**题目：基于周杰伦歌词的情感分析与歌词生成**

小组成员：刘志伟 董琛扬

1. **背景**

6月12日零点，周杰伦新歌《MOJITO》首发，一上线就让音乐平台瘫痪，2000年出道的他，20年后还拥有如此高的人气和关注度，这与他的歌曲曲风和词风是分不开的。而歌词是一首歌曲的灵魂所在，整段文字承载了一首歌的感情基调与主旋律。要了解一位歌手，对于其用字、用词风格的深入挖掘是至关重要的。

本项目将基于python语言，利用数据挖掘可视化方法、深度学习方法，尝试深入挖掘周杰伦歌曲的用字、用词习惯并构建歌词情感预测模型和歌词生成器。

1. **项目要求**

**Task 1** 收集相关数据，可视化周杰伦歌词用字、用词习惯；

**Task 2** 对歌词进行情感分析，构建歌词情感预测模型；

**Task 3** 构建“周氏词风”歌词自动生成器

1. **项目实现流程与结果**

(一) 数据获取与预处理

1. 数据来源

本项目中所用到的数据来自qq音乐官方网站，通过官网的搜索引擎可以获取任意歌手的歌曲、所属专辑名以及相关歌词等信息。

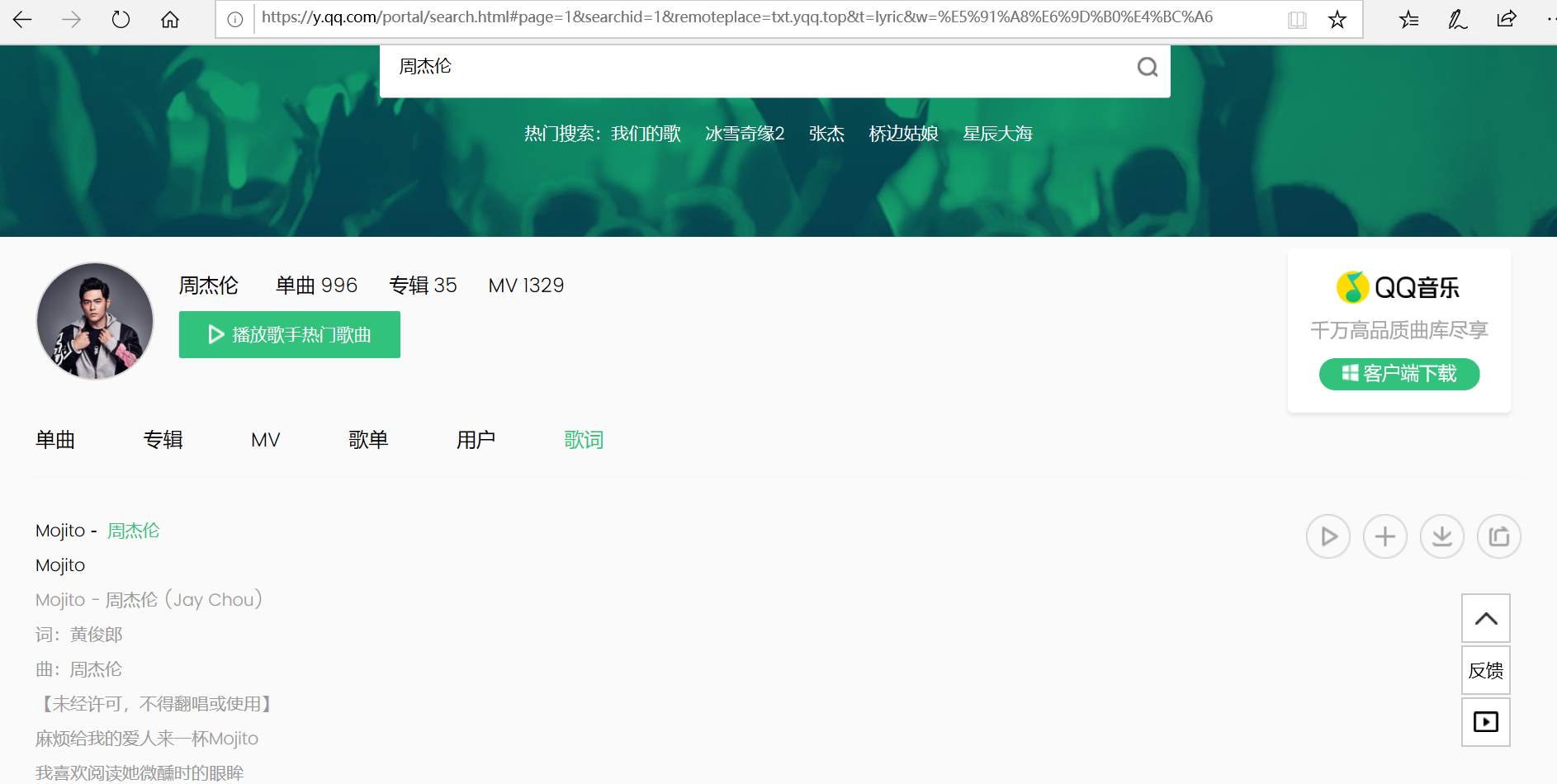


图1.QQ音乐搜索引擎

2. 数据爬取与解析

