Crawling Hupu’s NBA Game Comments?

Overview

This article explains how to use Scrapy to crawl and scrape Hupu's hottest comments (高亮评论) after NBA games, along with player statistics and basic game information such as scores.

Background

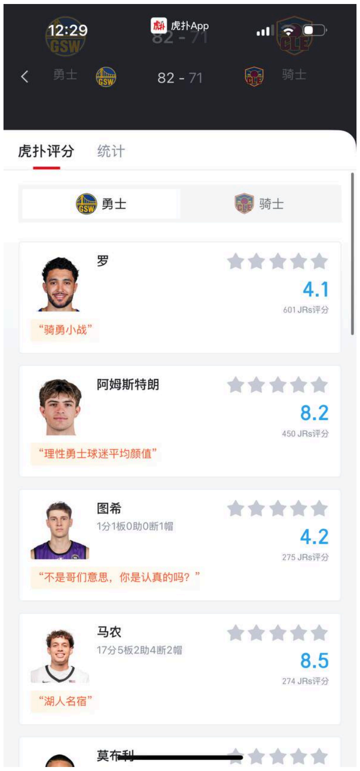
I've always wanted to fine-tune an LLM to mimic Hupu JRs' commenting style, which is typically concise, harsh, and direct in both criticism and praise. These comments are closely tied to game statistics and the specific players being discussed (for example, LeBron James haters have particular patterns when mocking him).

To build an effective training dataset for the LLM, we need to collect game information, player statistics, and the comments about individual players.

Method

Context

Hupu can be accessed in two ways: through the mobile app or through the website (www.hupu.com).
While Hupu's website provides a comprehensive database of game statistics, it lacks the game review and highlighted comment sections. The comment section I'm targeting appears under "虎扑评分" (as shown in the screenshot) and is only available in the Hupu mobile app.



罗
4.1
601 JRs评分
“骑勇小战”

阿姆斯特朗
8.2
450 JRs评分
“理性勇士球迷平均颜值”

图希
1分1板0助0断1帽
4.2
275 JRs评分
“不是哥们意思，你是认真的吗？”

马农
17分5板2助4断2帽
8.5
274 JRs评分
“湖人名宿”

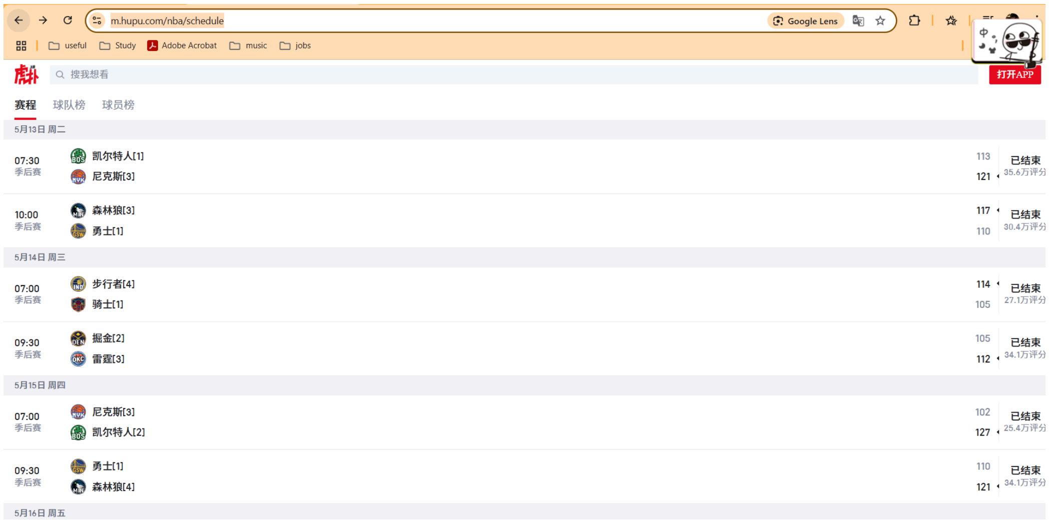
莫布利

Consequently, we need to crawl data from the Hupu app, which is more challenging than scraping the website since it's typically harder to locate the API endpoints that the app uses.

