虎扑每日赛后评论爬虫方法

概述

本文介绍如何使用 **Scrapy** 爬取虎扑 NBA 比赛后的热门评论（高亮评论），并获取球员数据和比赛基本信息（如比分，球员得分， 篮板， 助攻等）。

背景

作为乐子人，我一直想微调一个大语言模型（LLM），并尝试将他训练成虎扑网友评论的风格。这些评论通常与比赛数据和特定球员密切相关（eg. 詹姆斯的黑粉在嘲讽时有固定的起手式和段子）。

为了构建一个有效的训练数据集，我们需要收集：

- 比赛信息
- 球员数据
- 针对每个球员的评论

而虎扑的赛后评分则提供了一个完美的数据库，因为它既有大量的网友评论，又有球员对应的数据，能够为微调的过程提供比较有用的 ‘prompt’ 和 ‘completion‘。

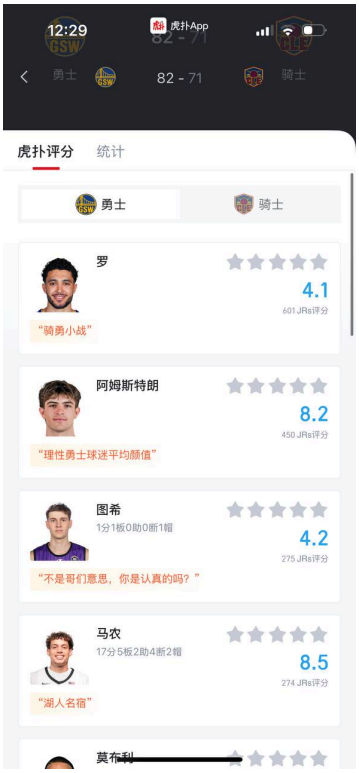
方法

数据来源

虎扑有两种访问方式：

1. **手机 App**
2. **网页版**（www.hupu.com）

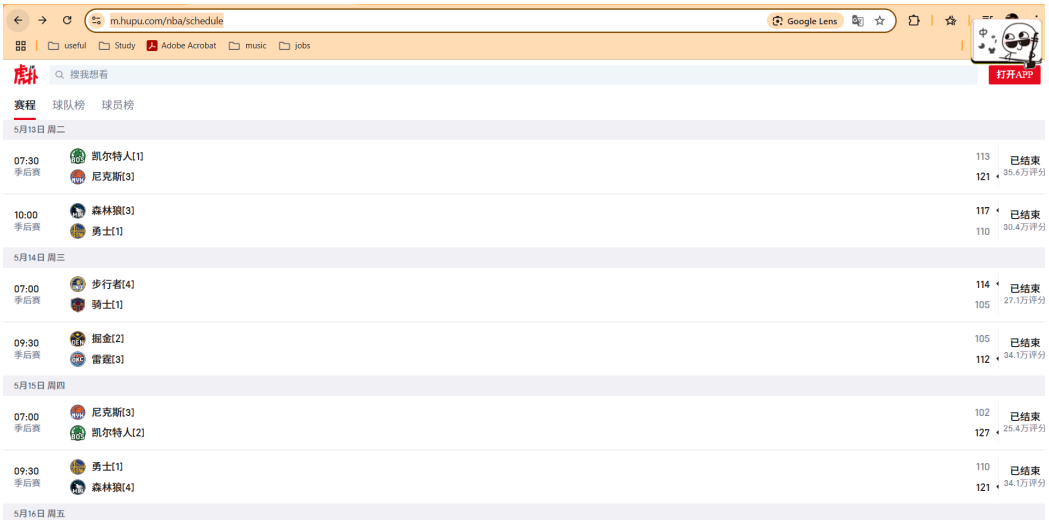
网页版提供了完整的比赛数据，但并没有比赛回顾和高亮评论区。我想要爬取的评论区位于 **“虎扑评分”** 下方（见截图），只在虎扑手机 App 中可见。



