**项目进展报告**

**数据获取及预处理**

利用的是NeteaseCloudMusicApi开源项目，对网易云音乐的歌单数据进行爬取。本次实验爬取下来的网易云音乐歌单数据以JSON格式进行存储。

下图为本次实验中爬取的一个歌单数据的具体格式:

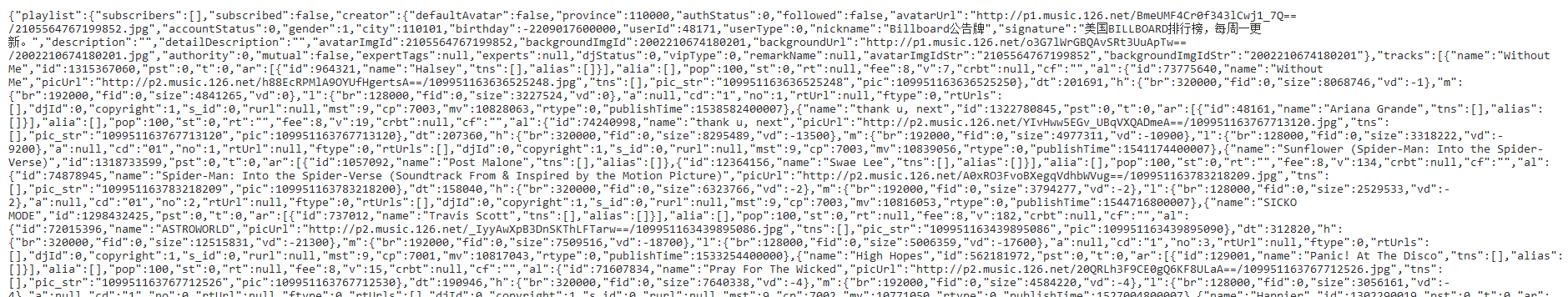


图 爬虫部分效果图

我们对爬取下来的数据进行数据预处理，并进行数据提取、特征选取。本智能歌单系统最终推荐效果的好坏不仅仅取决于算法的好坏和数据的准确性，还取决于选取歌单的特征。

具体步骤包括以下3个步骤:

1. Parser步骤。该步骤的功能主要是进行数据清洗，具体步骤为读取JSON格式的网易云音乐歌单数据，选取以下字段的数据，歌单名、歌单id、歌单tag、歌单收藏次数、歌曲名、歌曲id、歌曲演唱者;初步处理后，最终用于训练的歌单数据为24951，大小为145M，歌曲数量为801194。

2. Dummper步骤。该步骤的主要功能是将使用python的pickle模块将歌单数据和歌曲数据序列化，并保存到本地，数据持久化。

3. processor步骤。该步骤的主要功能是将数据清洗后的数据处理成python surprise库所需要的格式。

**数据分析与可视化**

具体效果如下图所示：

输入歌单名，系统为用户提供和输入歌单相似的歌单。

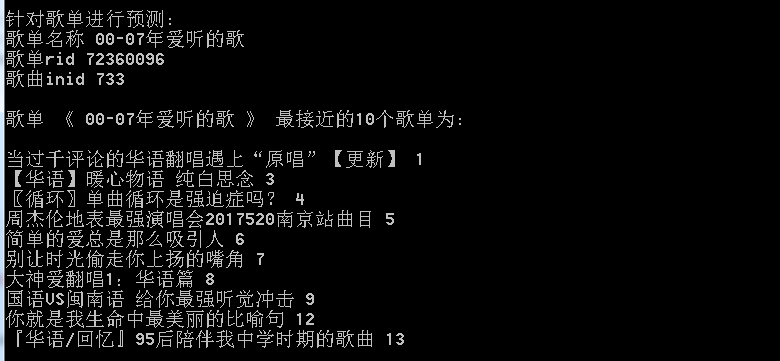


图 歌单推荐效果图

系统为用户推荐的歌曲。

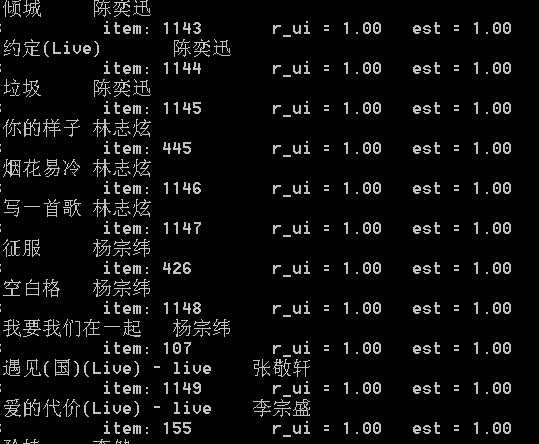


图 歌曲推荐效果图

**数据预处理流程**：

1. 从原始的JSON数据中挑选出下列关键字

歌单名 歌单id 歌单tag 歌单收藏次数

歌曲名 歌曲id 歌曲演唱者

1. 将不符合需求格式的歌单数据全部剔除，将结果存在txt文件，最终用于训练的歌单数据为24951，大小为145M
2. 序列化歌单数据，并保存到本地
3. 将数据转化成符合格式要求的数据集

模型选取

基于用户的协同过滤 KNN

基于物品的协同过滤 KNN

基于模型的协同过滤(SVD, PMF, SVD++, NMF等)

**存在的问题**

目前我们的歌单推荐系统在面对海量的歌曲数据、更复杂的推荐要求时，普遍暴露出准确性低、系统稳定性差、处理速度迟缓等问题。

**下一步工作**

在服务器端，由于网络上的音乐信息每天都会大量更新，因此需要有一个专门的爬虫系统，使得能每过一段时间来更新一遍数据库中的歌单数据。在算法方面，当前的推荐功能还有一些不足之处，准确率有待改进。