**温州大学计算机与人工智能学院**

**Python应用开发 实验报告**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **课程名称：** | Python应用开发 | | | | |
| **班 级：** | 23网工2 | **姓名：** | 王益民 | **学号：** | 23211835201 |
| **实验地点：** | 5b105 | **日期：** | 4.11 | | |

**一、程序简介**

bbs.nga.cn是一个较大的互联网游戏论坛，需要登陆账号来访问帖子数据。

1.     通过Cookie移植的方式实现论坛50页历史数据的爬取，获取数据包括每个帖子的链接、发布时间、标题、楼主ID、帖子内容、跟帖ID、跟帖时间、跟帖内容（仅限文字部分，图片可不进行处理）。

2.针对上述爬取数据，将数据同时存储到MySQL数据库或者内存数据库Redis。

**二、开发要点**

1.cookie的获取：使用cookie实现更方便的登录

2.数据的处理：对评论中的引用，表情等特殊内容的处理

**三、运行效果、运行关键步骤截图**

1.get\_cookie.py

登录成功后使用driver.get\_cookies函数保留cookie数据到Cookies.json文件中



Cookies.json数据的部分截图



1. get\_link.py

获取程序员职业交流论坛前50页的数据



将获取到的数据存放到link.txt中



Link.txt的部份数据展示



1. get\_data.py

使用cookies.json实现登录状态



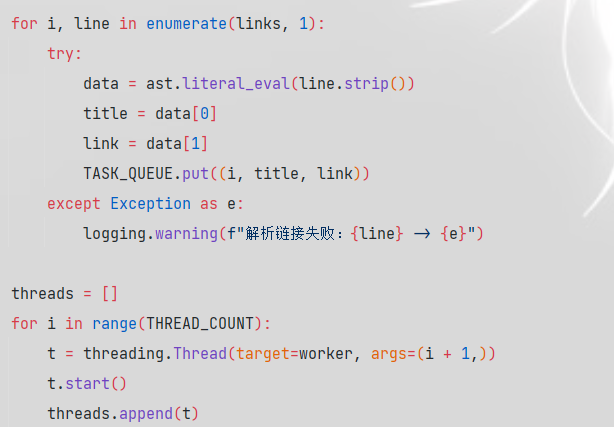
遍历link.txt将每个帖子的数据爬取并一一存储到./data/post-{i}.json

文件中

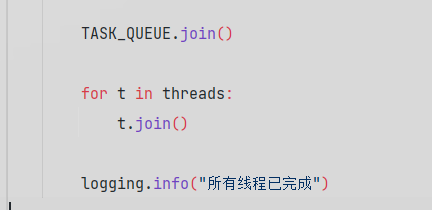


使用多线程的消费者-生产者模式以便适当的提供效率

生产者：



确保所有任务都完成



获取到的数据的部分截图



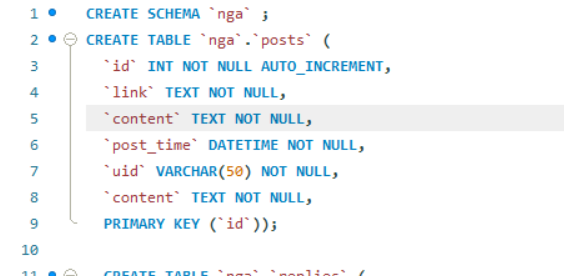


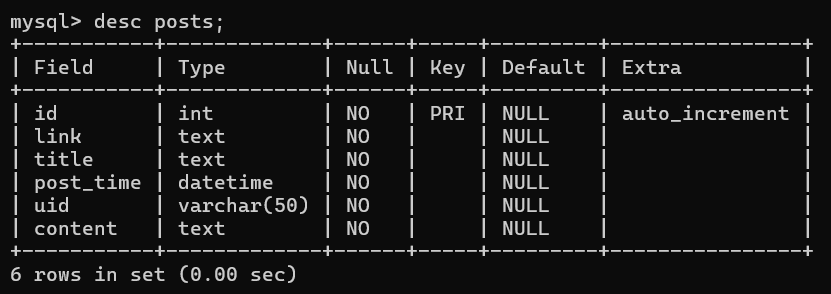
1. save\_data.py

建立数据库nga

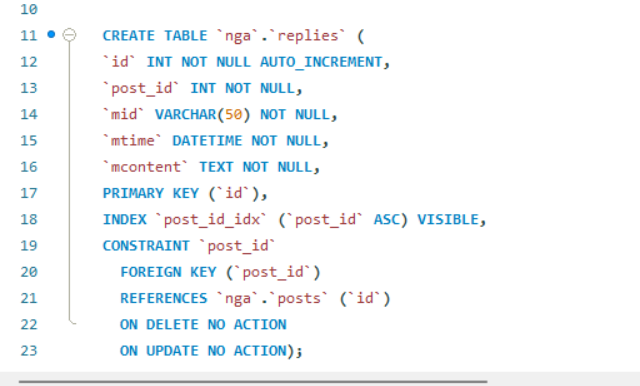
建立表table posts:记录每条帖子的数据；replies:每个帖子里面的回帖数据。将posts的id与replies的post\_id关联起来

