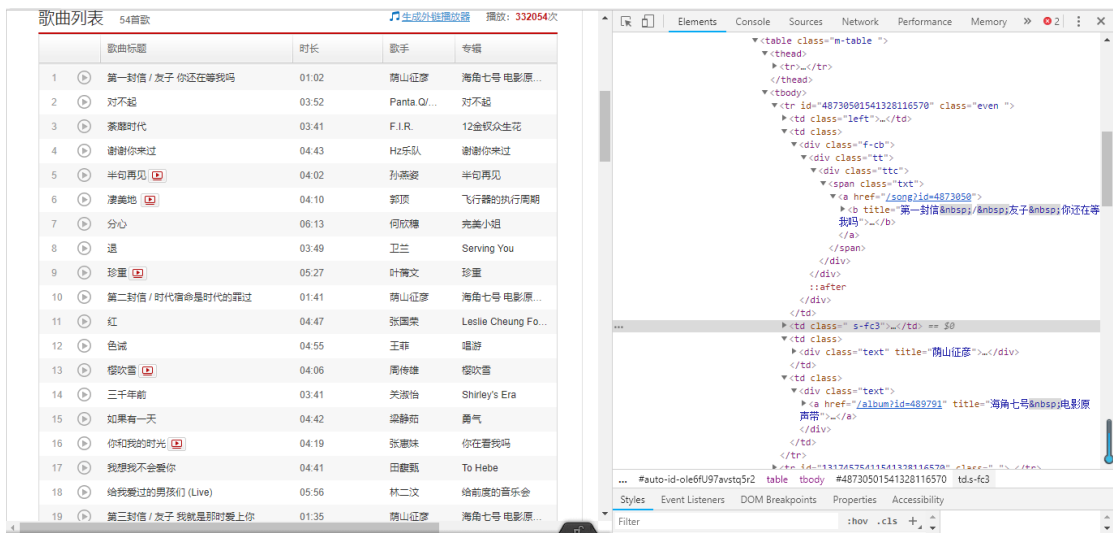
首先，关于 BeautifulSoup 库的安装：

```
C:\Users\71434>pip install beautifulsoup4
Collecting beautifulsoup4
  Downloading https://files.pythonhosted.org/packages/21/0a/47fdf541c97fd9b6a610cb5fd518175308a7cc60569962e776ac52420387/beautifulsoup4-4.6.3-py3-none-any.whl (90kB)
    | 30kB 3.2kB/s eta 0:00:1
    | 40kB 2.4kB/s eta 0:
    | 51kB 2.2kB/s eta
    | 61kB 2.6kB/s
    | 71kB 2.5
    | 81kB
    | 9
2kB 2.6kB/s
Installing collected packages: beautifulsoup4
Successfully installed beautifulsoup4-4.6.3
```

2. 关于爬取的前期准备：



在网易云音乐的歌单界面可以进行开发者选项查看网页的结构，可以发现网易云歌单是存放在类名 class=“u-cover u-cover-1”的<div>标签中且 href 属性中可以看出来每个歌单都有一个属于自己的 id，并且使用 url=’<https://music.163.com/#/playlist?id=2483999981>’能够跳转到自己的歌单界面。



然后在如图所示的标签中就能够找到自己所需要的歌单内容。

我们在歌单的首页点击翻页就可以发现每个页面有 35 个歌单，这样就可以进行所有歌单的 id 的获取：

← → ↻ 🔒 安全 | <https://music.163.com/#/discover/playlist/?order=hot&cat=全部&limit=35&offset=0>

← → ↻ 🔒 安全 | <https://music.163.com/#/discover/playlist/?order=hot&cat=全部&limit=35&offset=35>

3. 爬取函数的定义：

首先，建立 getHTMLText(url) 函数用于进行网页的信息获取；

建立 getSongList(slist, url) 函数用于获得歌单 id 信息

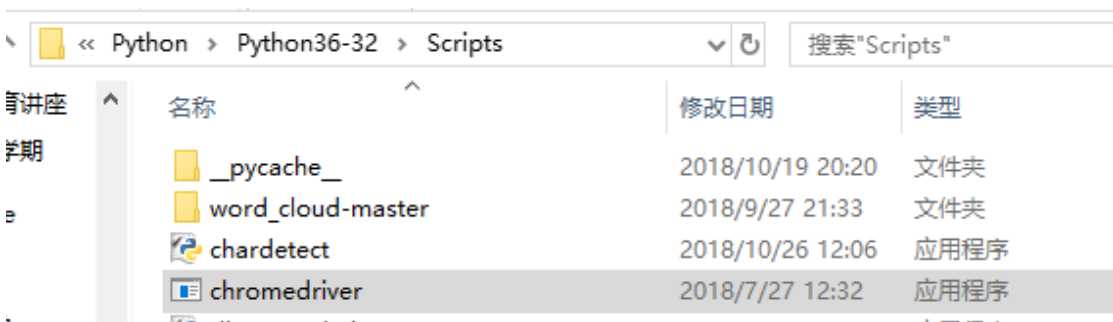
建立 writeSongInfo(slist, fpath) 函数用于获得歌单的详细信息，并将所获得的歌单存入文件中。

4. 爬取

第一步就遇到了困难，因为在网易云歌单的源代码中并不能找到歌曲信息，这是一个动态页面！

```
C:\Users\71434>pip install Selenium
Collecting Selenium
  Downloading https://files.pythonhosted.org/packages/80/d6/429f13e17190289f9d0613b0a44e5dd6a7f5ca98459853/selenium-3.141.0-py2.py3-none-any.whl (904kB)
    29% |#####| 266kB 28kB
```

安装 Selenium，下载 chromedriver 准备进行动态页面爬取：



```

def getHTMLText(url):
    try:
        r=webdriver.Chrome()
        r.get(url)
        r.switch_to.frame('g_iframe')
        return r.page_source
    except:
        return ""

```

5. 将数据存入文件:

```

        continue
def writeSongInfo(lst,fpath):
    for sid in lst:
        try:
            with open(fpath, 'a',encoding = 'utf-8') as f:
                f.write(str(sid) + '\n')
        except:
            traceback.print_exc()
            continue

```