爬取豆瓣关于电影复仇者联盟3:无限战争影评

计划步骤：

1. 用scrapy爬取评论页面所有评论，存入云端MySQL
2. 分析数据
3. 生成图分析

分析：

评论地址：<https://movie.douban.com/subject/24773958/comments?status=P>

<div class="comment">

<h3>

<span class="comment-vote">

<span class="votes">25492</span>

<input value="1364036972" type="hidden">

<a href="javascript:;" class="j a\_show\_login" onclick="">有用</a>

</span>

<span class="comment-info">

<a href="https://www.douban.com/people/lingrui1995/" class="">凌睿</a>

<span>看过</span>

<span class="allstar50 rating" title="力荐"></span>

<span class="comment-time " title="2018-04-26 00:03:05">

2018-04-26

</span>

</span>

</h3>

<p class="">

<span class="short">混沌初开，宇宙出现了六颗象征尊贵身份的宝石，一旦集齐就有毁天灭地的能力。它们分别是QQ空间黄钻（心灵宝石）、QQ秀红钻（现实宝石）、QQ炫舞紫钻（力量宝石）、QQ游戏蓝钻（空间宝石）、QQ音乐绿钻（时间宝石）、QQ橙钻（灵魂宝石）。为了得到六大宝石，灭霸终于对马化腾，哦不，复联下手了。</span>

</p>

<div class="comment-report"><a rel="nofollow" href="javascript:void(0)">举报</a></div></div>

获取后页：

<a href="?start=80&amp;limit=20&amp;sort=new\_score&amp;status=P&amp;percent\_type=" data-page="" class="next">后页 &gt;</a>

用户名：<a href="https://www.douban.com/people/lingrui1995/" class="">凌睿</a>

Vote：<span class="votes">2</span>星级： <span class="allstar50 rating" title="力荐"></span>

日期：<span class="comment-time " title="2018-04-26 00:03:05">

评论： <span class="short">混沌初开，宇宙出现了六颗象征尊贵身份的宝石，一旦集齐就有毁天灭地的能力。它们分别是QQ空间黄钻（心灵宝石）、QQ秀红钻（现实宝石）、QQ炫舞紫钻（力量宝石）、QQ游戏蓝钻（空间宝石）、QQ音乐绿钻（时间宝石）、QQ橙钻（灵魂宝石）。为了得到六大宝石，灭霸终于对马化腾，哦不，复联下手了。</span></p>

mysql表格设计

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Username | vote | star | data | comment |
| 字符串 | 0~ | 1到5 | 日期 | 字符串 |

CREATE TABLE IF NOT EXISTS `douban\_1`(

`userid` INT UNSIGNED AUTO\_INCREMENT,

`username` VARCHAR(60) NOT NULL,

`vote` INT NOT NULL,

`star` TINYINT NOT NULL,

`date` DATE,

`comment` LONGTEXT NOT NULL,

PRIMARY KEY ( `userid` )

)ENGINE=InnoDB DEFAULT CHARSET=utf8;

目前达成（2018.10.15 23:01）

远程mysql连接成功，数据爬取成功

问题：

2018-10-15 23:00:19 [scrapy.spidermiddlewares.httperror] INFO: Ignoring response <403 https://movie.douban.com/subject/24773958/comments?start=340&limit=20&sort=new\_score&status=P>: HTTP status code is not handled or not allowed

2018-10-15 23:00:20 [scrapy.core.engine] DEBUG: Crawled (403) <GET https://movie.douban.com/subject/24773958/comments?start=360&limit=20&sort=new\_score&status=P> (referer: https://www.taobao.com/)

2018-10-15 23:00:20 [scrapy.core.engine] DEBUG: Crawled (403) <GET https://movie.douban.com/subject/24773958/comments?start=320&limit=20&sort=new\_score&status=P> (referer: https://www.taobao.com/)

2018-10-15 23:00:20 [scrapy.spidermiddlewares.httperror] INFO: Ignoring response <403 https://movie.douban.com/subject/24773958/comments?start=360&limit=20&sort=new\_score&status=P>: HTTP status code is not handled or not allowed

2018-10-15 23:00:20 [scrapy.core.engine] DEBUG: Crawled (403) <GET https://movie.douban.com/subject/24773958/comments?start=380&limit=20&sort=new\_score&status=P> (referer: https://www.taobao.com/)

2018-10-15 23:00:20 [scrapy.core.engine] DEBUG: Crawled (403) <GET https://movie.douban.com/subject/24773958/comments?start=220&limit=20&sort=new\_score&status=P> (referer: https://www.taobao.com/)

2018-10-15 23:00:20 [scrapy.spidermiddlewares.httperror] INFO: Ignoring response <403 https://movie.douban.com/subject/24773958/comments?start=320&limit=20&sort=new\_score&status=P>: HTTP status code is not handled or not allowed

