**第九周进度报告**

18124399 冯秦萱

2021/3/3

**作业选题及方向：**

探索bilibili番剧信息，研究剧集数，追番人数，投币数，点赞数，弹幕数，播放数，标签等对评分的影响程度，构建评分预测模型；根据每年的平均数据，预测2021年平均播放量，番剧数量；提取简介关键词，标签等信息，总结人气番特点。

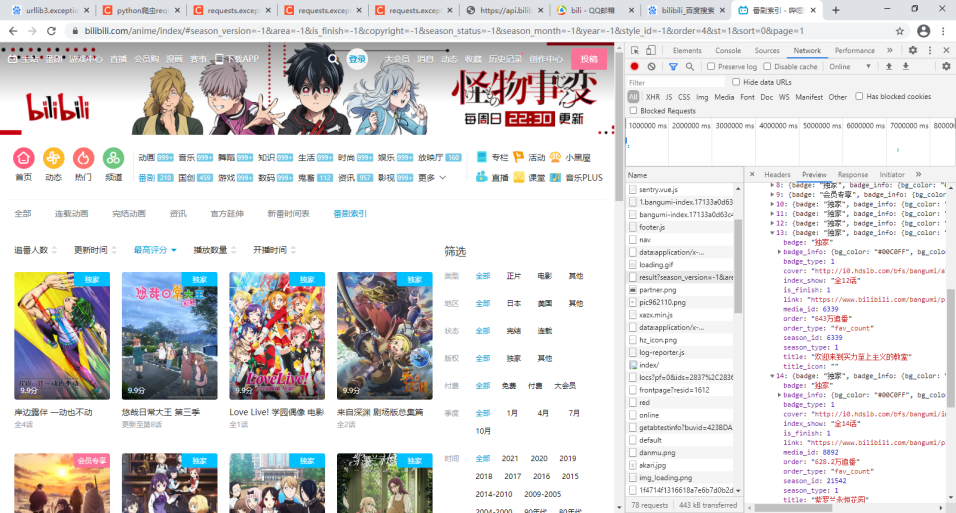
**已完成工作：**

**1.数据爬取**：

爬取B站**番剧索引页，播放页和详情页**中的数据, 爬取100页，每页20个共2000个番剧信息。分两部分爬取后合并。

**Part 1 (bilibili.py)**

番剧索引页爬取**番剧名，剧集数，分级，评分，是否完结，播放链接，番剧id，详情id**



播放页爬取**系列追番人数，追番人数，投币数，点赞数，弹幕数，播放数，**



**Part2 (bilibili\_detail.py)**

详情页**爬取标签，评分人数，发布日期，简介**



**2.数据集概述**：

bilibili\_data.csv

**爬取来源：**bilibili番剧索引

<https://www.bilibili.com/anime/index/#season_version=-1&area=-1&is_finish=-1&copyright=-1&season_status=-1&season_month=-1&year=-1&style_id=-1&order=4&st=1&sort=0&page=1>

