本周进度

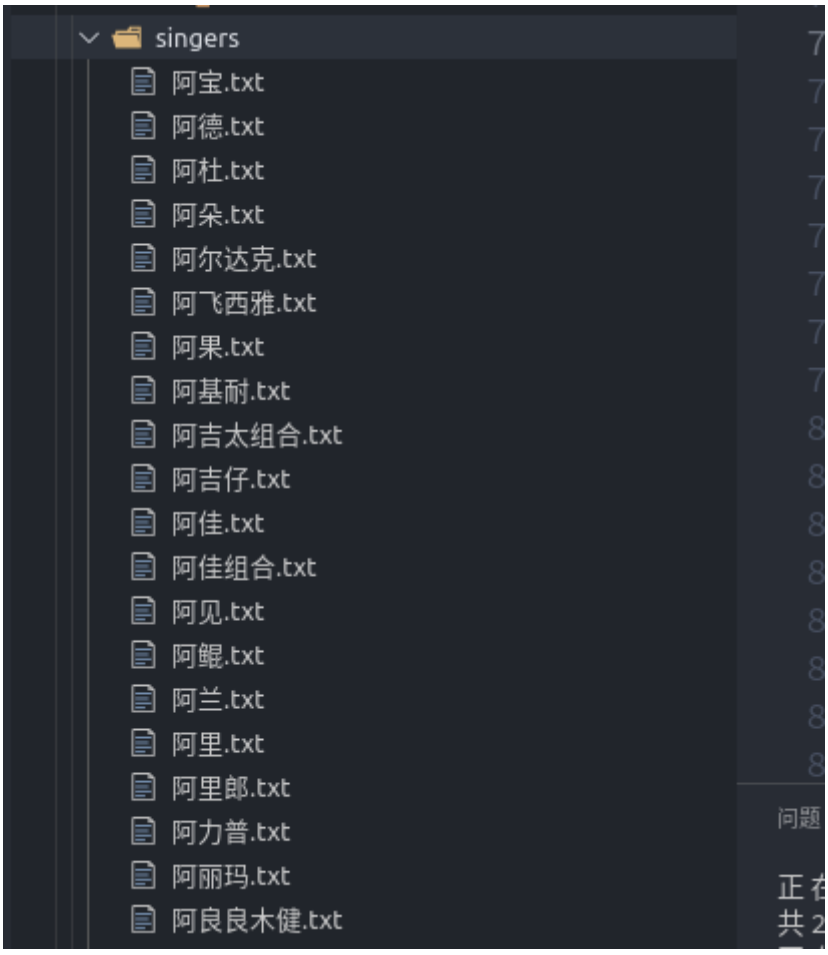
- 继续学习 GPT2-Chinese 模型
- 学习爬虫
- 查看 spider163 项目爬取网易云音乐数据
- 看论文