爬虫技术是一种利用程序自动访问指定浏览器的数据获取技术，而python中的requests和json库分别用于初步爬取网页源码信息和对爬取到的信息进一步解析，两者的配合能够实现爬取指定的目标数据。

基于python语言，爬取网页中的数据步骤如下：

Step 1 获取网址和用户设备、浏览器信息

网址信息在网页搜索框中即可获取；用户设备、浏览器信息可通过在目标网页下，点击F12通过开发者工具获取。



图2.获取QQ音乐官网网址



图3.获取QQ音乐搜索引擎所在网页网址和用户相关信息

Step 2通过requests库爬取网页源码

基于上一步获取到的两个网址和用户信息，通过requests.get（）函数，初步获取QQ音乐中所有周杰伦歌曲相关信息。

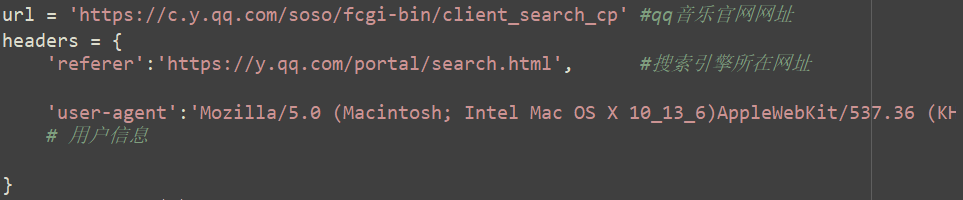




图4.相关指令



图5.部分源码信息

Step 3 通过json 解析数据、提取指标与保存数据

源码信息中有大量无用字符，通过json.loads（）函数解析源码，去除无用字符，提取歌名、所属专辑、年份、歌词等信息 ，并将结果保存为csv文件到本地

但由于需要下载qq音乐客户端并开通vip权限才能查看QQ音中周杰伦的全部歌曲的信息，由于爬虫技术有限，本项目爬取到的歌曲为当前热度排行Top 60的歌曲， 利用pandas打开数据框如下：



