网易云数据爬取

1. 数据爬取目的

本小组的所做的项目是音乐推荐系统，为了能够给用户推荐歌曲，我们需要大量歌曲的各项信息，这样确保我们能够根据歌曲的信息与用户的喜好的相似度来给用户推荐歌曲。以上是确保系统能够在运行初期、用户数据较少、不足以支持协同过滤算法时能够正常运行的关键。因此从哪里获取信息、需要获取哪些信息等问题是我们组工作初期的关键问题。

2. 目标信息

经过本小组讨论后，我们决定在网易云音乐上获取我们需要的信息，我们认为对实现推荐系统有着重要作用的是以下信息：

歌曲名：作为给用户推荐的结果

歌曲作者：用于匹配用户所喜欢的音乐家

歌曲类型：用于匹配用户所喜欢的歌曲风格

播放量：用于分析歌曲流行程度

3. 数据爬取步骤

(1)初始化信息（要爬取的信息，页数等）

(2)填写请求头

(3)Froms 构造表格

(4)获取页面源码

(5)匹配要爬取的数据

(6)数据处理（删除低播放量歌单等）

(7)将歌单数据保存到 csv 文件中，提取歌单链接，然后循环爬取歌单中歌曲信息

部分代码如下：

```python
from urllib import parse
from lxml import etree
from urllib3 import disable_warnings
import requests
import csv


class Wangyiyun(object):

    def __init__(self, **kwargs):
        # 歌单风格
        self.types = kwargs['types']
        # 歌单类型
        self.years = kwargs['years']
        # 爬取十页
        self.pages = pages
        # 页数
        self.limit = 35
        self.offset = 35 * self.pages - self.limit
        # url
        self.url = "https://music.163.com/discover/playlist/?"

    # 设置请求头部信息
    def set_header(self):
        self.header = {
            "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/73.0.
            "Referer": "https://music.163.com/",
            "Upgrade-Insecure-Requests": '1',
        }
        return self.header

    # 设置请求表格信息
    def set_froms(self):
        self.key = parse.quote(self.types)
        self.froms = {
            "cat": self.key,
            "order": self.years,
            "limit": self.limit,
            "offset": self.offset,
        }
        return self.froms
```

```python
    # 解析代码，获取有用的数据
    def parsing_codes(self):
        page = etree.HTML(self.code)
        # 标题
        self.title = page.xpath('//div[@class="u-cover u-cover-1"]/a[@title]/@title')
        # 作者
        self.author = page.xpath('//p/a[@class="nm nm-icn f-thide s-fc3"]/text()')
        # 阅读量
        self.listen = page.xpath('//span[@class="nb"]/text()')
        # 歌单链接
        self.link = page.xpath('//div[@class="u-cover u-cover-1"]/a[@href]/@href')
        # 将数据保存为csv文件
        data = list(zip(self.title, self.author, self.listen, self.link))
        with open('yinyue.csv', 'a', encoding='utf-8', newline='') as f:
            writer = csv.writer(f)
            # writer.writerow(header)
            writer.writerows(data)

    # 获取网页源代码
    def get_code(self):
        disable_warnings()
        self.froms['cat'] = self.types
        disable_warnings()
        self.new_url = self.url + parse.urlencode(self.froms)
        self.code = requests.get(
            url=self.new_url,
            headers=self.header,
            data=self.froms,
            verify=False,
        ).text

    # 爬取多页时刷新offset
    def multi(self, page):
        self.offset = self.limit * page - self.limit


if __name__ == '__main__':
    # 歌单的歌曲风格
    types = "说唱"
    # 歌单的发布类型:最热=hot，最新=new
    years = "hot"
    # 指定爬取的页数
    pages = 10
    # 通过pages变量爬取指定页面
```

4. 数据爬取结果

结果中包括以下信息：歌单名、歌单作者、歌单播放量、歌单链接、歌曲名、歌曲作者、时长、所属专辑。

部分结果如下所示：

247	「欧美嘻哈」Flexing大佬们带你随时起飞	-Pony_Boi-	58357
248	Trap丨不客观毒性幻觉单	Pomegranatea	57563
249	欧美新潮流行说唱New Wave Rap	Juden_Wittgenstein	53720
250	「复古经典」东西海岸的OG们	-Pony_Boi-	49252
251	Lil Tecca ☢	Dot_RhAM	48953
252	『史上最佳25首Hip-Hop歌曲』	RealGoldLit	47811
253	经典Hip-hop伴奏	lolliporn	41944
254	精选说唱伴奏‖freestyle beats	ZiJianHan	36962
255	tom macdonald	鸽补法	36715
256	2019年度最佳说唱	MeloNoLove	35182
257	『Trap』炸裂硬核重金属风的 “噪音说唱”	暖色乐章	34367
258	毛子专属 hardbass	O_gtr	33649
259	Pulse\\All.Hiphop All.Night（周日更新）	TX3iX	32292
260	RAP TRAP day	_joshkwan	30914
261	Roddy Ricch	humorous-sliver	27176
262	$殺気爆棚 超凶trap$	psy_扭不开的奥利奥	27000
263	Uyghur Rap丨精选说唱集合	楼兰Music	25568
264	New School Hiphop/R&B	Yandhiii	25379
265	China Mac 亇【华裔匪帮rapper NO.1】	黑白鱼眼	24934
266	英式说唱的进化UK drill	Juden_Wittgenstein	24728
267	JalalEnwer歌曲集	Jalal-Anwar	24270
268	欧美丨真诚说唱，被遗忘在世外仙境的珠宝	IiIuzivert	24131
269	【HIPHOP说唱】必 听 的 经 典 钻 石 单 曲	迷人的混蛋是西西	21868
270	张颜齐循环一百次	Fivvvve	21065
271	硬核 蒙古说唱	BaynErhet	18919
272	「欧美嘻哈」迅速崛起的超新星们	-Pony_Boi-	18709
273	公主rap&后妈联盟	_山鬼_乔乔	18596
274	Young-Nash Nashtarr【美国维吾尔说唱歌手	PRADA-802	15064
275	2019十首最炸flow说唱(英文)	1LIFT	14131
276	NiuLai_Music Free beat合集	NiuLai_Music	12707
277	CHN饶社 原创说唱作品合集	CHN桉树	11989

5. 缺陷与不足

本次数据爬取工作未能成功获取歌曲的类型信息，接下来我们将会讨论一个可行方案来解决此问题。