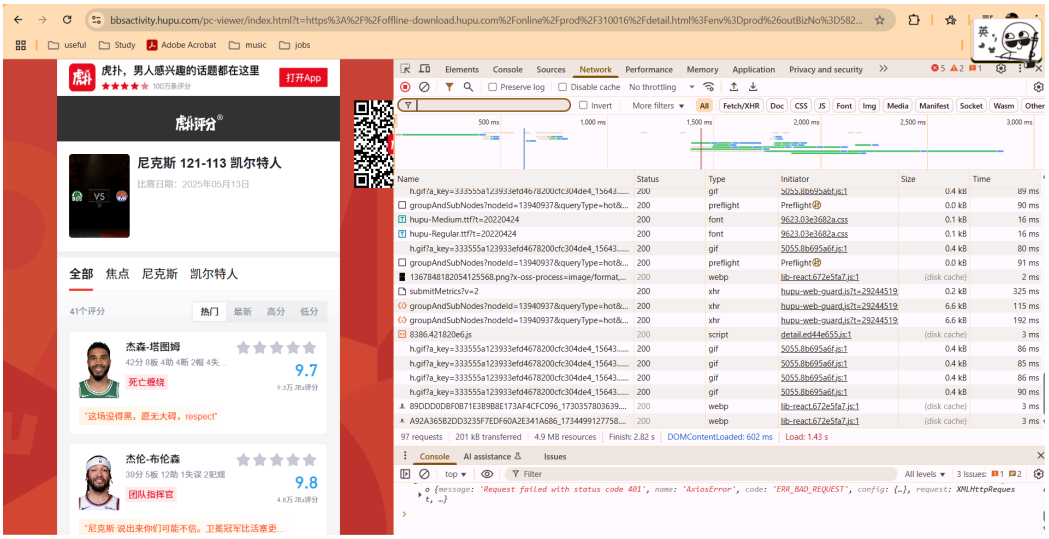
由于 虎扑 mobile app 的数据接口更难直接找到，我们可以利用虎扑的移动端网页版（m.hupu.com）来分析 API 请求。

爬取流程

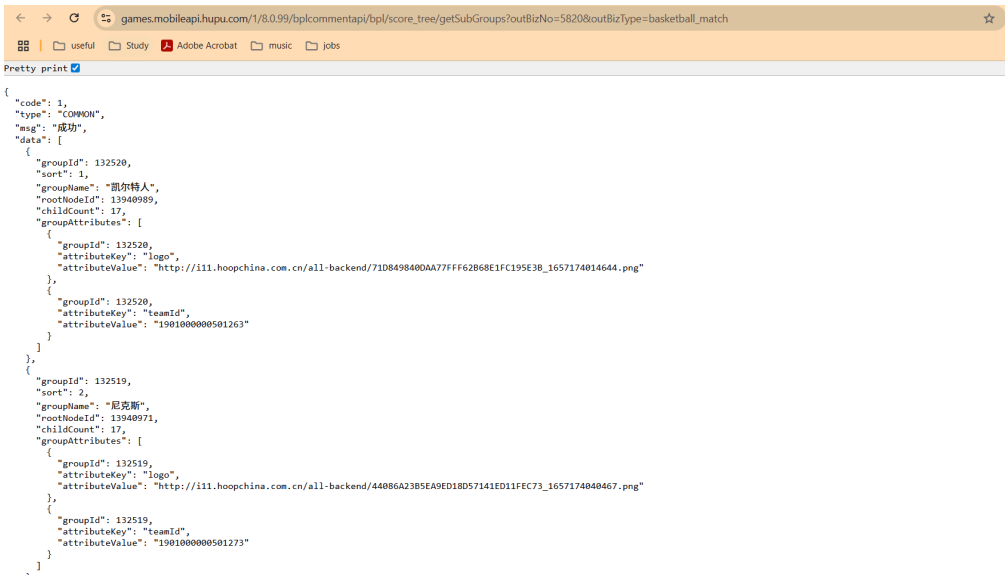
- 首先，找到比赛列表
  访问：https://m.hupu.com/nba/schedule
  （注意：该页面只显示有限时间范围内的比赛，大部分比赛是隐藏的。）



- 然后，我们分析网络请求：选择任意一场比赛，打开浏览器开发工具 → network 面板，观察网络请求。



- 以下的这些XML请求是我们的重点关注对象
  - https://games.mobileapi.hupu.com/1/8.0.99/bplcommentapi/bpl/score_tree/getSubGroups?outBizNo=5820&outBizType=basketball_match
    通过观察， 这里的***outBizNo***对应着每场比赛的***rootNodeID***（一个8位数）。我们可以通过遍历outBizNo来获取所有的rootNodeID。通过试错，outBizNo的最大值是5800左右。（最大值对应更为近期的比赛）

