资料仅供参考，切勿用于非法用途，请在24小时内学习后删除。
造成任何后果，与up无关!

1 说在前面

因为我这个是 jupyter，比较智能，不需要 print() 就可以输出东西
如果是 pycharm 之类的，需要写 print(xxx) 才能输出东西到

In [1]:

```
5+5
```

Out[1]:

```
10
```

1.1 分类

·搜索引擎爬虫
·精确爬虫

1.2 robots 协议（君子协议）

Robots协议（也称为爬虫协议、机器人协议等）的全称是“网络爬虫排除标准”，robots.txt是搜索引擎访问网站时第一个查看的文件，当我们网站有部分内容不希望收搜索引擎抓取时，就可以通过Robots协议来告诉搜索引擎哪些页面是不能抓取的，大多用来保护网站的隐私，以及一些死链、重复页面等等。
https://www.zhihu.com (https://www.zhihu.com)
https://www.zhihu.com/robots.txt (https://www.zhihu.com/robots.txt)

1.3 不要进橘子了

破坏计算机信息系统罪
侵犯公民个人信息罪

所以说，个人信息坚决不要碰（哪怕是只有电话号码，其他什么信息的都没有的）
政府网站一定要慢，有些年久失修的网站很容易爬崩

1.4 爬虫的本质

所见即所得，不可见不可得
能爬的能爬，不能爬的不能爬；如果不能爬的能爬，那么能爬的都能爬，所以你到底爬不爬

1.5 推荐课程

1.5.1 入门

https://www.bilibili.com/video/BV1i54y1h75W (https://www.bilibili.com/video/BV1i54y1h75W)

2021年最新Python爬虫教程+实战项目案例（最新录制）

UP 路飞学城IT · 2021-3-5

辅以：《Python 3网络爬虫开发实战 》崔庆才著.pdf

1.5.2 进阶

Python 3反爬虫原理与绕过实战 by 韦世东 (z-lib.org).pdf
正则指引 by 余晟 (z-lib.org).pdf
js逆向的资源

2 开始爬了

2.1 第0步：导库

In [2]:

```
import requests
```

In [3]:

```
pip install requests
```

```
Requirement already satisfied: requests in e:\pythonenvironment\lib\site-packages (2.28.1)
Requirement already satisfied: certifi>=2017.4.17 in e:\pythonenvironment\lib\site-packages
(from requests) (2022.6.15)
Requirement already satisfied: charset-normalizer<3,>=2 in e:\pythonenvironment\lib\site-pa
ckages (from requests) (2.0.12)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in e:\pythonenvironment\lib\site-packag
es (from requests) (1.26.9)
Requirement already satisfied: idna<4,>=2.5 in e:\pythonenvironment\lib\site-packages (from
requests) (3.3)
Note: you may need to restart the kernel to use updated packages.
```

2.2 第一步：请求信息

2.2.1 请求方法

> 请求方法一般有两种，post 请求、get 请求

2.2.1.1 get 请求

https://baike.baidu.com/item/%E6%88%90%E9%83%BD%E4%BF%A1%E6%81%AF%E5%B7%A5%E7%A8%8
(https://baike.baidu.com/item/%E6%88%90%E9%83%BD%E4%BF%A1%E6%81%AF%E5%B7%A5%E7%A8%8

> 网页源代码里面看得到（Ctrl+U），每一个操作，网页的网址都会改变（豆瓣）

In [7]:

```python
resp = requests.get("https://baike.baidu.com/item/%E6%88%90%E9%83%BD%E4%BF%A1%E6%81%A
resp.encoding = "utf-8" # 更改网页的编码
resp.text
```

