文本作业 1

张配天-2018202180

2020 年 4 月 28 日

1 代码结构

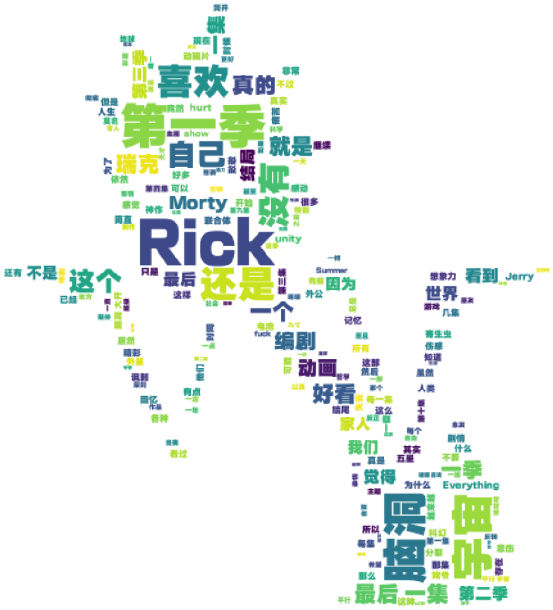
```
/my
├── /my
│   ├── /spiders
│   │   └── mySpider.py
│   ├── settings.py
│   ├── items.py
│   ├── pipelines.py
│   └── middlewares.py
├── scrapy.cfg
├── multi_JSON_transfer_TXT_jieba_wordCloud.py
├── item.json
└── result.txt
```

2 代码说明

- 在 settings.py 中设置 USER_AGENT 列表, 设置 Cookie(用来登录账户, 否则只能爬取前 *200* 条评论), 设置 ITEM_PIPELINES 为自己定义的 MyPipeline
- 在 items.py 中设置 author、reply 属性
- 在 pipelines.py 中将 item 处理方式设置为将非空的作者和其评论保存到 item.json(其中含有多个 *json*)
- 在 middlewares.py 中将 process_start_requests 方法定义为从 USER_AGENT 列表中随机选取一个作为 requests 的阅览器标头 (否则豆瓣评论页不允许访问, 或几波爬取后容易被识别发生 *403* 错误)
- 在爬虫文件 mySpider.py 中完成对 parse 方法的重定义, 通过 xpath 获取 author 和 reply 属性
- 在 multi_JSON_transfer_TXT_jieba_wordCloud.py 中实现对 item.json 的读取 (利用 jsonlines)、分词 (存入 result.txt)、绘制词云

3 词云

图 1: 选取 rick and morty 第二季的 500 条评论绘制, 以 rick 为背景



4 insights

i. 实现登陆的 cookie 要时常更新。。。应该和 expiration date 有关

ii. 登录后爬取可以考虑加等待时间, 要不然容易被 ban……

iii. 考虑添加 IP 池, 应该效果更好, 我中间就被封过一次号 (当时想爬 10000 条, 结果没那么多)