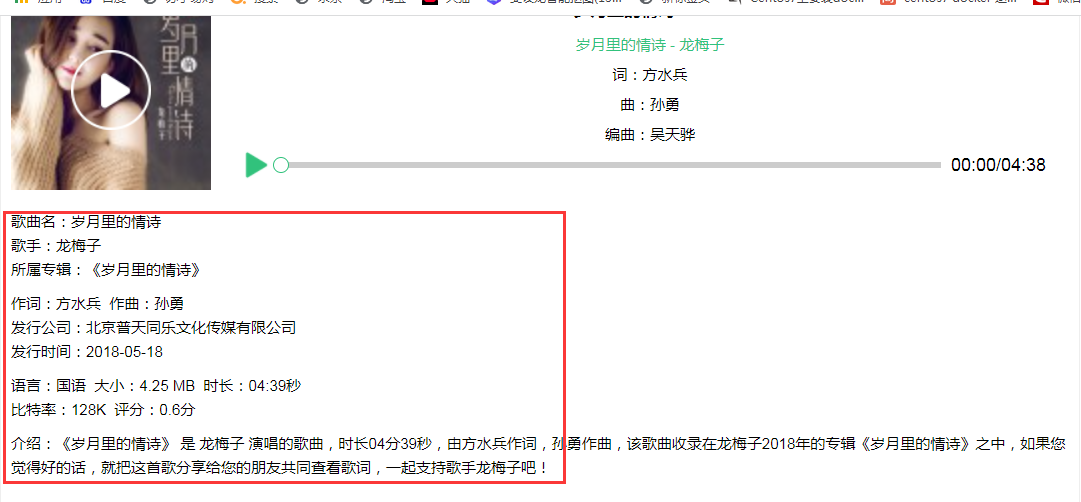
1. **数据采集**

https://www.8lrc.com/

采集【歌词搜索榜】、【上周歌词点击榜】栏目下的所有歌曲：



每一首歌词需采集如下字段

要求：

1、分别使用requests/scrapy实现

2、将爬取的歌词数据分别保存到数据库和csv文件中。

**二、数据清洗（spark）**

歌曲信息表如下（CSV格式）

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 歌曲名 | 歌手 | 所属专辑 | 作词 | 作曲 | 发行公司 | 发行时间 | 时长 | 评分 | 歌曲风格 | 上月播放次数 | 本月播放次数 |
| 岁月里的情诗 | 龙梅子 | 《岁月里的情诗》 | 方水兵 | 作曲A | 公司B | 2018 | 04:39 | 6 | 风格C | 2 | 4 |
| 才能证明这是终点歌词 | 歌手A | 专辑A | 作词A、作词B | 作曲B | 公司A | 2019 | 06:39 | 7 | 风格A | 3 | 9 |
| 琥珀色浪漫 洛少爷歌词 | 歌手C | 专辑B | 作词C | 作曲C | 公司A | 2021 | 05:39 | 8 | 风格X | 2 | 88 |
| 就只剩离分歌词 | 歌手G | 专辑C | 作词D | 作曲D | 公司C | 2020 | 04:39 | 11 | 风格C 风格H | 3 | 77 |
| 梦里的人啊在哪里歌词 | 歌手B | 《专辑D》 | 作词E | 作曲E | 公司D | 2018 | 04:32 | 3 | 风格A 风格B | 32 | 3 |
| 幸福会拥抱你拥抱你歌词 | 歌手B | 专辑D | 作词A  作词H | 作曲F | 公司B | 2019 | 01:44 | 2 |  | 44 | 67 |
| 穷叉叉歌词 | 歌手C | 专辑E | 作词X | 作曲G | 公司E | 2021 | 02:39 | 6 | 风格M 风格K | 2 | 3 |
| 我怕来者不是你歌词 | 歌手D | 专辑F | 作词Z | 作曲H | 公司E | 2020 | 04:39 | 7 | 风格A 风格C | 34 | 45 |
| 半生雪 | 歌手B | 专辑A | 作词N | 作曲X | 公司C | 2017 | 04:39 | 4 | 风格E | 23 | 45 |
| 哭泣站台歌词 | 歌手F | 专辑C | 作词N、作词K | 作曲K | 公司G | 2018 | 04:39 | 6 | 风格A 风格H | 43 | 12 |
| 问心 | 歌手E | 专辑D | 作词J，作词T | 作曲X | 公司C | 2019 | 04:39 | 7 | 风格B 风格C | 32 | 24 |
| Room歌词 | 歌手H | 专辑E | 作词C | 作曲E | 公司F | 2020 | 04-60 | 7 | 风格A | 23 | 24 |

任务1 按如下规则对数据进行清洗，将清洗后的数据保存到aa.csv文件中：

1）专辑字段的内容为“《专辑D》”的则改为“专辑D”。

2）作词有多个人的，人名之间应该以空格分隔，比如“张三 李四”，将以顿号和逗号分隔的人名改为空格。

3）时长的格式为MM:SS，将不符合该格式的歌曲删除，如时长为“04-60”的歌曲；

4）评分字段的范围为[0,10]，将不在该服务的评分字段修改为0；

在任务1完成的基础上，基于aa.csv文件，完成如下功能：

1、统计本表中总共有几个歌手？

2、统计每一年份发行的歌曲总数。

3、找出发行歌曲数最低的2个年份。

4、一首歌可以对应多个风格，统计每个风格出现的次数，并给出数量最高的3个风格。

5、统计每首歌的累计播放量，并将统计数据导出到文件xx.csv中。xx.csv的字段为歌曲名称、歌手、上月播放次数、本月播放次数、总播放量。

1. 对比每首歌曲上月播放次数和本月播放次数，找出播放量增长最快和播放量下降最快的歌曲。

**三、数据可视化（flask+echarts）**

在aa.csv文件的基础上完成如下任务：

1、将aa.csv导入到数据库；

2、使用饼状图显示每个发行公司的歌曲数量；

3、使用条形图显示每种风格出现的次数；

4、使用柱状对比图显示每首歌曲的上月播放次数、本月播放次数和总排放量；