

多来源爬虫汇总

此处以中美贸易战为要爬取的内容(共爬取数据约 15 万)

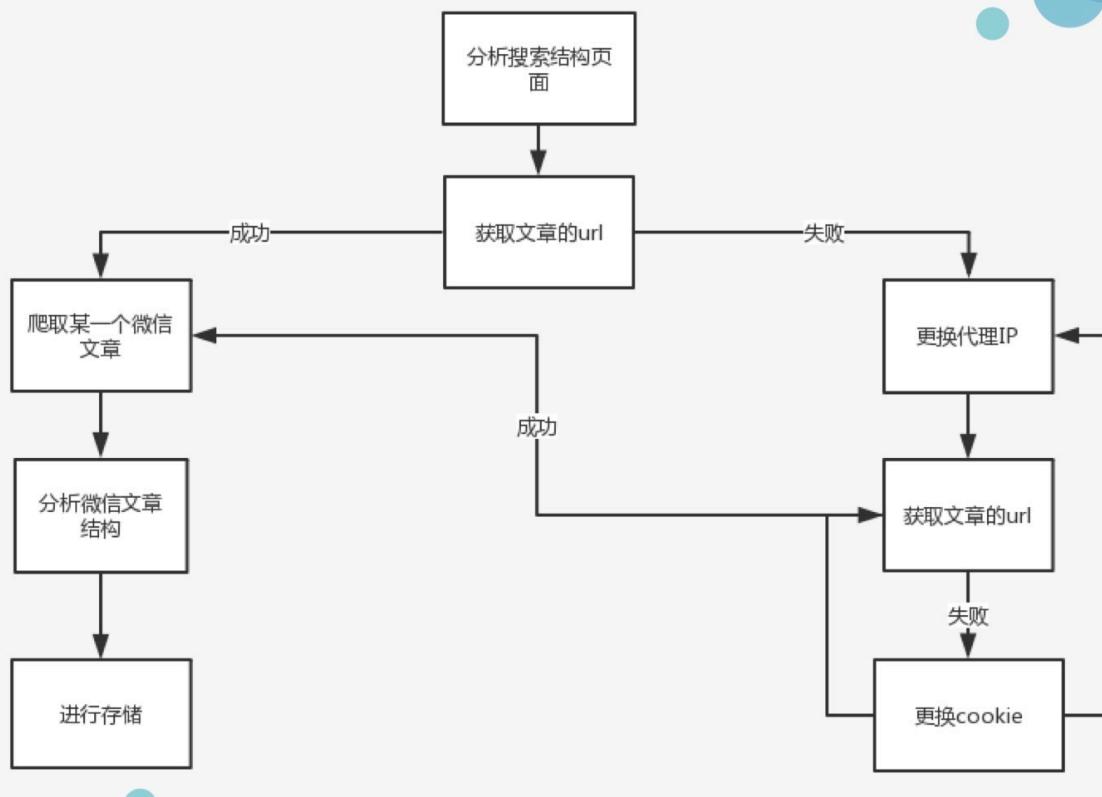
爬虫来源 1

微信文章的爬取

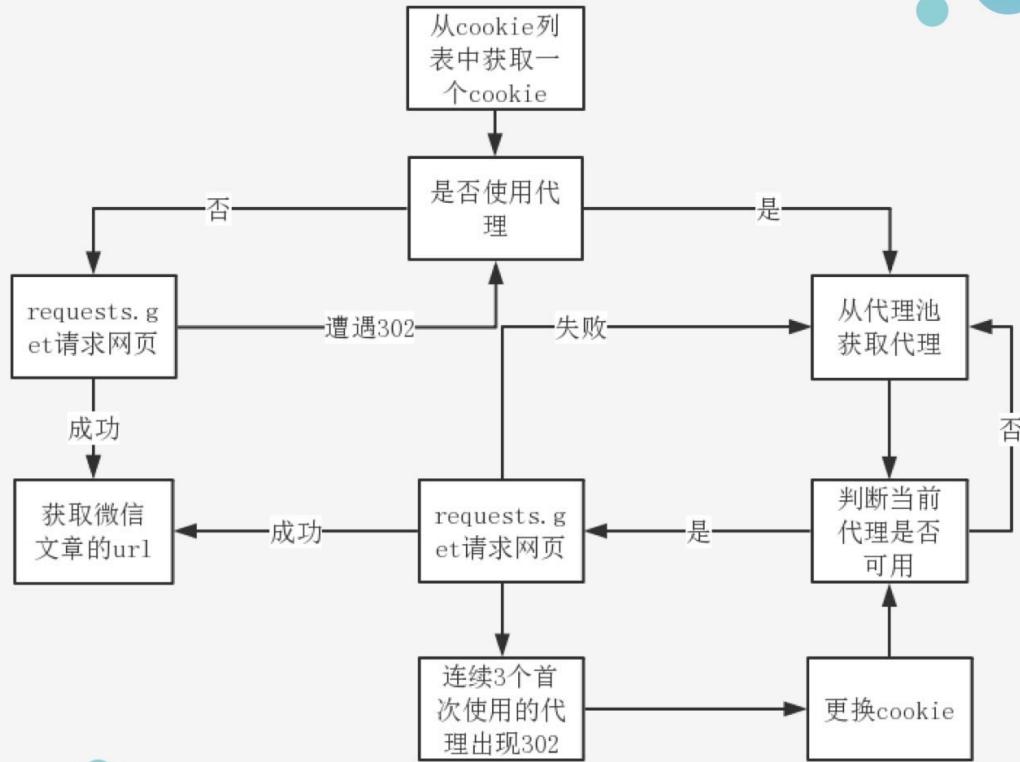
爬取策略: 因为微信文章的爬取时,微信对于每一个 ip 都有爬取有限制

解决方案: 所以这里要用到代理池的方法,代理池是 Redis 数据库中存储着大量的 IP 地址以供当一个 ip 被封了的时候,我么通过在 Redis 中存储的 ip 来替换,如果替换成功,则继续爬取,如果替换失败,则再在 Redis 中寻找下一个 Redis 进行爬取.