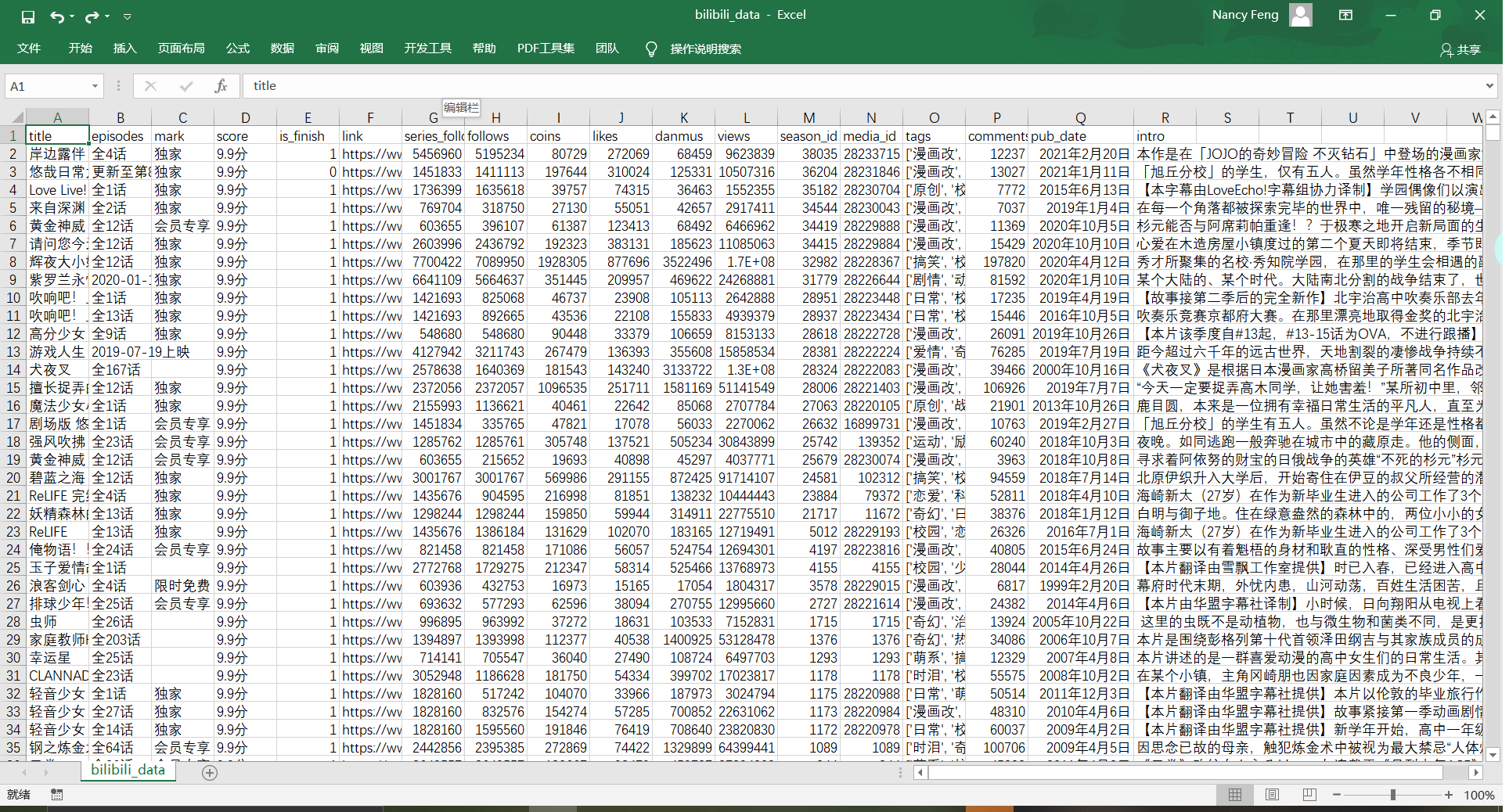
**工具：**python（主要使用api和selenium）

**大小：**2000 rows x 18 columns

|  |  |  |
| --- | --- | --- |
| **列名（EN）** | **列名（CN）** | **举例** |
| title | 番剧名 | 黄金神威 第二季 |
| episode | 剧集数 | 全12话 |
| mark | 分级 | 会员专享 |
| score | 评分 | 9.9分 |
| is\_finish | 是否完结 | 1 |
| link | 播放链接 | https://www.bilibili.com/bangumi/play/ss25679 |
| series\_follow | 系列追番人数 | 603655 |
| follows | 追番人数 | 215652 |
| coins | 投币数 | 19693 |
| likes | 点赞数 | 40898 |
| danmus | 弹幕数 | 45297 |
| views | 播放数 | 4037771 |
| season\_id | 番剧id | 25679 |
| media\_id | 详情id | 28230074 |
| tags | 标签 | ['漫画改', '搞笑', '冒险', '历史'] |
| comments | 评分人数 | 3963 |
| pub\_date | 发布日期 | 2018年10月8日 |
| intro | 简介 | 寻求着阿依努的财宝的日俄战争的英雄“不死的杉元”杉元佐一和阿依努族少女阿席莉帕。在追踪着身上刺青含有藏宝地提示的24名逃犯的过程中，逐渐发现了新的事实。抢走了宝藏、杀了阿席莉帕父亲的野篦坊，正是阿席莉帕的父亲本人。为了确认事情的真伪，只能去网走监狱中和他想见。“逃狱王”白石由竹和奇罗蓝克也加入了两人的旅行，从小樽到札幌最后前往夕张。另一边，目标同样是宝藏的第七师团的鹤见中尉与新选组“鬼之副长”土方岁三一派，也怀抱着各自的野心在行动着。明治时代后期，北方大地上发生的争夺财宝的生存战，更加激烈—— |

**数据预览：**

共2000条



**3.部分数据处理**：

episodes列进行数值化，去除异常值，mark列空值填充为‘免费’，不同级别0-1编码，score列数值化，tags列去除空值，分词，0-1编码。

**后续工作:**

* 继续完成数据处理
* 探索，可视化分析
* 特征工程
* 预测模型
* 评估总结