Fortunately, there is a website version of Hupu's mobile app (m.hupu.com). This allows us to identify the endpoints where Hupu fetches the data.

Crawling Process

- First, we navigate to the NBA schedule page (https://m.hupu.com/nba/schedule). Note that this page only displays games within a limited time range, with most games hidden.



m.hupu.com/nba/schedule

搜我想看

打开APP

赛程 球队榜 球员榜

5月13日 周二

07:30 季后赛	凯尔特人[1] 尼克斯[3]	113 121	已结束 35.6万评分
10:00 季后赛	森林狼[3] 勇士[1]	117 110	已结束 30.4万评分

5月14日 周三

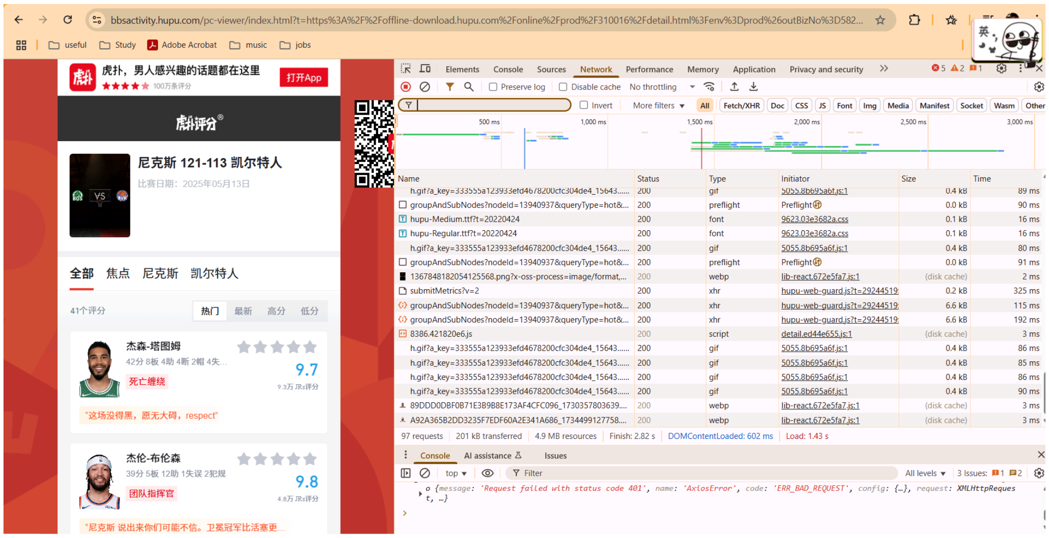
07:00 季后赛	步行者[4] 骑士[1]	114 105	已结束 27.1万评分
09:30 季后赛	掘金[2] 雷霆[3]	105 112	已结束 34.1万评分

5月15日 周四

07:00 季后赛	尼克斯[3] 凯尔特人[2]	102 127	已结束 25.4万评分
09:30 季后赛	勇士[1] 森林狼[4]	110 121	已结束 34.1万评分

5月16日 周五

- Next, we select any game, open the developer tools, and examine the network tab to analyze the traffic:



bbsactivity.hupu.com/pc-viewer/index.html?t=https%3A%2F%2Foffline-download.hupu.com%2Fonline%2Fprod%2F310016%2Fdetail.html%3Fenv%3Dprod%26outBizNo%3D582...

虎扑，男人感兴趣的话题都在这里 ★★★★★ 100万条评分 打开App

虎扑评分®

尼克斯 121-113 凯尔特人

比赛日期：2025年05月13日

全部 焦点 尼克斯 凯尔特人

41个评分 热门 最新 高分 低分

杰森-塔图姆 42分 8板 4助 4断 2帽 4失... 9.7 9.3万 JRs评分 死亡缠绕

“这场没得黑，愿无大碍，respect”

杰伦-布伦森 39分 5板 12助 1失误 2犯规 9.8 4.8万 JRs评分 团队指挥官

“尼克斯 说出来你们可能不信。卫冕冠军比活塞更...