流程图



获取文章url

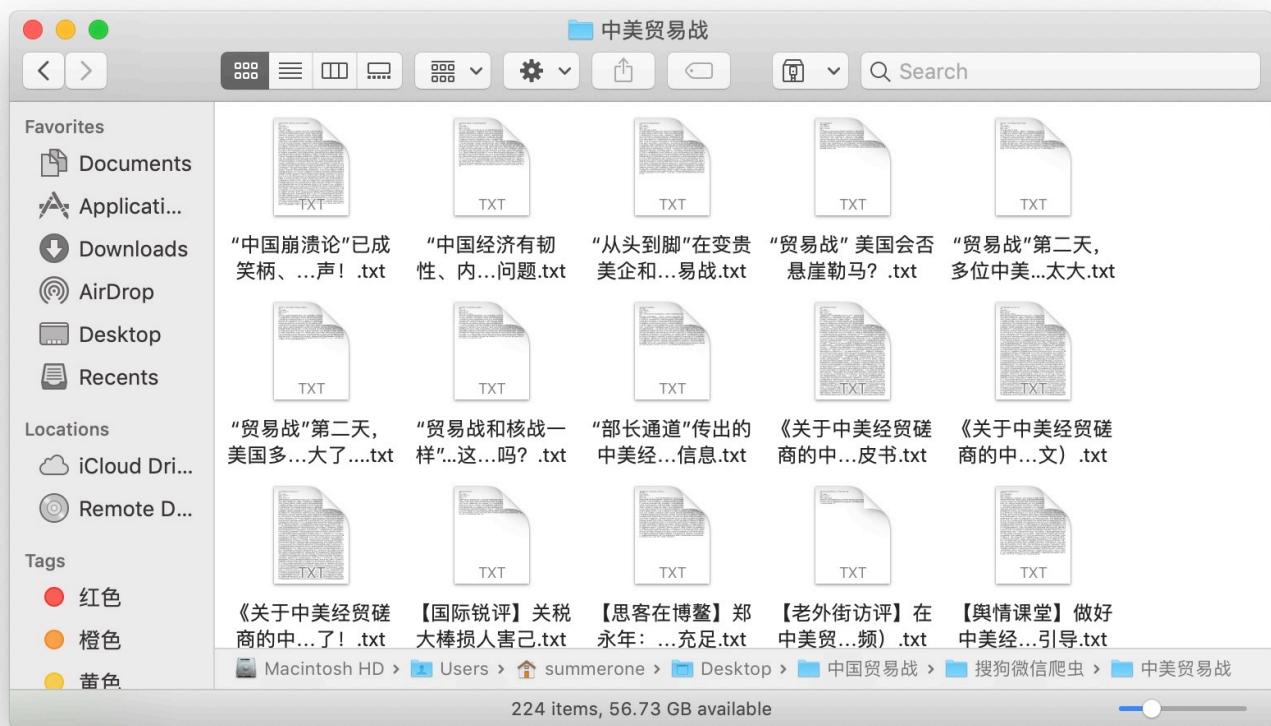


Redis 数据库代理池:

2.Redis数据库

row	value
1	121.232.194.196:9000
2	121.232.194.196:9000
3	125.40.109.154:61875
4	219.159.38.201:56210
5	222.128.9.235:33428
6	163.204.246.231:9999
7	125.40.109.154:61875

爬取结果:



爬虫来源 2

| 中新网(爬取了 300 多篇)

爬取策略:

