

本周进度

- 学习斯坦福深度自然语言处理课程
 - 第一讲：NLP与深度学习入门
- 修改 Spider163 爬虫程序 -- 部分完成
- 论文查找和阅读

1. 网易云数据爬取

- 爬取非精彩评论

目前只修改部分源码，可以按照时间线进行评论的爬取，需要对爬取数据进一步解析，未进行文件保存

```
/usr/local/lib/python2.7/dist-packages/spider163/spider » spider163 comment -s 411214279
正在执行抓取歌曲 411214279 热门评论计划
时间对不上，有请下一位演员
抓取完成
```

2. 问题总结

- 爬取数据中包含所以评论和其他比较乱的信息，需要进一步进行解析处理

```
正在执行抓取歌曲 411214279 热门评论计划
[{'content': u'\u65f6\u95f4\u5b9d\u4e0d\u5f0c\u6709\u8bf7\u4e0b\u4e00\u4f4d\u6f14\u5458', 'status': 0, 'beRepliedCommentId': 1666968606, 'user': {'liveInfo': None, 'remarkName': None, 'expertTags': None, 'avatarUrl': u'http://p1.music.126.net/QtRaj-kvLUuP0jMAgtEXjQ=/109951164313662039.jpg', 'userId': 1769014993, 'locationInfo': None, 'vipRights': None, 'userType': 0, 'experts': None, 'authStatus': 0, 'nickname': u'\u9727\u6739', 'vipType': 0}, 'expressionUrl': None}]
抓取完成
```

下周进度

- 继续学习课程
- 修改爬虫程序
- 同步学习 pytorch 与 transformers