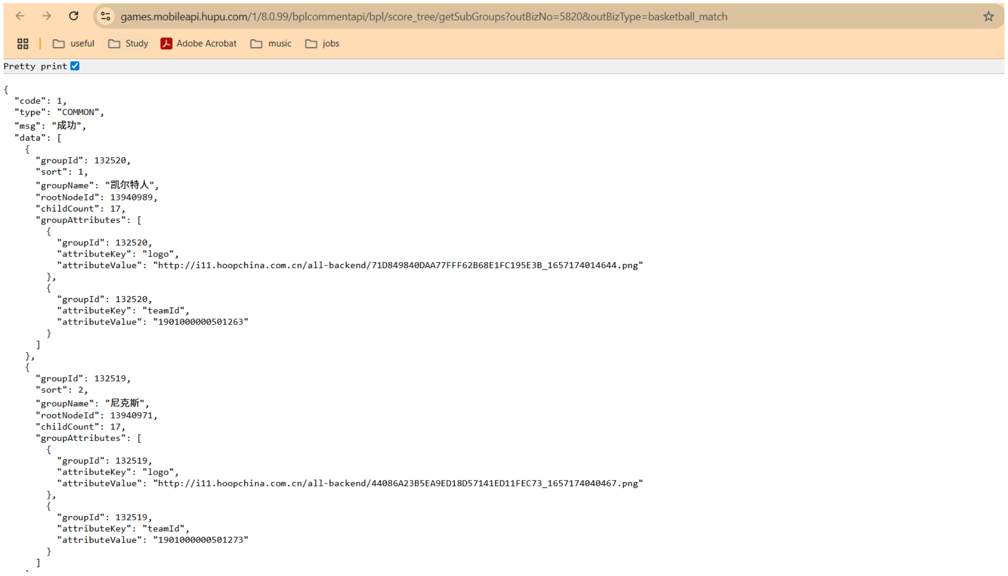
Name	Status	Type	Initiator	Size	Time
h.gif?a_key=333555a123933efd4678200cfc304de4_15643......	200	gif	5055.8b695a6f.js:1	0.4 kB	89 ms
groupAndSubNodes?nodeId=13940937&queryType=hot&...	200	preflight	Preflight	0.0 kB	90 ms
hupu-Medium.ttf?t=20220424	200	font	9623.03e3682a.css	0.1 kB	16 ms
hupu-Regular.ttf?t=20220424	200	font	9623.03e3682a.css	0.1 kB	16 ms
h.gif?a_key=333555a123933efd4678200cfc304de4_15643......	200	gif	5055.8b695a6f.js:1	0.4 kB	80 ms
groupAndSubNodes?nodeId=13940937&queryType=hot&...	200	preflight	Preflight	0.0 kB	91 ms
1367848182054125568.png?x-oss-process=image/format,...	200	webp	lib-react.672e5fa7.js:1	(disk cache)	2 ms
submitMetrics?v=2	200	xhr	hupu-web-guard.js?t=29244519	0.2 kB	325 ms
groupAndSubNodes?nodeId=13940937&queryType=hot&...	200	xhr	hupu-web-guard.js?t=29244519	6.6 kB	115 ms
groupAndSubNodes?nodeId=13940937&queryType=hot&...	200	xhr	hupu-web-guard.js?t=29244519	6.6 kB	192 ms
8386.421820e6.js	200	script	detail.ed44e655.js:1	(disk cache)	3 ms
h.gif?a_key=333555a123933efd4678200cfc304de4_15643......	200	gif	5055.8b695a6f.js:1	0.4 kB	86 ms
h.gif?a_key=333555a123933efd4678200cfc304de4_15643......	200	gif	5055.8b695a6f.js:1	0.4 kB	85 ms
h.gif?a_key=333555a123933efd4678200cfc304de4_15643......	200	gif	5055.8b695a6f.js:1	0.4 kB	86 ms
h.gif?a_key=333555a123933efd4678200cfc304de4_15643......	200	gif	5055.8b695a6f.js:1	0.4 kB	90 ms
89DDD0DBF0B71E3B9B8E173AF4CFC096_1730357803639....	200	webp	lib-react.672e5fa7.js:1	(disk cache)	3 ms
A92A365B2DD3235F7EDF60A2E341A686_1734499127758....	200	webp	lib-react.672e5fa7.js:1	(disk cache)	3 ms

97 requests | 201 kB transferred | 4.9 MB resources | Finish: 2.82 s | DOMContentLoaded: 602 ms | Load: 1.43 s

```
▸ o {message: 'Request failed with status code 401', name: 'AxiosError', code: 'ERR_BAD_REQUEST', config: {…}, request: XMLHttpReques
  t, …}
```

- Several notable XML requests appear:
  - https://games.mobileapi.hupu.com/1/8.0.99/bplcommentapi/bpl/score_tree/getSubGroups?outBizNo=5820&outBizType=basketball_match

Based on observation, the ***outBizNo*** parameter appears to be a numbering system corresponding to the ***groupID***. We will find out what ***groupID*** is in the next step.