1. 首先通过爬取中新网主页面上关于中美贸易战的所有新闻的 url 并存储在中新网.txt中



```
http://www.chinanews.com/cj/2019/06-22/8872123.shtml
http://www.chinanews.com/cj/2019/06-22/8872086.shtml
http://www.chinanews.com/gj/2019/06-22/8871931.shtml
http://www.chinanews.com/gn/2019/06-21/8871286.shtml
http://www.chinanews.com/stock/2019/06-21/8871165.shtml
http://www.chinanews.com/business/2019/06-21/8871119.shtml
http://www.chinanews.com/ll/2019/06-18/8868224.shtml
http://www.chinanews.com/gn/2019/06-18/8867728.shtml
http://www.chinanews.com/gj/2019/06-18/8867724.shtml
http://www.chinanews.com/gn/2019/06-17/8866760.shtml
http://www.chinanews.com/sh/2019/06-17/8866418.shtml
http://www.chinanews.com/cj/2019/06-16/8865993.shtml
http://www.chinanews.com/gn/2019/06-15/8865663.shtml
http://www.chinanews.com/gn/2019/06-13/8864320.shtml
http://www.chinanews.com/gn/2019/06-13/8864312.shtml
http://www.chinanews.com/gn/2019/06-13/8863858.shtml
http://www.chinanews.com/business/2019/06-13/8863622.shtml
http://www.chinanews.com/gn/2019/06-12/8863006.shtml
http://www.chinanews.com/gn/2019/06-12/8862988.shtml
http://www.chinanews.com/gn/2019/06-11/8861039.shtml
http://www.chinanews.com/kong/2019/06-10/8860305.shtml
http://www.chinanews.com/ll/2019/06-10/8860027.shtml
http://www.chinanews.com/gn/2019/06-08/8859153.shtml
http://www.chinanews.com/sh/2019/06-07/8858721.shtml
http://www.chinanews.com/gn/2019/06-06/8858426.shtml
http://www.chinanews.com/gn/2019/06-06/8858282.shtml
http://www.chinanews.com/gn/2019/06-06/8858132.shtml
http://www.chinanews.com/gn/2019/06-06/8857488.shtml
http://www.chinanews.com/gn/2019/06-06/8857440.shtml
http://www.chinanews.com/gn/2019/06-04/8855989.shtml
http://www.chinanews.com/cj/2019/06-04/8855372.shtml
http://www.chinanews.com/gn/2019/06-03/8855240.shtml
http://www.chinanews.com/tw/2019/06-03/8854608.shtml
http://www.chinanews.com/gn/2019/06-03/8854415.shtml
http://www.chinanews.com/gn/2019/06-03/8854387.shtml
http://www.chinanews.com/cj/2019/06-02/8853894.shtml
http://www.chinanews.com/cj/2019/06-02/8853846.shtml
http://www.chinanews.com/gn/2019/06-01/8853687.shtml
http://www.chinanews.com/gj/2019/06-01/8853379.shtml
http://www.chinanews.com/gn/2019/06-01/8853332.shtml
http://www.chinanews.com/gn/2019/05-31/8853272.shtml
http://www.chinanews.com/gn/2019/05-31/8852585.shtml
http://www.chinanews.com/mil/shipin/cns/2019/05-30/news817843.shtml
http://www.chinanews.com/mil/2019/05-30/8852151.shtml
http://www.chinanews.com/mil/2019/05-30/8851731.shtml
http://www.chinanews.com/gn/2019/05-30/8851665.shtml
http://www.chinanews.com/gn/2019/05-29/8850196.shtml
http://www.chinanews.com/gn/2019/05-29/8850012.shtml
http://www.chinanews.com/cj/2019/05-28/8848772.shtml
http://www.chinanews.com/ll/2019/05-27/8848492.shtml
http://www.chinanews.com/gn/2019/05-27/8847763.shtml
http://www.chinanews.com/gn/2019/05-27/8847762.shtml
http://www.chinanews.com/cj/2019/05-26/8847408.shtml
http://www.chinanews.com/gj/2019/05-25/8846889.shtml
http://www.chinanews.com/gn/2019/05-24/8846500.shtml
```

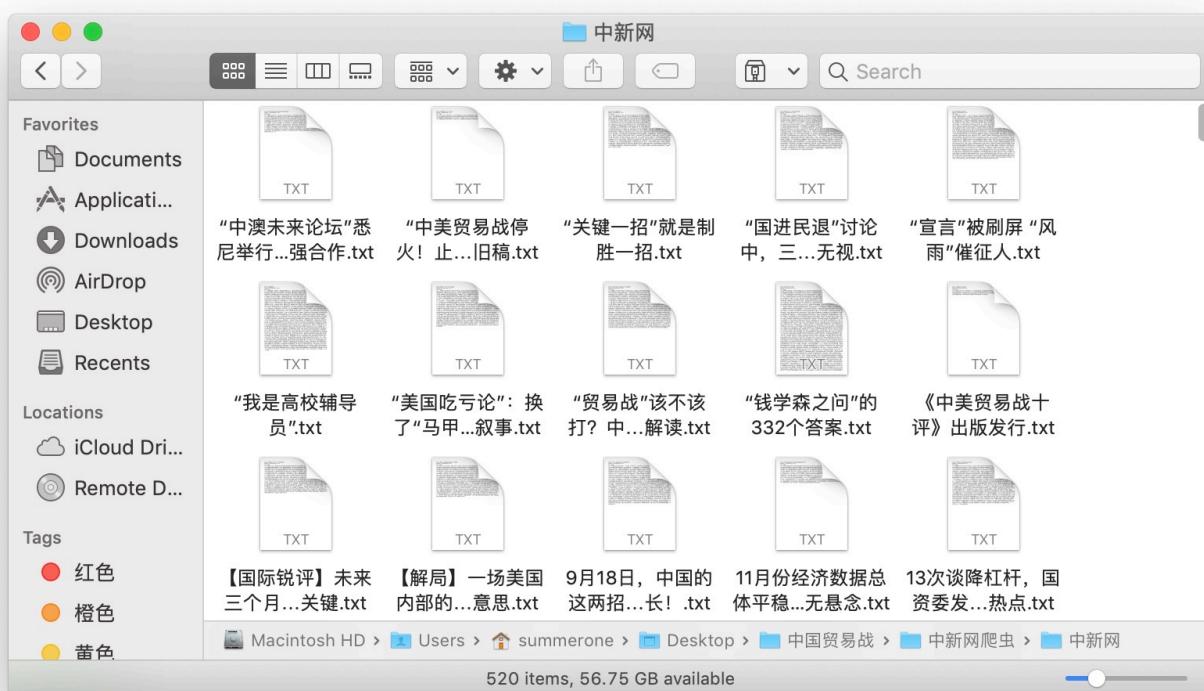
2. 然后对 txt 中每一个 url 进行爬取(详见中新网.py)