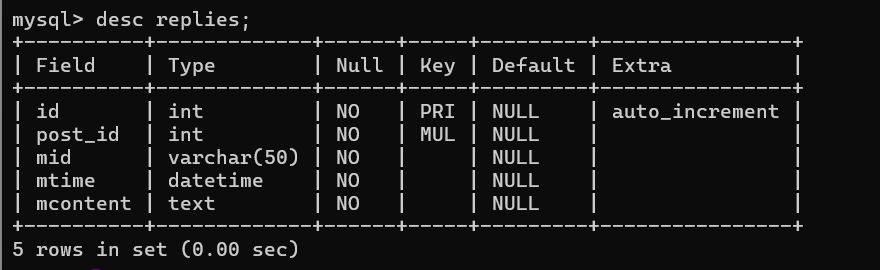
Posts表



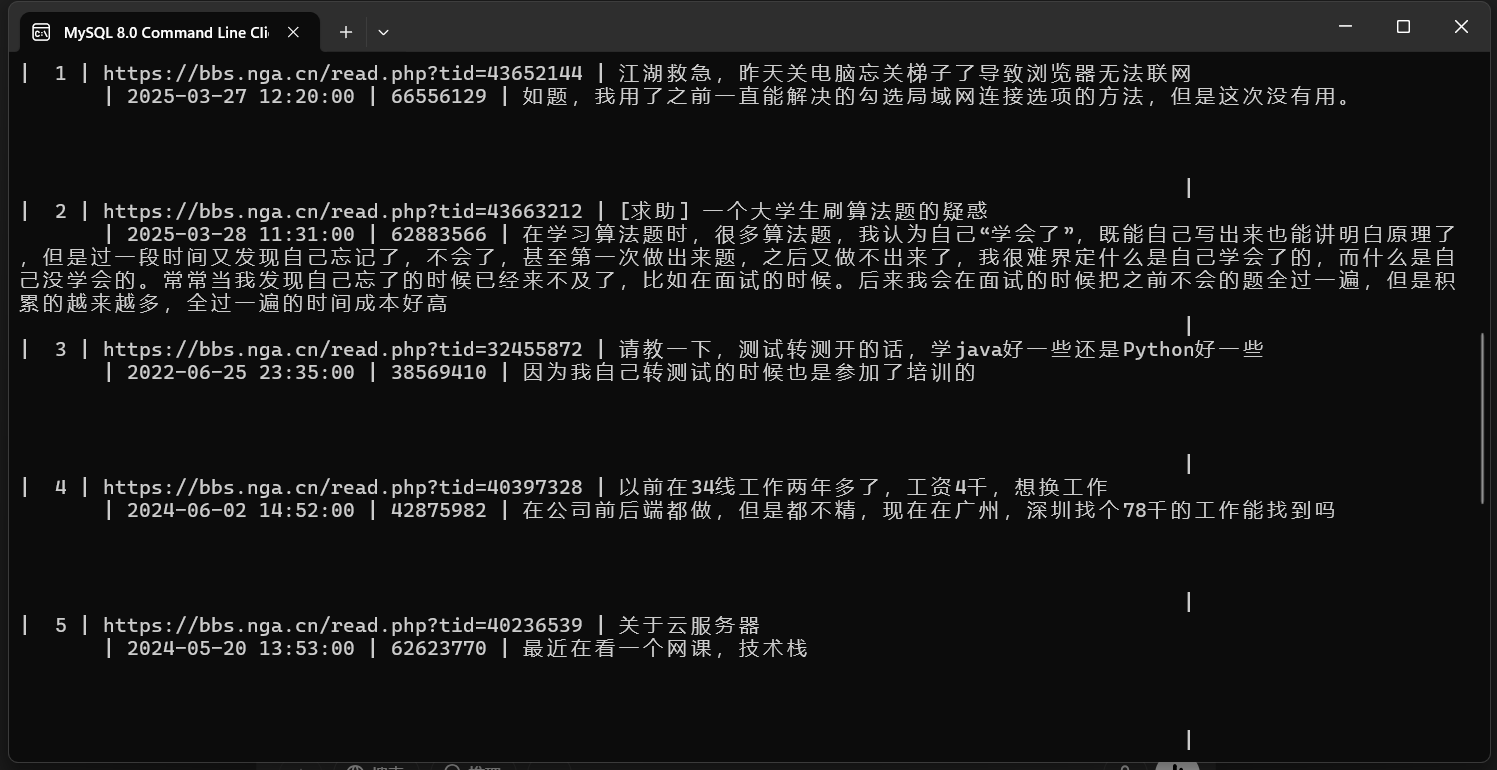


Replies表





Posts部分数据展示



Replies部分数据展示