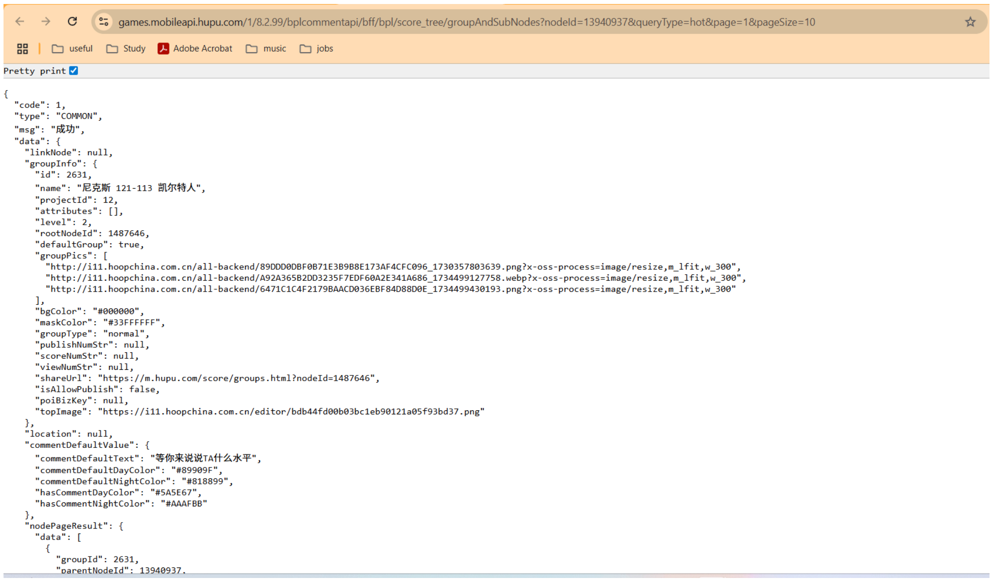

games.mobileapi.hupu.com/1/8.0.99/bplcommentapi/bpl/score_tree/getSubGroups?outBizNo=5820&outBizType=basketball_match
useful
Study
Adobe Acrobat
music
jobs
Pretty print ☑

```
{
  "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
  By inserting the groupID into the "nodeID=<groupID>" parameter, we can retrieve a JSON response containing information about a specific game (see attached screenshot). Note that this only shows one team's stats, meaning each game has two different groupIDs—one for each team.

  From this JSON response, we can extract:
  1. Game scores for each team
  2. Individual player statistics
  3. Each player's individual ***bizID*** (which becomes crucial later)

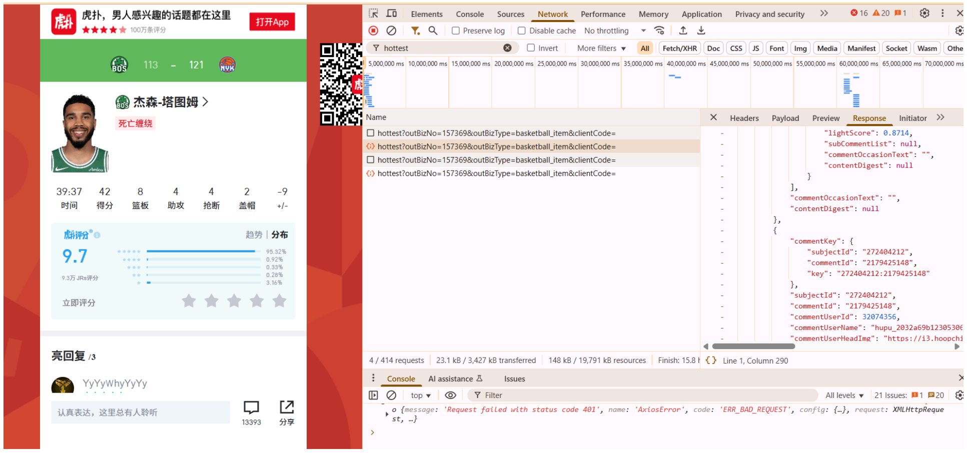
  However, this page only displays one top comment. Since I want to collect the top 3 comments for each player, I need to find something that display all the comments.



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