其中出现的问题:爬取中新网不同网页的编码方式不同,所以爬到了很多的乱码

解决方法:通过对不同编码方式的页面进行识别,观察到若是"gb2312"进行编码的网页,则会在网页中出现中文的"<--pc 和手机适配代码开始 -->",其余的网页都是 utf-8 编码,所以在对网页的内容进行解析时,通过 beautifulsoup 匹配<!-- --!>中的内容,来判断新闻的内容

```
r.encoding = 'utf-8'
test=re.findall(re.compile(r'<!--(.*)-->', re.S), r.text)[1]
print(test)
if test!='pc和手机适配代码开始':
    r.encoding='gb2312'
```

3. 爬取结果



爬取来源 3

新浪财经(爬取接近 2 万条)

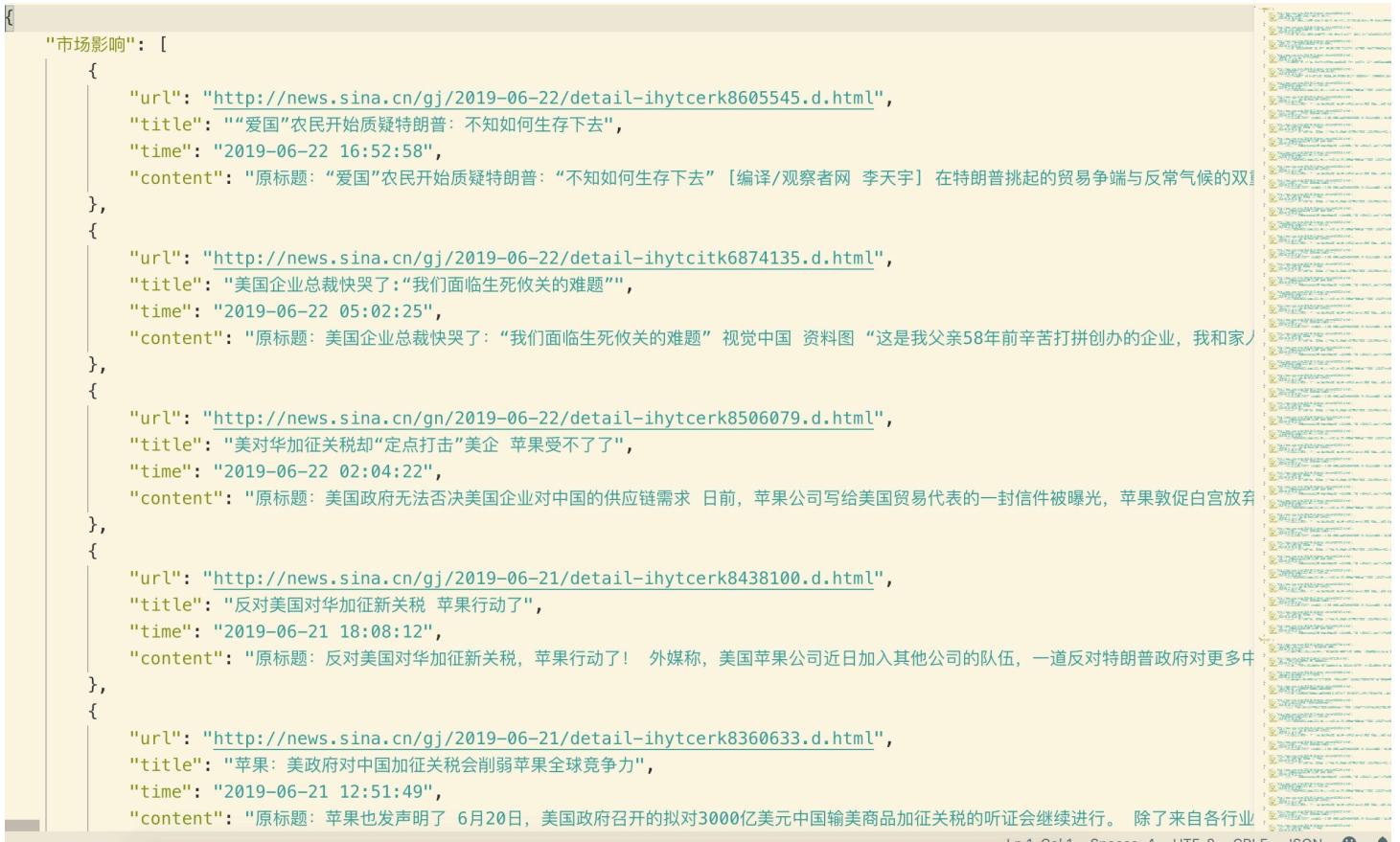
爬取策略与以上中新网相似,但是通过观察,我发现并不是所有的新闻都是我们所需要的,我们更加关心的是例如"市场影响","不可靠实体清单"这些比较敏感言论的新闻,所以在新浪财经中我的爬取策略是对我所需要的有特殊言论的新闻进行爬取

```

for i in range(8):
    url = 'https://interface.sina.cn/wap_api/wap_std_subject_feed_list.d.json?component_id=_conf_34|wap_zt_std_theme_feed|http://news.sina.cn/zt_d/maoyi0109&page=2&_=1561260054831&callback=Zepto1561260046'+str(a)
    print(url)
    get2(url,'市场影响',a)
    a = a + 2
for i in range(7):
    url = 'https://interface.sina.cn/wap_api/wap_std_subject_feed_list.d.json?component_id=_conf_34|wap_zt_std_theme_feed|http://news.sina.cn/zt_d/maoyi0109&page=2&_=1561260054831&callback=Zepto1561260046'+str(a)
    print(url)
    get2(url,'分析评论',a)
    a = a + 2
for i in range(4):
    url = 'https://interface.sina.cn/wap_api/wap_std_subject_feed_list.d.json?component_id=_conf_34|wap_zt_std_theme_feed|http://news.sina.cn/zt_d/maoyi0109&page=2&_=1561260054831&callback=Zepto1561260046'+str(a)
    print(url)
    get2(url,'中方回应',a)
    a = a + 2
for i in range(1):
    url = 'https://interface.sina.cn/wap_api/wap_std_subject_feed_list.d.json?component_id=_conf_34|wap_zt_std_theme_feed|http://news.sina.cn/zt_d/maoyi0109&page=2&_=1561260054831&callback=Zepto1561260046'+str(a)
    print(url)
    get2(url,'不可靠实体清单制度',a)
    a = a + 2
for i in range(2):
    url = 'https://interface.sina.cn/wap_api/wap_std_subject_feed_list.d.json?component_id=_conf_34|wap_zt_std_theme_feed|http://news.sina.cn/zt_d/maoyi0109&page=2&_=1561260054831&callback=Zepto1561260046'+str(a)
    print(url)
    get2(url,'赴美提醒',a)
    a = a + 2
for i in range(1):
    url = 'https://interface.sina.cn/wap_api/wap_std_subject_feed_list.d.json?component_id=_conf_34|wap_zt_std_theme_feed|http://news.sina.cn/zt_d/maoyi0109&page=2&_=1561260054831&callback=Zepto1561260046'+str(a)
    print(url)
    get2(url,'中国经济底气',a)
    a = a + 2