1. 网易云数据爬取

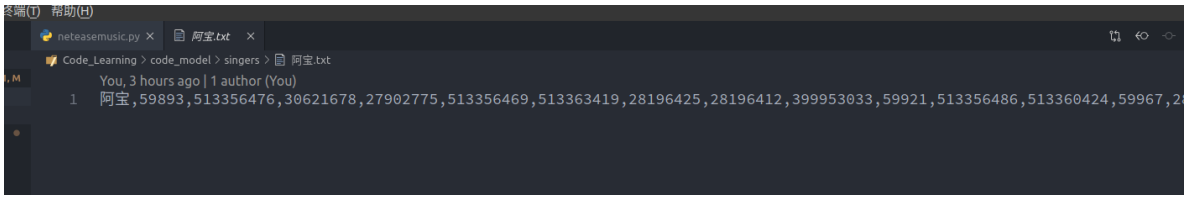
1.1 数据处理

将原文件通过歌手切分为多个文件，提取出每个歌手的歌曲对应id，进行爬取数据。

- 分割后的数据：



- 数据内容



1.2 源码修改

对spider163进行源码修改，使得返回数据最终保存到对应文件中。

1.3 数据爬取和保存

通过循环读取处理后的文件进行歌曲信息爬取，最终以json形式保存到对应文件中。

1.4 爬取结果

```
Code_Learning > code_model > singers_save > {} 阿良良木健.json > {} 0 > abc lyric
[
    {
        "album": "洛天依作品2号：V",
        "name": "绝体绝命",
        "author": "阿良良木健,洛天依",
        "lyric": "[by:雪落与掌心]\n[00:00.000] 作曲 : 阿良良木健\n[00:00.015] 作词 : 阿良良木健\n[00:00.47]调教: 阿良良木健\n[00:12.61]编曲/混音: 阿良良木健\n
        [00:17.11]曲绘: 诗驯\n[00:20.87]PV: 睡狸\n[00:24.39]没有人曾体会 没有人曾了解\n[00:27.87]没有人曾感受 我喜与悲\n[00:31.8]我被肆意踏践 丢弃全部尊严\n[00:33.82]
        人们路过 笑过 骂过 不留一声抱歉\n[00:43.51]我等着你污蔑 我看着你诬陷\n[00:46.57]我观赏这喜剧 是你导演\n[00:49.84]即分不清现实 也听不进奉劝\n[00:52.84]用一分偏执
        换三分喜悦\n[00:56.65]你看不懂 你看不见\n[00:59.68]我悄悄 流的眼泪\n[01:02.51]在无尽深夜 用自责 去自毁\n[01:05.90]像刀片 将精神 层层割裂\n[01:09.61]来来 我最亲
        爱的朋友\n[01:12.12]来看我 毁灭 毁灭\n[01:14.84]请让我去自受自虐 承受这污点\n[01:18.54]反复鞭笞我的罪 一遍又一遍\n[01:22.38]来来 我最牵挂的朋友\n[01:24.68]请为
        我 悼念 悼念\n[01:27.45]我身上恶疾已蔓延 无尽的繁衍\n[01:31.11]这世界 我已厌倦\n[01:33.41]「请别将我救援」\n[01:46.84]人们都在沉醉 人们都已忘却\n[01:49.73]人们
        都装作 看懂这结尾\n[01:52.92]一味陷入争辩 无人聆听箴言\n[01:56.8]该可悲可泣 或该叹可怜\n[01:59.69]一差二错 三分愚昧\n[02:02.90]四分五裂 到毁灭\n[02:05.88]谁曾想
        语言 能掀起 这灾变\n[02:08.86]谁曾说 祸起萧墙 是妄言\n[02:12.43]来来 我最亲爱的朋友\n[02:14.99]来看我 毁灭 毁灭\n[02:17.92]请让我去自受自虐 承受这污点\n
        [02:21.39]反复鞭笞我的罪 一遍又一遍\n[02:25.29]来来 我最牵挂的朋友\n[02:27.66]请为我 悼念 悼念\n[02:30.51]我身上恶疾已蔓延 无尽的繁衍\n[02:34.4]这世界 我已厌倦
        \n[02:36.68]「别将我救援」\n[02:37.82]来来 我最亲爱的朋友\n[02:40.20]请听我 最后 遗言\n[02:43.13]我不想就这样沦陷 迷失在黑夜\n[02:46.58]我想燃烧这生命 就算再壮
        烈\n[02:50.47]拜拜 我最牵挂的朋友\n[02:53.3]请不要 为我 悼念\n[02:55.90]如果风降临在春天 请与我相见\n[02:59.25]这世界 我仍依恋\n[03:02.30] 「请你将我救援」\n",
        "id": 550059604,
        "cmts": [
            {
                "liked": "1113",
                "txt": "龙牙和声，最后的please don't follow me why you why you follow me是up和的\n\n把我顶上去啊[撇嘴][撇嘴][撇嘴]",
                "author": "墨水青蓝绿"
            },
            {
                "liked": "1143",
                "txt": "鲁迅:‘‘我他妈没说过这句话’’",
                "author": "滑稽的滑稽鱼"
            },
            {
                "liked": "1144",
                "txt": "老年天依是《春风来》里的主角，少女天依是《恋爱理论》专辑里的主角。老年天依身中一种病毒，如果不通过“特殊的方法”治疗，就会很快死去，而少女天依就是因此被绑
                架了过来。但故事并非如此简单，完整剧情：  https://m.weibo.cn/status/4226202388205729",
                "author": "次元幻想"
            },
```

2. 问题总结

- ~~源码修改过程中出现一些配置错误 --已解决~~
- 服务器安装包配置出现问题 -- 未解决

3. 下周任务

- 论文阅读
- GPT2-Chinese 学习
- 整合爬取数据