```
{
  "code": 1,
  "type": "COMMON",
  "msg": "成功",
  "data": {
    "linkNode": null,
    "groupInfo": {
      "id": 2631,
      "name": "尼克斯 121-113 凯尔特人",
      "projectId": 12,
      "attributes": [],
      "level": 2,
      "rootNodeId": 1487646,
      "defaultGroup": true,
      "groupPics": [
        "http://i11.hoopchina.com.cn/all-backend/89DDD0DBF0B71E3B9B8E173AF4CFC096_1730357803639.png?x-oss-process=image/resize,m_lfit,w_300",
        "http://i11.hoopchina.com.cn/all-backend/A92A365B2DD3235F7EDF60A2E341A686_1734499127758.webp?x-oss-process=image/resize,m_lfit,w_300",
        "http://i11.hoopchina.com.cn/all-backend/6471C1C4F2179BAACD036EBF84D88D0E_1734499430193.png?x-oss-process=image/resize,m_lfit,w_300"
      ],
      "bgColor": "#000000",
      "maskColor": "#33FFFFFF",
      "groupType": "normal",
      "publishNumStr": null,
      "scoreNumStr": null,
      "viewNumStr": null,
      "shareUrl": "https://m.hupu.com/score/groups.html?nodeId=1487646",
      "isAllowPublish": false,
      "poiBizKey": null,
      "topImage": "https://i11.hoopchina.com.cn/editor/bdb44fd00b03bc1eb90121a05f93bd37.png"
    },
    "location": null,
    "commentDefaultValue": {
      "commentDefaultText": "等你来说说TA什么水平",
      "commentDefaultDayColor": "#89909F",
      "commentDefaultNightColor": "#818899",
      "hasCommentDayColor": "#5A5E67",
      "hasCommentNightColor": "#AAAFBB"
    },
    "nodePageResult": {
      "data": [
        {
          "groupId": 2631,
          "parentNodeId": 13940937,
```

- https://games.mobileapi.hupu.com/1/8.2.99/bplcommentapi/bpl/comment/list/primarySingleRow/hottest?outBizNo=157369&outBizType=basketball_item&clientCode=
  After clicking into each player's individual tab (see screenshot), I observed this XML request appearing in the network traffic. The ***outBizNo*** here corresponds to the ***bizID*** of the player.



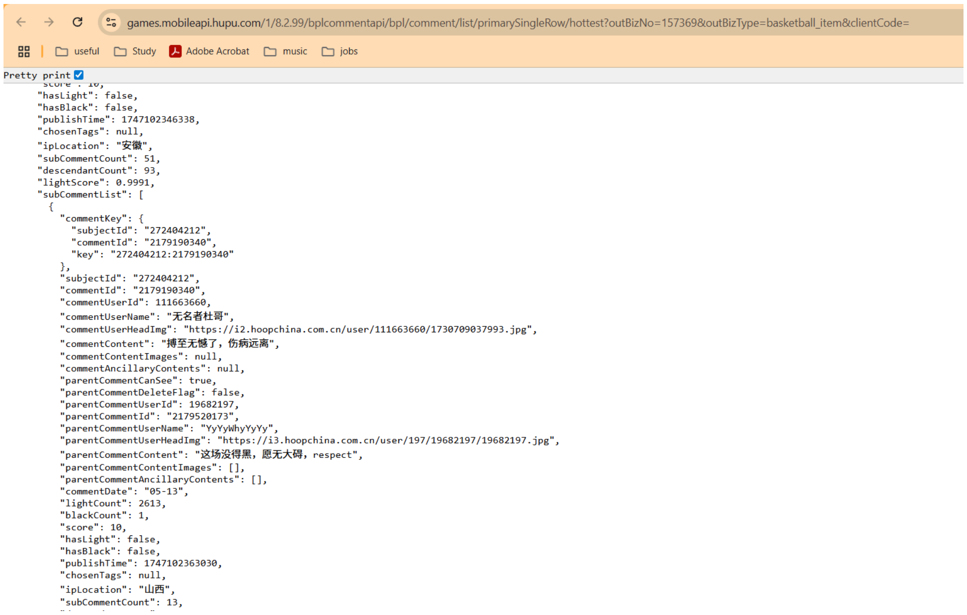
```
hottest?outBizNo=157369&outBizType=basketball_item&clientCode=
hottest?outBizNo=157369&outBizType=basketball_item&clientCode=
hottest?outBizNo=157369&outBizType=basketball_item&clientCode=
hottest?outBizNo=157369&outBizType=basketball_item&clientCode=
```