```

最终以 json 的格式进行存储



```

{
  "market影響": [
    {
      "url": "http://news.sina.cn/gj/2019-06-22/detail-ihytcerk8605545.d.html",
      "title": "“爱国”农民开始质疑特朗普：不知如何生存下去",
      "time": "2019-06-22 16:52:58",
      "content": "原标题：“爱国”农民开始质疑特朗普：“不知如何生存下去”【编译/观察者网 李天宇】在特朗普挑起的贸易争端与反常气候的双重夹击下，美国不少农民开始对特朗普政府失去信心。"
    },
    {
      "url": "http://news.sina.cn/gj/2019-06-22/detail-ihytcitk6874135.d.html",
      "title": "美国企业总裁快哭了：“我们面临生死攸关的难题”",
      "time": "2019-06-22 05:02:25",
      "content": "原标题：美国企业总裁快哭了：“我们面临生死攸关的难题” 视觉中国 资料图 “这是我父亲58年前辛苦打拼创办的企业，我和家人...”
    },
    {
      "url": "http://news.sina.cn/gn/2019-06-22/detail-ihytcerk8506079.d.html",
      "title": "美对华加征关税却“定点打击”美企 苹果受不了了",
      "time": "2019-06-22 02:04:22",
      "content": "原标题：美国政府无法否决美国企业对中国的供应链需求 日前，苹果公司写给美国贸易代表的一封信件被曝光，苹果敦促白宫放弃...”
    },
    {
      "url": "http://news.sina.cn/gj/2019-06-21/detail-ihytcerk8438100.d.html",
      "title": "反对美国对华加征新关税 苹果行动了",
      "time": "2019-06-21 18:08:12",
      "content": "原标题：反对美国对华加征新关税，苹果行动了！ 外媒称，美国苹果公司近日加入其他公司的队伍，一道反对特朗普政府对更多中...”
    },
    {
      "url": "http://news.sina.cn/gj/2019-06-21/detail-ihytcerk8360633.d.html",
      "title": "苹果：美政府对中国加征关税会削弱苹果全球竞争力",
      "time": "2019-06-21 12:51:49",
      "content": "原标题：苹果也发声明了 6月20日，美国政府召开的拟对3000亿美元中国输美商品加征关税的听证会将继续进行。 除了来自各行业...”
    }
  ]
}

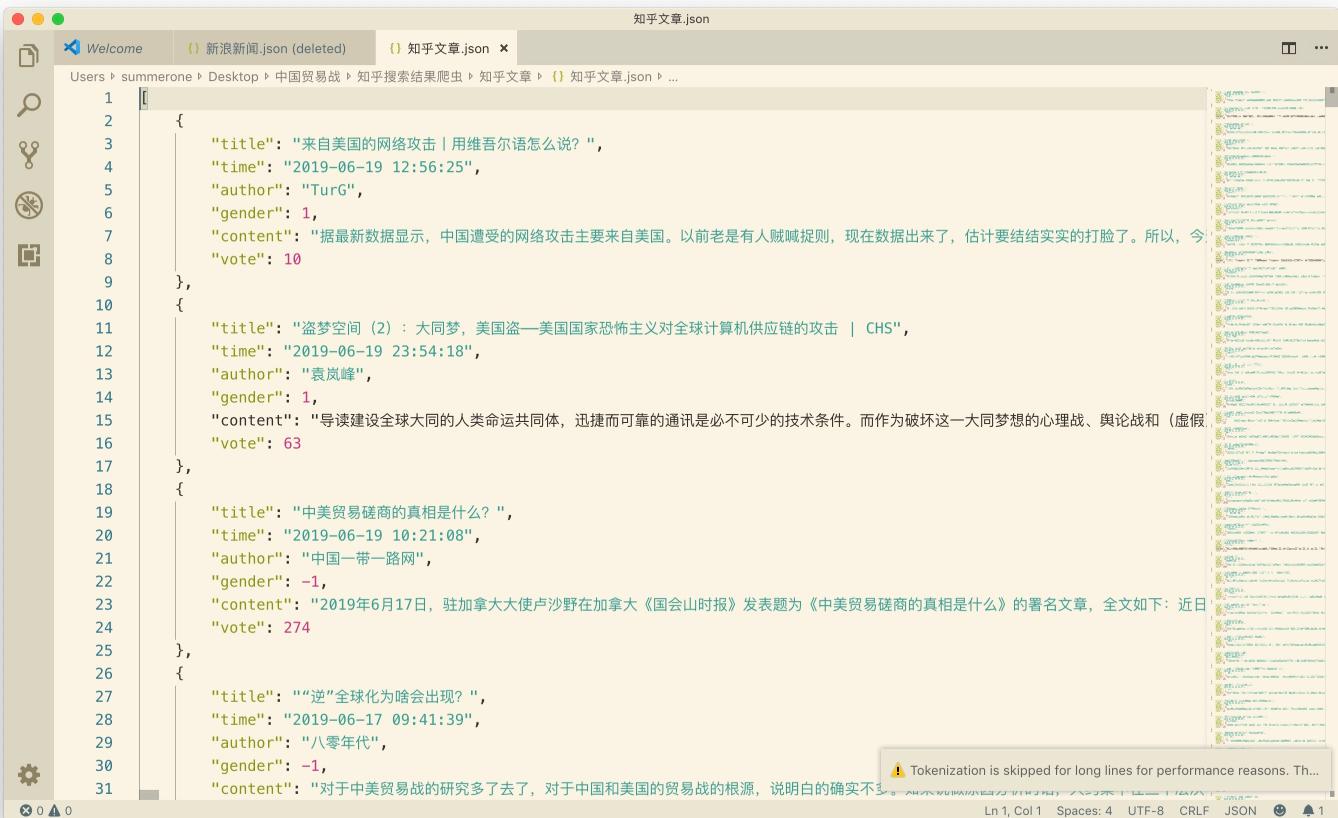
```