图5.部分数据

3. 数据标注

由于Task 2需要构建情感预测模型，而目前得到的数据是缺少情感标签的，因此需要做数据标注，即给现有的数据打上标签。

本项目根据百度百科中对于周杰伦的歌曲鉴赏，将周杰伦歌词情感分为积极（label=2）、中性（label=1）和消极（label=0）三类：



图6. 周杰伦的歌曲鉴赏

按照该分类准则，将原始数据打上标签，如下图：

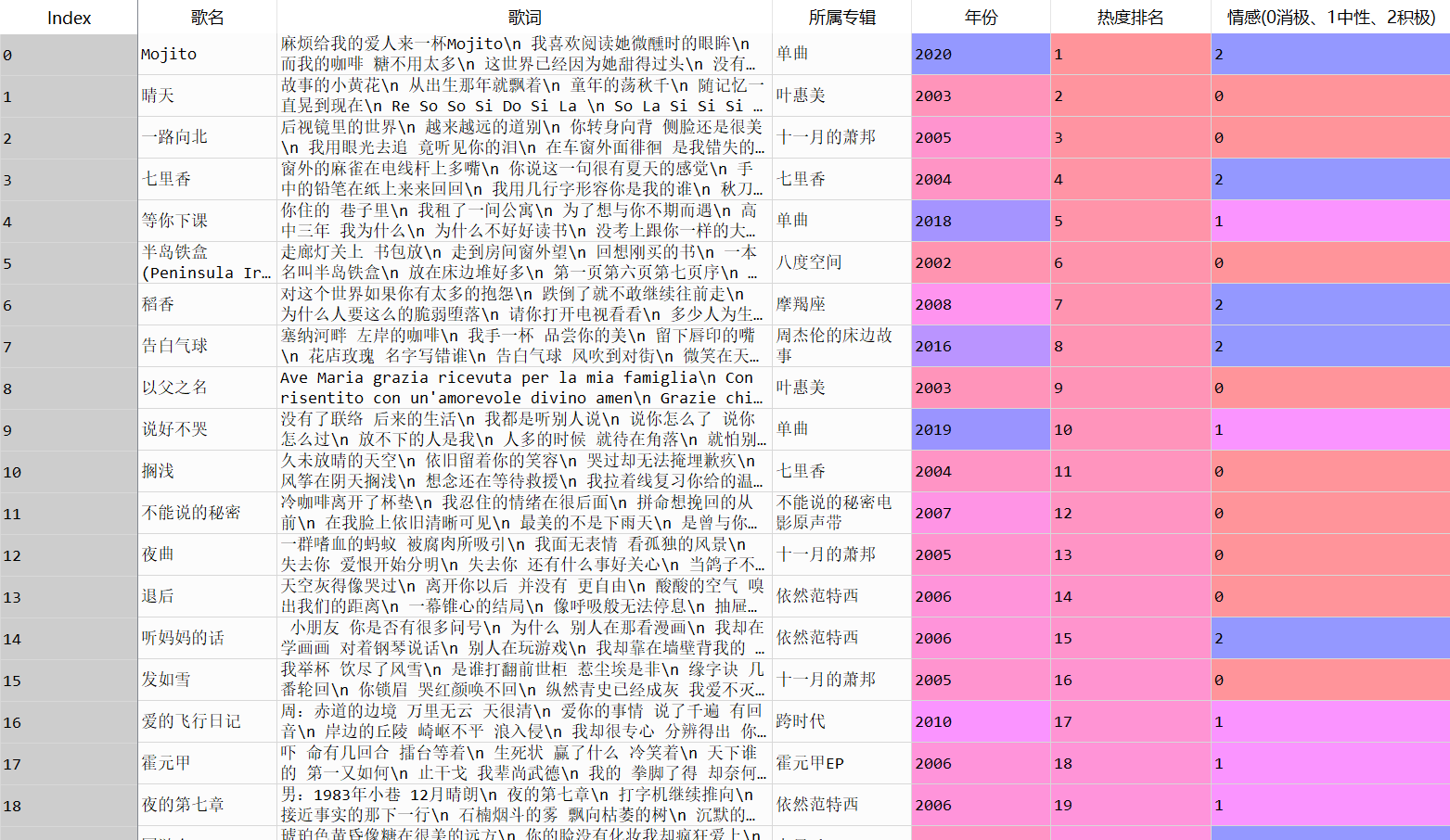


图7.打过标签后的数据

(二) 描述性统计分析

(三) 歌词情感分析

(四) 歌词生成器的构建

1. **不足与改进点**
2. 数据量不足，应扩充样本量，大数据更适合深度学习模型；
3. 人工数据标注虽然参考百度歌曲鉴赏内容，但仍具有主观性；

**五、代码展示**