```
                "lightScore": 0.8714,
                "subCommentList": null,
                "commentOccasionText": "",
                "contentDigest": null
            }
        ],
        "commentOccasionText": "",
        "contentDigest": null
    },
    {
        "commentKey": {
            "subjectId": "272404212",
            "commentId": "2179425148",
            "key": "272404212:2179425148"
        },
        "subjectId": "272404212",
        "commentId": "2179425148",
        "commentUserId": 32074356,
        "commentUserName": "hupu_2032a69b1230530(
        "commentUserHeadImg": "https://i3.hoopchi
```

```
4 / 414 requests | 23.1 kB / 3,427 kB transferred | 148 kB / 19,791 kB resources | Finish: 15.8 h
Line 1, Column 290
```

```
o {message: 'Request failed with status code 401', name: 'AxiosError', code: 'ERR_BAD_REQUEST', config: {…}, request: XMLHttpReque
st, …}
```

- The response (see screenshot below) contains all the comments from Hupu users about this specific player in this game, with each comment accessible through the "commentContent" parameter.


games.mobileapi.hupu.com/1/8.2.99/bplcommentapi/bpl/comment/list/primarySingleRow/hottest?outBizNo=157369&outBizType=basketball_item&clientCode=
useful
Study
Adobe Acrobat
music
jobs

Pretty print ☑

```
      "score": 10,
      "hasLight": false,
      "hasBlack": false,
      "publishTime": 1747102346338,
      "chosenTags": null,
      "ipLocation": "安徽",
      "subCommentCount": 51,
      "descendantCount": 93,
      "lightScore": 0.9991,
      "subCommentList": [
        {
          "commentKey": {
            "subjectId": "272404212",
            "commentId": "2179190340",
            "key": "272404212:2179190340"
          },
          "subjectId": "272404212",
          "commentId": "2179190340",
          "commentUserId": 111663660,
          "commentUserName": "无名者杜哥",
          "commentUserHeadImg": "https://i2.hoopchina.com.cn/user/111663660/1730709037993.jpg",
          "commentContent": "搏至无憾了，伤病远离",
          "commentContentImages": null,
          "commentAncillaryContents": null,
          "parentCommentCanSee": true,
          "parentCommentDeleteFlag": false,
          "parentCommentUserId": 19682197,
          "parentCommentId": "2179520173",
          "parentCommentUserName": "YyYyWhyYyYy",
          "parentCommentUserHeadImg": "https://i3.hoopchina.com.cn/user/197/19682197/19682197.jpg",
          "parentCommentContent": "这场没得黑，愿无大碍，respect",
          "parentCommentContentImages": [],
          "parentCommentAncillaryContents": [],
          "commentDate": "05-13",
          "lightCount": 2613,
          "blackCount": 1,
          "score": 10,
          "hasLight": false,
          "hasBlack": false,
          "publishTime": 1747102363030,
          "chosenTags": null,
          "ipLocation": "山西",
          "subCommentCount": 13,
```

- Now, we have all the necessary responses needed to gather the information. We need to design a crawling script to generate all the ***groupIDs*** of basketball matches, get the game stats and the individual ***bizID*** of the player, request for the json response, retrieve the top 3 comments of the player and finally plug them all into a .csv file.

The actual implementation using scrapy can be found here: https://github.com/alexgoexercise/hupu-comments-crawler.git
Have fun crawling!