Ln 1. Col 1 Spaces: 4 CRLF JSON

爬虫来源 4

知乎的话题和评论

我们知道,知乎的结构和新闻是不同的,知乎是基于话题和评论的,而不是像新闻那样是结构性的,所以我在对知乎进行爬取时,分别对关于中美贸易战的话题和问答进行爬取
文章内容存储:



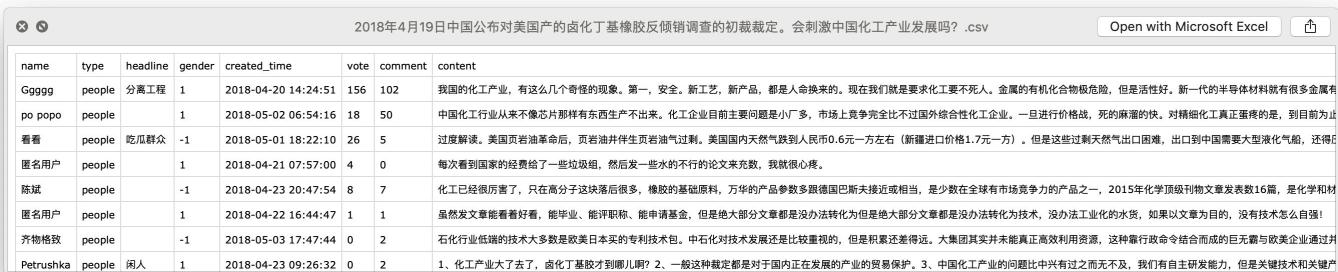
```
1 {  
2   "title": "来自美国的网络攻击 | 用维吾尔语怎么说?",  
3   "time": "2019-06-19 12:56:25",  
4   "author": "TurG",  
5   "gender": 1,  
6   "content": "据最新数据显示,中国遭受的网络攻击主要来自美国。以前老是有人贼喊捉贼,现在数据出来了,估计要结结实实的打脸了。所以,今  
7   "vote": 10  
8 },  
9 {  
10   "title": "盗梦空间 (2) : 大同梦, 美国盗—美国国家恐怖主义对全球计算机供应链的攻击 | CHS",  
11   "time": "2019-06-19 23:54:18",  
12   "author": "袁岚峰",  
13   "gender": 1,  
14   "content": "导建设全球大同的人类命运共同体,迅捷而可靠的通讯是必不可少的技术条件。而作为破坏这一大同梦想的心理战、舆论战和(虚假  
15   "vote": 63  
16 },  
17 {  
18   "title": "中美贸易磋商的真相是什么?",  
19   "time": "2019-06-19 10:21:08",  
20   "author": "中国一带一路网",  
21   "gender": -1,  
22   "content": "2019年6月17日,驻加拿大使卢沙野在加拿大《国会山时报》发表题为《中美贸易磋商的真相是什么?》的署名文章,全文如下: 近日  
23   "vote": 274  
24 },  
25 {  
26   "title": "“逆”全球化为啥会出现?",  
27   "time": "2019-06-17 09:41:39",  
28   "author": "八零年代",  
29   "gender": -1,  
30   "content": "对于中美贸易战的研究多了去了,对于中国和美国的贸易战的根源,说明白的确实不多。我自己的理解是,美国的逆全球化政策  
31 }  
32 }
```

⚠ Tokenization is skipped for long lines for performance reasons. This may result in inaccurate search results.