Out[7]:

```
'<!doctype html>\r\n<html lang="en">\r\n<head>\r\n    <meta charset="UTF-8">\r\n    <title>
百度百科——全球领先的中文百科全书</title>\r\n    <style>\r\n    p {\r\n        margin: 0;\r\n    }\r
\n    .baikeLogo {\r\n        width: 780px;\r\n        height: 50px;\r\n        margin: 150px auto 75px;\r
\n        text-indent: -9999em;\r\n        background: url(https://img.baidu.com/img/baike/logo-ba
ike.png) 50% 50% no-repeat;\r\n    }\r\n    /* S-- errorBox */\r\n        .errorBox {\r\n            width:
780px;\r\n            margin: 0 auto 65px;\r\n            text-align: center;\r\n            font-family: "Micro
soft yahei";\r\n        }\r\n        .errorBox .timeOut {\r\n            color: #666;\r\n            font-size: 16p
x;\r\n        }\r\n        .errorBox .timeOut a {\r\n            color: #136ec2;\r\n            text-decoration:n
one;\r\n        }\r\n        .errorBox .countdown {\r\n            font-weight: 700;\r\n        }\r\n    /* E-- er
rorBox */\r\n\r\n    /* S-- sorryBox */\r\n        .sorryBox {\r\n            position: relative;\r\n
margin-bottom: 10px;\r\n        }\r\n        .sorryBox .sorryTxt {\r\n            color: #559ee7;\r\n        }
\r\n        .sorryBox .sorryCont {\r\n            color: #333;\r\n            font-size: 35px;\r\n        }\r\n
.sorryBox .sorryBubble {\r\n            position: absolute;\r\n            left: 98px;\r\n            top: -35p
x;\r\n            width: 72px;\r\n            height: 37px;\r\n            background: url(/static/common/im
g/error_bubble_7da2966.jpg) no-repeat 50% 50%;\r\n        }\r\n    /* E-- sorryBox */\r\n\r\n    /*
S-- footer */\r\n    .ft {\r\n            width: 780px;\r\n            margin: 0 auto 65px;\r\n            paddi
ng-top: 20px;\r\n            padding-bottom: 20px;\r\n            line-height: 20px;\r\n            color: #66
6;\r\n            font-size: 12px;\r\n            text-align: center;\r\n            background-color: #f8f8f8;\r
\n        }\r\n        .ft a{\r\n            color: #2d64b3;\r\n            text-decoration: none;\r\n        }\r\n
.ft a:hover {\r\n            text-decoration: underline;\r\n        }\r\n        .feedBackWays .ul {\r\n
margin: 0;\r\n            padding: 0;\r\n        }\r\n        .feedBackWays .li {\r\n            list-style: none;\r
\n        }\r\n        .ftCont {\r\n            margin-top: 20px;\r\n            color: #2d64b3;\r\n        }\r\n    /*
E-- footer */\r\n    </style>\r\n</head>\r\n<body>\r\n    <div id="bd">\r\n        <h1 class="baike
Logo">\r\n            百度百科错误页\r\n        </h1>\r\n        <div class="errorBox">\r\n            <!-- 主
体 -->\r\n            <div class="sorryBox">\r\n                <div class="sorryBubble"></div>\r\n
<p class="sorryCont"><span class="sorryTxt">抱歉</span>，您所访问的页面不存在...</p>\r\n
</div>\r\n            <div class="timeOut">\r\n                <p><span class="countdown" id="countdo
wn">3</span>秒后自动跳转到<a href="http://baike.baidu.com/">百科首页</a></p>\r\n
</div>\r\n            <!-- /主体 -->\r\n        </div>\r\n    </div>\r\n    <div id="ft">\r\n        <div class
="ft">\r\n            <div class="feedBackWays">\r\n                <ul class="ul">\r\n                    <li class
="li">如果想提出功能问题或意见建议,请到<a href="http://baike.baidu.com/feedback" target
="_blank">意见反馈</a>;</li>\r\n                    <li class="li">如果您要举报侵权或违法信息,请到<a
href="http://help.baidu.com/newadd?prod_id=10&category=1" target="_blank">投诉中心</a
>;</li>\r\n                    <li class="li">其他问题请访问<a href="http://tieba.baidu.com/f?kw=%B
0%D9%B6%C8%B0%D9%BF%C6" target="_blank">百度百科吧</a>。</li>\r\n                </ul>\r
\n            </div>\r\n            <div class="ftCont">\r\n                &copy;<span id="copyYear"></span>
Baidu <a href="http://www.baidu.com/duty/" target="_blank">使用百度前必读</a> | <a href
="http://help.baidu.com/question?prod_en=baike&class=89&id=1637" target="_blank">百科
协议</a> | <a href="http://baike.baidu.com/hezuo/" target="_blank">百度百科合作平台</a>\r
\n            </div>\r\n        </div>\r\n    </div>\r\n    <script type="text/javascript">\r\n    window.o
nload = function(){\r\n        var time = 3,\r\n            year = new Date().getFullYear();\r\n\r\n
document.getElementById("copyYear").innerHTML = year;\r\n        setInterval(function(){\r\n
if (time == 0){\r\n                window.location = "http://baike.baidu.com/";\r\n                return;\r\n
}\r\n            document.getElementById("countdown").innerHTML = time;\r\n            time--;\r\n
}, 1000);\r\n    }\r\n    </script>\r\n</body>\r\n</html>'
```