2018-10-15 23:00:20 [scrapy.core.engine] DEBUG: Crawled (403) <GET https://movie.douban.com/subject/24773958/comments?start=400&limit=20&sort=new\_score&status=P> (referer: https://www.taobao.com/)

2018-10-15 23:00:20 [scrapy.spidermiddlewares.httperror] INFO: Ignoring response <403 https://movie.douban.com/subject/24773958/comments?start=380&limit=20&sort=new\_score&status=P>: HTTP status code is not handled or not allowed

2018-10-15 23:00:20 [scrapy.spidermiddlewares.httperror] INFO: Ignoring response <403 https://movie.douban.com/subject/24773958/comments?start=220&limit=20&sort=new\_score&status=P>: HTTP status code is not handled or not allowed

2018-10-15 23:00:20 [scrapy.core.engine] DEBUG: Crawled (403) <GET https://movie.douban.com/subject/24773958/comments?start=420&limit=20&sort=new\_score&status=P> (referer: https://www.taobao.com/)

2018-10-15 23:00:20 [scrapy.spidermiddlewares.httperror] INFO: Ignoring response <403 https://movie.douban.com/subject/24773958/comments?start=400&limit=20&sort=new\_score&status=P>: HTTP status code is not handled or not allowed

2018-10-15 23:00:20 [scrapy.spidermiddlewares.httperror] INFO: Ignoring response <403 https://movie.douban.com/subject/24773958/comments?start=420&limit=20&sort=new\_score&status=P>: HTTP status code is not handled or not allowed

2018-10-15 23:00:20 [scrapy.core.engine] DEBUG: Crawled (403) <GET https://movie.douban.com/subject/24773958/comments?start=460&limit=20&sort=new\_score&status=P> (referer: https://www.taobao.com/)

2018-10-15 23:00:20 [scrapy.spidermiddlewares.httperror] INFO: Ignoring response <403 https://movie.douban.com/subject/24773958/comments?start=460&limit=20&sort=new\_score&status=P>: HTTP status code is not handled or not allowed

2018-10-15 23:00:21 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://movie.douban.com/subject/24773958/comments?start=20&limit=20&sort=new\_score&status=P> (referer: https://www.taobao.com/)

2018-10-15 23:00:21 [scrapy.core.engine] DEBUG: Crawled (403) <GET https://movie.douban.com/subject/24773958/comments?start=440&limit=20&sort=new\_score&status=P> (referer: https://www.taobao.com/)