Ln 1, Col 1 Spaces: 4 UTF-8 CRLF JSON

问答内容:

此处以"中国公告对美反倾销裁定"的回答为例

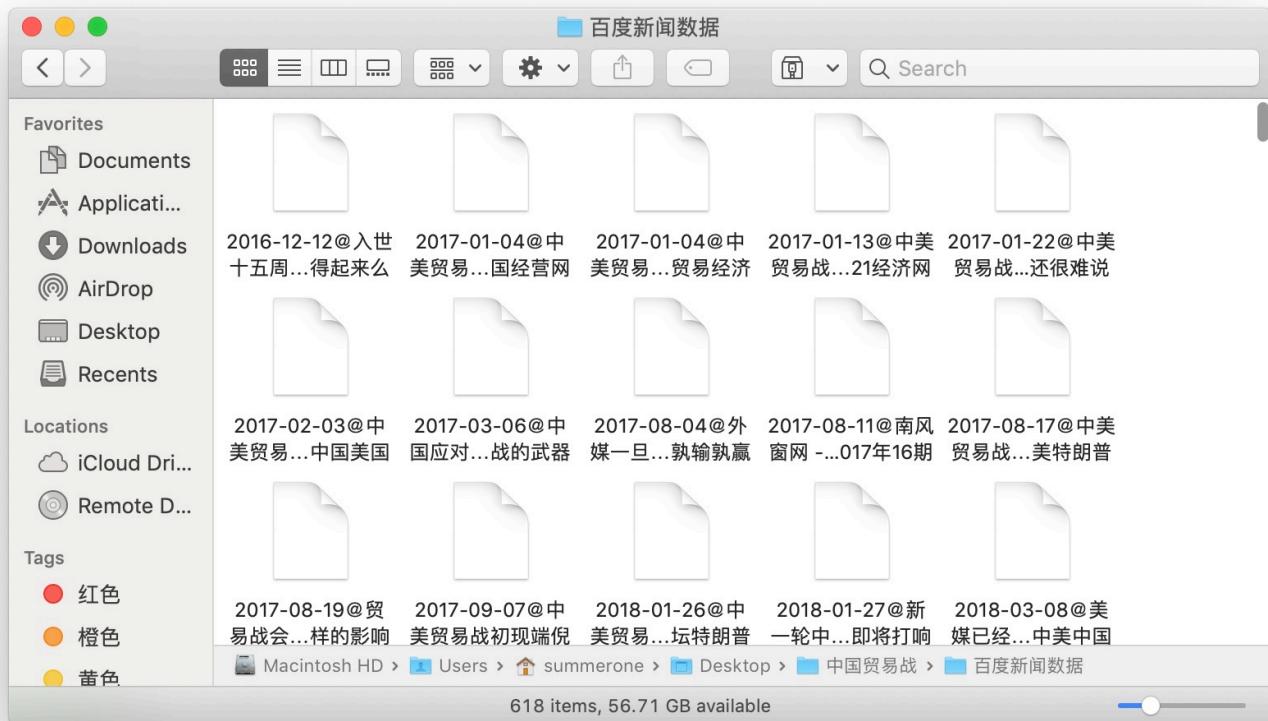


2018年4月19日中国公布对美国产的卤化丁基橡胶反倾销调查的初裁裁定。会刺激中国化工产业发展吗?.csv							Open with Microsoft Excel
name	type	headline	gender	created_time	vote	comment	content
Ggggg	people	分离工程	1	2018-04-20 14:24:51	156	102	我国的化工产业,有这么几个奇怪的现象。第一,安全。新工艺,新产品,都是人命换来的。现在我们就是要求化工要不死人。金属的有机化合物极危险,但是活性好。新一代的半导体材料就有很多金属有
po popo	people		1	2018-05-02 06:54:16	18	50	中国化工行业从来不像芯片那样有东西生产不出来。化工企业目前主要问题是小厂多,市场上竞争完全比不过国外综合性化工企业。一旦进行价格战,死的麻溜的快。对精细化工真正蛋疼的是,到目前为止
看看	people	吃瓜群众	-1	2018-05-01 18:22:10	26	5	过度解读。美国页岩油革命后,页岩油井伴生页岩气过剩。美国内天然气跌到人民币0.6元一方左右(新疆进口价格1.7元一方)。但是这些过剩天然气出口困难,出口到中国需要大型液化气船,还得
匿名用户	people		1	2018-04-21 07:57:00	4	0	每次看到国家的经费给了些垃圾组,然后发一些水的不行的论文来充数,我就很心疼。
陈斌	people		-1	2018-04-23 20:47:54	8	7	化工已经很厉害了,只在高分子这块落后很多。橡胶的基础原料,万华的产品参数多跟德国巴斯夫接近或相当,是少数在全球有市场竞争力的产品之一。2015年化学顶级刊物文章发表数16篇,是化学和材
匿名用户	people		1	2018-04-22 16:44:47	1	1	虽然发文章能看着好看,能毕业、能评职称、能申请基金,但是绝大部分文章都是没办法转化为技术,但是绝大部分文章都是没办法转化为技术,没办法工业化的干货,如果以文章为目的,没有技术怎么自强!
齐物格致	people		-1	2018-05-03 17:47:44	0	2	石化行业低端的技术大多数是欧美日本买的专利技术包。中石化对技术发展还是比较重视的,但是积累还差得远。大集团其实并未能真正高效利用资源,这种靠行政命令结合而成的巨无霸与欧美企业通过并
Petrushka	people	闲人	1	2018-04-23 09:26:32	0	2	1、化工产业大了去了,卤化丁基胶才到哪儿啊?2、一般这种裁定都是对于国内正在发展的产业的贸易保护。3、中国化工产业的问题比中兴有过之而不及,我们有自主研发能力,但是关键技术关键产