2.2.1.2 post 请求

http://www.gov.cn/zhuanti/2021yqfkgdzc/index.htm#/ (http://www.gov.cn/zhuanti/2021yqfkgdzc/index.htm#/)

> 浏览器地址栏的url不会改变，需要通过抓包获取真实的url

In [19]:

```python
import time
time.time()
```

Out[19]:

```
1659277521.9730966
```

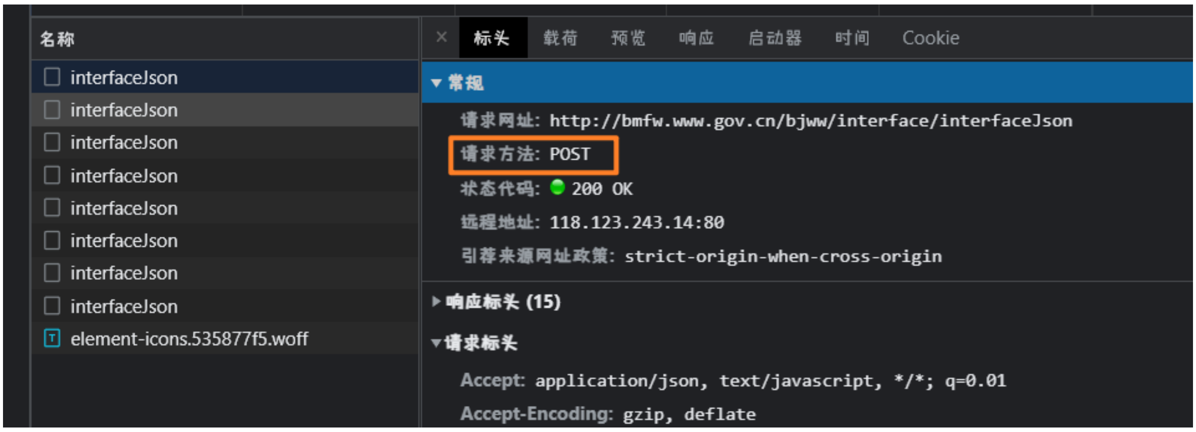
In [65]:

```
# 这个网站有 sha265 加密

import hashlib
import os
import time
import pandas as pd
import requests
import fake_useragent
import tqdm
import random
import re
import datetime


def get_signatureHeader(timestamp, token="23y0ufFl5YxIyGrI8hWRUZmKkvtSjLQA"):
    timestamp = timestamp
    nonce = "123456789abcdefg"
    data = timestamp + token + nonce + timestamp
    data_sha = hashlib.sha256(data.encode("utf-8")).hexdigest()
    return data_sha


def get_x_wif_signature(timestamp):
    data = timestamp + "fTN2pfuisxTavbTuYVSsNJHetwq5bJvCQkjjtiLM2dCratiA" + timestamp
    data_sha = hashlib.sha256(data.encode("utf-8")).hexdigest()
    return data_sha


def get_data(city_code):
    ti = str(time.time()).split(".")[0]
    js = {
        "appId": "NcApplication",
        "code": city_code,
        "key": "6C3C60DC1BF54982A54D5A8CB4D1817D",
        "nonceHeader": "123456789abcdefg",
        "paasHeader": "zdww",
        "signatureHeader": get_signatureHeader(timestamp=ti).upper(),
        "timestampHeader": ti
    }

    ua = fake_useragent.UserAgent()
    headers = {
        "Accept": "application/json, text/javascript, */*; q=0.01",
        "Accept-Encoding": "gzip, deflate",
        "Accept-Language": "en-GB,en;q=0.9,zh-CN;q=0.8,zh;q=0.7",
        "Cache-Control": "no-cache",
        "Connection": "keep-alive",
        "Content-Length": "251",
        "Content-Type": "application/json; charset=UTF-8",
        "Host": "bmfw.www.gov.cn",
        "Origin": "http://www.gov.cn",
        "Pragma": "no-cache",
        "Referer": "http://www.gov.cn/",
        "User-Agent": ua.random,
        "x-wif-nonce": "QkjjtiLM2dCratiA",
        "x-wif-paasid": "smt-application",
        "x-wif-signature": get_x_wif_signature(timestamp=ti).upper(),
        "x-wif-timestamp": ti
    }
```

```python

    url = "http://bmfw.www.gov.cn/bjww/interface/interfaceJson"
    resp_json = requests.post(url=url, headers=headers, json=js).json()
    return resp_json


get_data("321300")
```

