环球网网站分析（<http://www.huanqiu.com>）

1. 专题类：<http://world.huanqiu.com/special/SCO2017/index.html>

<http://finance.huanqiu.com/special/beltandroad/index.html>

<http://sports.huanqiu.com/special/nbafinal2017/index.html>

第一话题：<http://opinion.huanqiu.com/debate/> 其子模块是否抓取？非.html页面

1. 新闻内容类：.html结尾网页 图解新闻（新闻正文内容为图片）是否抓取？

例：<http://world.huanqiu.com/exclusive/2017-06/10797129.html>

* 1. 标题：//\*[@class="conText"]/h1
  2. 发布时间：//\*[@id="pubtime\_baidu"] 日期格式：2017-06-08 07:35:00
  3. 媒体名称：//\*[@id="source\_baidu"]/a
  4. 正文：//\*[@id="text"]/p
  5. 评论获取方法：
     1. 在网页html源码中搜寻[id=” commentJs”]节点，如下图所示：



<script type="text/javascript" src="http://himg2.huanqiu.com/js/commentiframe.js?aid=10797129&app=cms&type=cms" id="commentJs">

取得其中三个参数：aid, app, type

如果搜索不到commentJs节点，说明该新闻无评论区

* + 1. 根据上一步获取的参数值，调用url：

[http://himg2.huanqiu.com/js/commentiframe.js?aid={aid}&app={app}&type={type}](http://himg2.huanqiu.com/js/commentiframe.js?aid=%7baid%7d&app=%7bapp%7d&type=%7btype%7d)，返回js脚本数据，如下图：



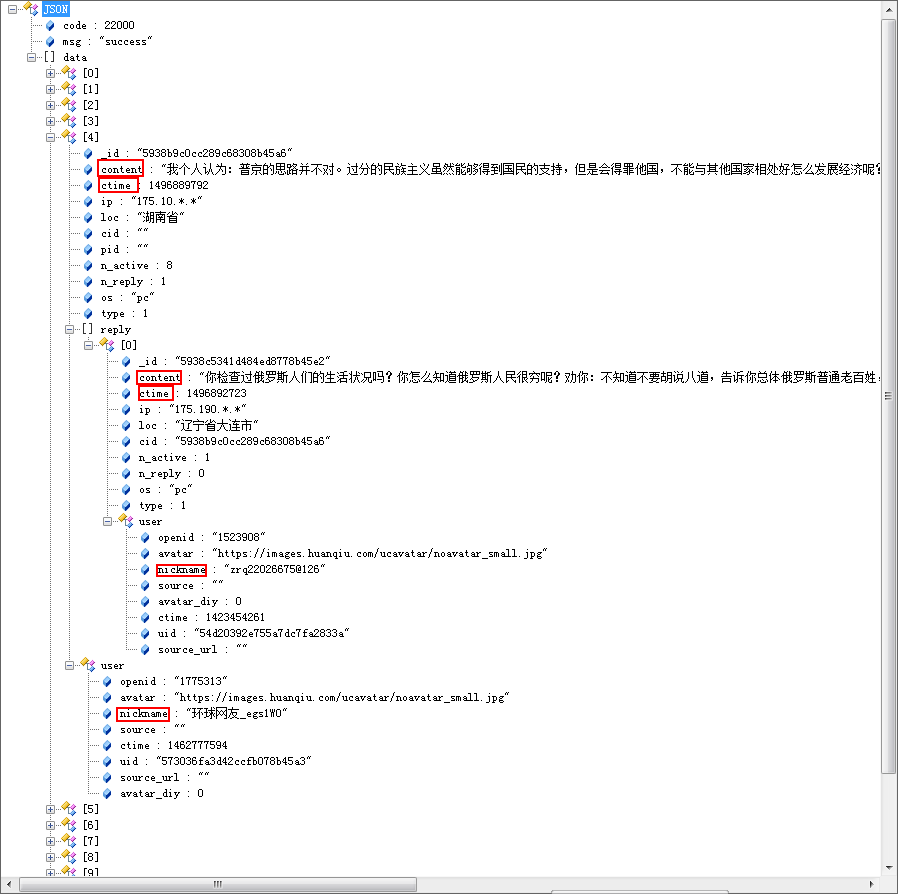
appidArr=["cms|e8fcff106c8f","zhuanti|e8fcff106c8f","cmspic|e8fcff106c8f","survey|e8fcff106c8f","sarons|5365386dfd0a","saronscms|5365386dfd0a","saronszt|5365386dfd0a","hqsb|GHif1mbXiH"]

* + 1. 在返回脚本内容中，搜索appidArr数组，并根据第一步获取的app值匹配相应的数值(如cms对应的为e8fcff106c8f)，定义为appid。调用url：<https://commentn.huanqiu.com/api/v2/async?a=comment&m=source_info&appid={appid}&sourceid={aid}&url={url}>，其中参数url为新闻的url。
    2. 上一步返回的json数据如下图所示：



* + 1. 根据返回值获取sid=data[\_id]，获取评论数data[n\_comment]，获取回复数data[n\_reply]，评论数与回复数之和为页面显示的跟贴数。
    2. 每页评论最多显示15条，由评论数计算出获取评论内容的次数。问评论url：<https://commentn.huanqiu.com/api/v2/async?a=comment&m=comment_list&sid={sid}&n=15&p={page}&appid={appid}&callback=comment_list>。返回的数据格式如下图所示：



将数据格式化，取comment\_list中的jason数据，格式化后的内容，如下图所示：

其中，评论内容：data[content]，评论时间毫秒数：data[ctime]，评论人昵称：data[user][nickname]，评论回复数：data[reply].length。

1. 图片类：
   1. 标题：//\*[@class="focus\_box"]/h1/strong
   2. 发布时间：//\*[@class="time"]/div/span 日期格式：2017-06-07 10:31
   3. 媒体名称：//\*[@class="from"]/div/span
   4. 正文：//\*[@id="img\_txt"]
   5. 评论获取方法：同新闻类