爬取来源 5

百度新闻

爬取这个来源不太费力,利用 Selenium 来模拟浏览器来进行爬取就可以爬下来



爬取来源 6

今日头条

爬取今日头条时,因为是动态的网页,每次都在重定向.所以不能利用以前爬取静态网页的爬取策略,所以这里我们利用 Ajax 异步加载返回来的页面来进行 url 的爬取来获得对应文章的url,通过构造不同文章的 url 来爬取动态加载的网页

如图为构造网页 url 的过程,通过观察对应新闻的 url 来构造对应的 url 即可,但是今日头条的 url 有加密算法,但是网上都有相关的破解方法

```
def parser_page_index(a):
    while True:
        url = 'https://www.toutiao.com/api/search/content/?aid=24&app_name=web_search&offset=' + str(a) +
        '&format=json&keyword=%E4%B8%AD%E7%BE%8E%E8%B4%B8%E6%98%93%E6%88%98&autoload=true&count=20&en_qc=1&cur_tab=1&from=search_tab&pd=synthesis'
        respond = requests.get(url, headers=headers)
        dict1=respond.json()
        with open('今日头条.json', 'w', encoding='utf-8')as f:
            json.dump(dict1, f, indent=4, ensure_ascii=False)
        id1 = re.findall(re.compile(r"'group_id': '(.*)'", re.S), str(dict1))
        id.extend(id1)
        a=a+20
        if dict1['has_more']!=1:
            break
```

爬取结果:

```
1  [
2  {
3      "type": "国际",
4      "title": "中美贸易战进入全新阶段，中美都有三个没想到！",
5      "time": "2018-09-24 17:51:04",
6      "news": "（一）来而不往非礼也。中国的这句古话，今天说来显得特别妥帖。今天，9月24日，中国传统佳节中秋节，美国送来了一份“厚礼”：对我2
7  },
8  {
9      "type": "国际",
10     "title": "中美贸易战打打停停，这次是否能再次迎来转机？",
11     "time": "2019-06-21 14:01:13",
12     "news": "今天，我们从一通重要电话说开去。6月18日深夜，习近平应约同美国总统特朗普通电话。中美贸易战打打停停，这次是否能再次迎来转机呢？
13  },
14  {
15      "type": "国际",
16      "title": "美国专家发现 中国如此巧妙应对美贸易战",
17      "time": "2019-06-20 12:04:52",
18      "news": "参考消息网6月20日报道美国《大西洋》月刊网站6月18日发表华盛顿彼得森国际经济研究所专家查德·鲍恩的文章称，中国通过降低除美国1
19  },
20  {
21      "type": "其他",
22      "title": "中美贸易战对百姓影响 对普通人生活带来什么",
23      "time": "2019-05-14 13:47:20",
24      "news": "美方挑起贸易战，中方不得不采取必要的反制措施。大家最关心的，莫过于中方的反制会给企业和老百姓带来多大影响？我们会用什么办法来
25  },
26  {
27      "type": "财经",
28      "title": "港媒：中美贸易战其实无关贸易，而是这两大事件在推动",
29      "time": "2018-08-20 13:34:43",
30      "news": "香港《南华早报》网站8月17日发表香港中文大学经济学教授刘遵义的文章《中美贸易战的背后是两国力量竞争与贸易保护主义浪潮高涨》称，中
31  },
```

Ln 1, Col 1 Spaces: 4 UTF-8 with BOM CRLF JSON

爬取来源 7

twitter

这一部分我不是太熟,直接就是翻墙然后利用 Selenium 模拟浏览器进行强行的爬,爬的效率比较慢,但是为了体现数据的多元性,也爬了一点

This screenshot shows a Microsoft Excel spreadsheet titled "中美贸易战Twitter数据". The data consists of approximately 100 rows of tweet information. Each row includes fields for ID, date, author, content, and various engagement metrics (likes, retweets, etc.). The content of the tweets spans from June 17, 2019, to June 18, 2019.

爬取来源 8

微博

```
headers = {
    'accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8',
    'accept-encoding': 'gzip, deflate, sdch, br',
    'cookie': '_T_WM=e44257ff22941fdb16a125fbac54c83a; ALF=1563517107; SCF=AijEvzIXGeu1jnCQ0KPxP2m2eH09-GvBRTN-veHdpEm8nkDvVtTavo05',
    'user-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.36'
}

client = pymongo.MongoClient('mongodb://localhost:27017/')
db = client['weibo']
collection = db['trade']
# 添加page数 (92)
base_url = 'https://weibo.cn/search/mblog?hideSearchFrame=&keyword=%E4%B8%AD%E7%BE%8E%E8%B4%B8%E6%98%93%E6%88%98&advancedfilter=1&s'
```

通过构造自己的头部,添加爬取相关爬取的参数,并利用 pyquery 进行解析爬取相关的网页,将爬取到的内容存到 MongoDB 中

```
try:
    inf = re.findall('赞\[.*?\]\.*转发\[.*?\]\.*评论\[.*?\]\.*收藏 (.*) 来自', c.text())[0]
except:
    inf = re.findall('赞\[.*?\]\.*转发\[.*?\]\.*评论\[.*?\]\.*收藏 (.*)', c.text())[0]
item['zan'] = inf[0]
item['zhuanfa'] = inf[1]
item['pinglun'] = inf[2]
item['time'] = inf[3]
item['content'] = content
item['_id'] = str(inf)
```

爬取来源 9

搜狗新闻

这个比较好爬,没有什么反爬措施,直接进行构造爬取网页的页数,请求页面就可以了