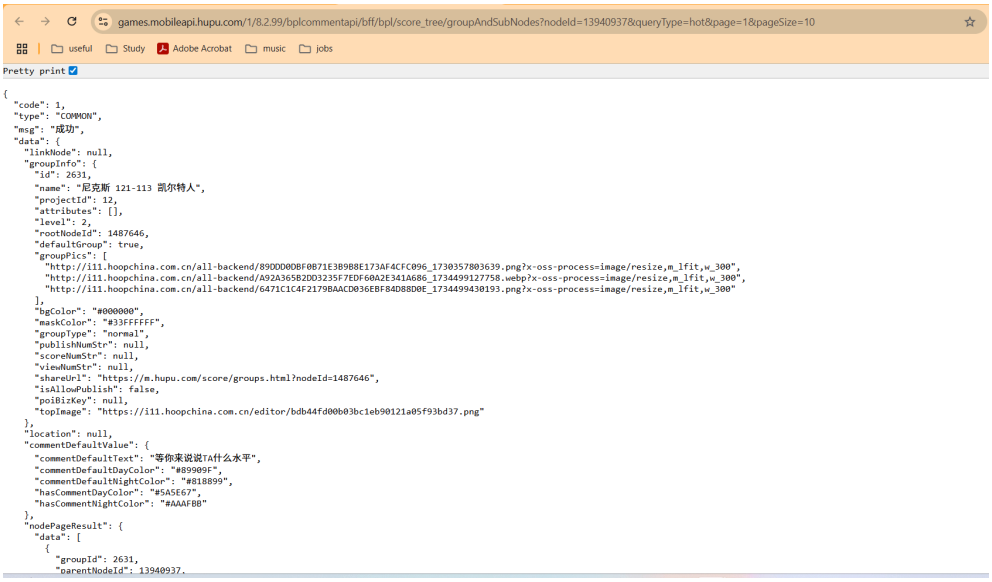



```
{
  "code": 1,
  "type": "COMMON",
  "msg": "成功",
  "data": [
    {
      "groupId": 132520,
      "sort": 1,
      "groupName": "凯尔特人",
      "rootNodeId": 13940989,
      "childCount": 17,
      "groupAttributes": [
        {
          "groupId": 132520,
          "attributeKey": "logo",
          "attributeValue": "http://i11.hoopchina.com.cn/all-backend/71D849840DAA77FFF62B68E1FC195E3B_1657174014644.png"
        },
        {
          "groupId": 132520,
          "attributeKey": "teamId",
          "attributeValue": "1901000000501263"
        }
      ]
    },
    {
      "groupId": 132519,
      "sort": 2,
      "groupName": "尼克斯",
      "rootNodeId": 13940971,
      "childCount": 17,
      "groupAttributes": [
        {
          "groupId": 132519,
          "attributeKey": "logo",
          "attributeValue": "http://i11.hoopchina.com.cn/all-backend/44086A23B5EA9ED18D57141ED11FEC73_1657174040467.png"
        },
        {
          "groupId": 132519,
          "attributeKey": "teamId",
          "attributeValue": "1901000000501273"
        }
      ]
```