部分网页无法爬取，看是否能否重复爬取得到结果，说是有14万条，展示出的很少只有25页，不知是否是程序限制，明天解决。

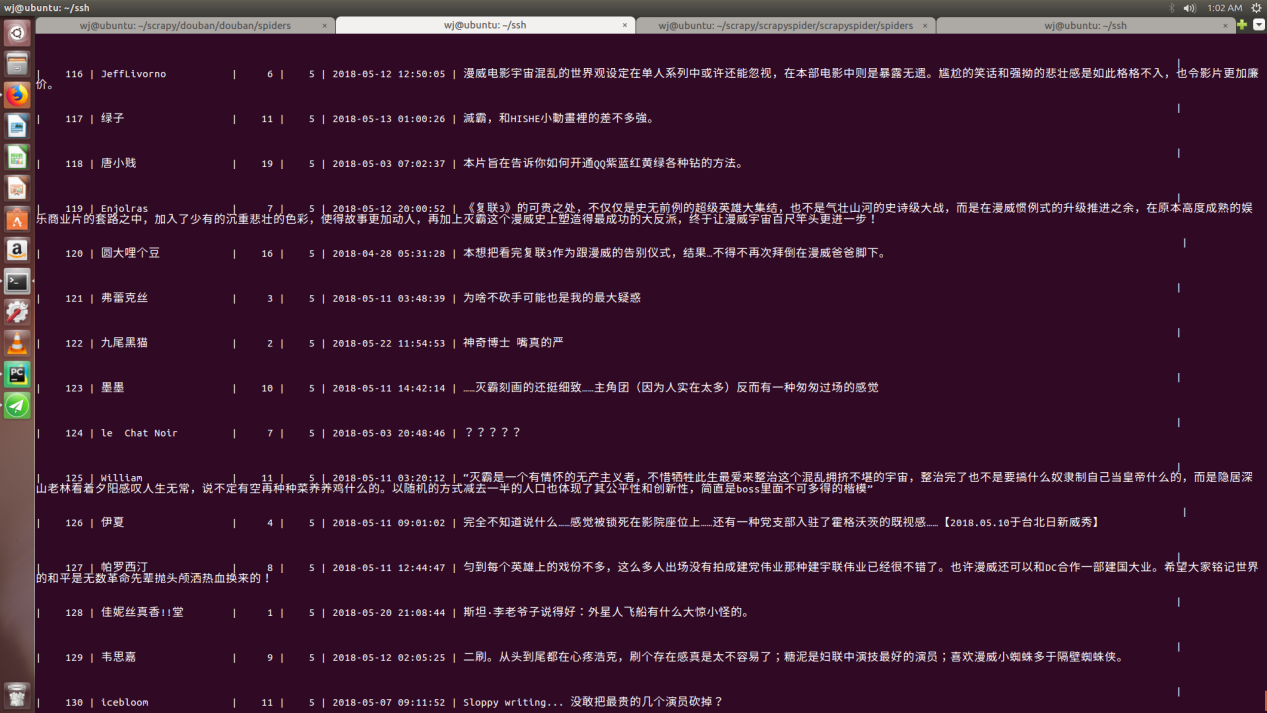
解决：Cookie问题，Cookie随时间变化，评论数量未解决

2018.10.17 0:59

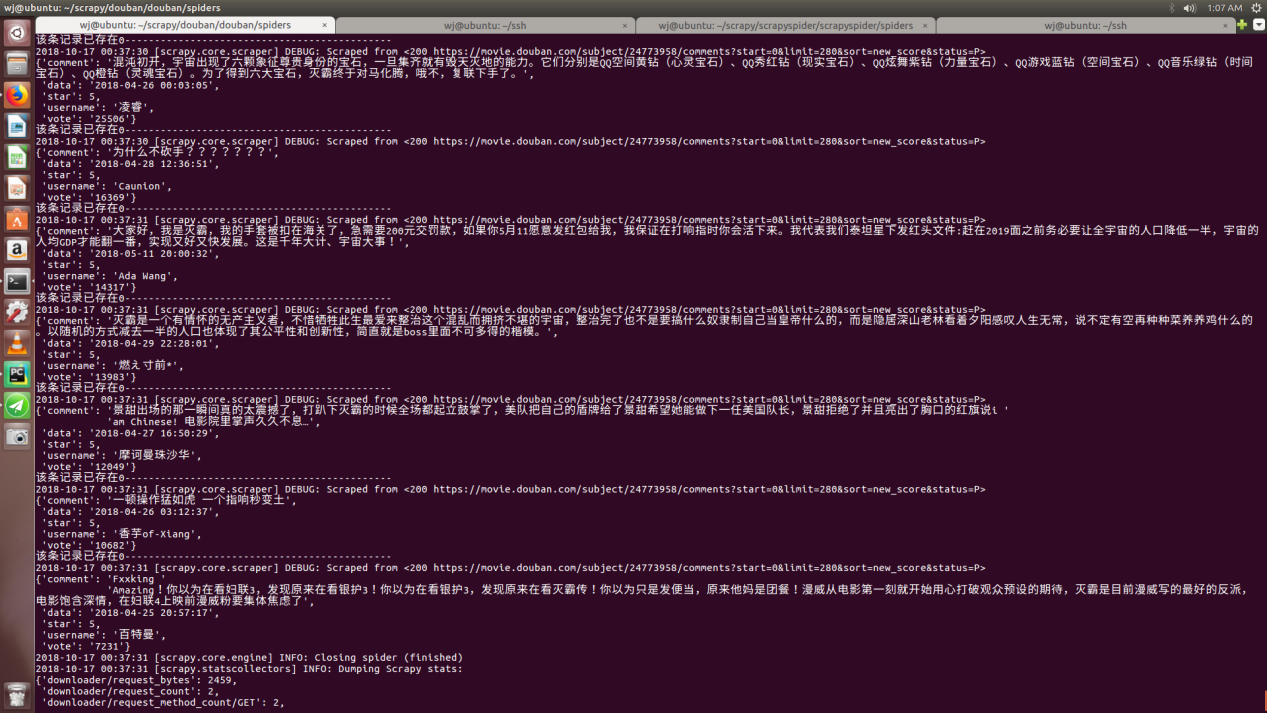
发现问题 Cookie中带有时间挫，明天进行匹配。

今日进展 数据有100+入库，由于Cookie问题，正确网址配置错误Cookie导致访问错误网页。

数据库图



程序运行图



预计数据在20\*24 480条左右，其余评论不知道是隐藏，正常访问无法访问到

2018.10.18

数据量只有150左右，scrapy的Cookie无法支撑项目爬取，决定自己写request爬取数据，通过pymysql入表。

2018.10.19 1:57

通过自己写request，通过模拟登陆抓取cookie，成功爬取到全部数据。哈哈哈

不过scrapy 应该可以通过配置cookie来获取，明天试试