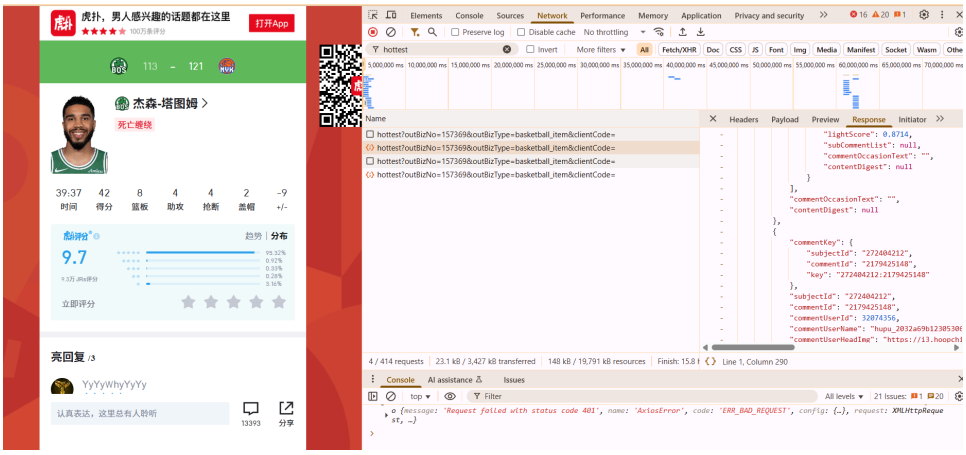

  - https://games.mobileapi.hupu.com/1/8.2.99/bplcommentapi/bff/bpl/score_tree/groupAndSubNodes?nodeId=13940937&queryType=hot&page=1&pageSize=10

    nodeID=<groupID>， 通过这个请求，我们可以得到某场比赛的JSON响应（见附图）
    从这个JSON响应中，我们可以得到
    1. 每支球队的比分
    2. 球员数据
    3. 每个球员各自的***BizID***（这个参数在后面会很重要）

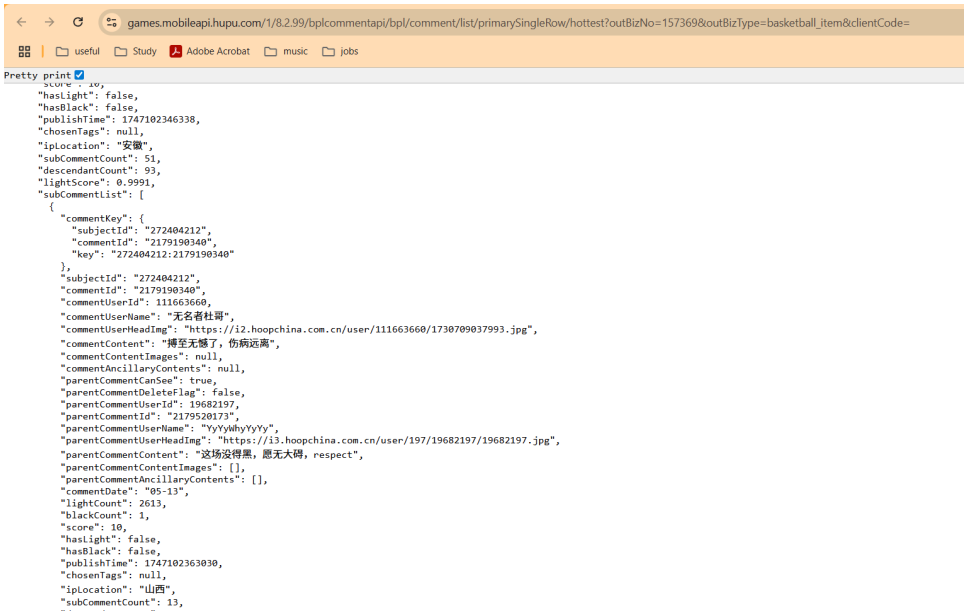
由于在本页只能看到一条热门评论，我们需要找到一个能够显示更多评论的接口来爬取更多的热门评论。



- https://games.mobileapi.hupu.com/1/8.2.99/bplcommentapi/bpl/comment/list/primarySingleRow/hottest?outBizNo=157369&outBizType=basketball_item&clientCode=
  在点进每一个球员的页面之后，我们可以看到所有的评论。
  在这里，outBizNo=<BizID>，而***bizID***正是我们前面所获取的参数。所以，我们只要利用***bizID***发送这个请求，就可以查看此球员在这场比赛收到的所有赛后评论。



- （看附图）所有的评论都在这个JSON响应里。



现在，我们已经获得了收集所需信息的全部响应数据。
接下来需要设计一个爬虫脚本，来完成以下步骤：

1. 生成所有篮球比赛的 **groupID**
2. 获取比赛数据以及每位球员的 **bizID**
3. 请求对应的 JSON 接口
4. 获取每位球员的 **前三条热门评论**
5. 最终将所有数据写入一个 .csv 文件中

爬虫脚本开源github repo: https://github.com/alexgoexercise/hupu-comments-crawler.git