Out[65]:

```
{'data': {'levelTag': '6',
 'list': [{'code': '321300',
   'province': '江苏省',
   'city': '宿迁市',
   'county': '',
   'provider': '宿迁市新冠肺炎疫情防控工作领导小组办公室',
   'leave_policy': '非必要不前往中高风险地区和有本土聚集性疫情所在的县（市、区、旗）。如
果确实有需要要前往这些地区，要在抵宿前向所在社区（村、单位、酒店）报告，或通过“宿康
宝”小程序进行网上申报，并配合做好健康管理措施。',
   'come_policy': '1.密切关注疫情动态，凡是有中高风险地区、本土聚集性疫情所在设区市的来
（返）宿人员，提前2天通过“宿康宝”小程序自主申报或向目的地单位、社区（村）或酒店报
备，抵宿后积极配合落实相关健康管理措施。\n\n2.对入境人员实行“7天集中隔离医学观察+3天
居家健康监测”管理措施。居家健康监测期间不外出，如就医等特殊情况必需外出时做好个人防
护，尽量避免乘坐公共交通工具。\n\n3.近7天内有高风险地区旅居史人员，要主动向所在社区
（村、酒店、单位等）报备，积极配合做好“7天集中隔离医学观察”，在集中隔离第1、2、3、
5、7天各开展一次核酸检测。\n\n4.近7天内有中风险地区旅居史人员，配合做好“7天居家隔离
医学观察”，居家隔离医学观察第1、4、7天各开展一次核酸检测；如不具备居家隔离医学观察条
件，采取集中隔离医学观察。\n\n5.近7天内有中高风险区所在县（区）低风险区旅居史的人员，
需持48小时核酸检测阴性证明。返宿后3天内应完成两次核酸检测（间隔24小时），并做好健康
监测。其他低风险区来宿（返）宿人员持绿码可自由流动。\n\n6.对于有疫情地区，高校内如没
有疫情、学校实施7天以上封闭管理、持48小时内核酸阴性证明和所在高校开具的相关证明，不
再进行集中隔离，到家后实施7天健康监测。\n\n7.有关防控措施将根据疫情形势动态调整，请广
大市民继续履行个人防控责任，做好戴口罩、勤洗手、多通风、勤消毒、保持社交距离等防护措
施，主动配合落实各项疫情防控措施。',
   'aviation': '',
   'railway': '1.对进站旅客检码测温，检查48小时核酸检测阴性证明；对健康码为红黄码、体温异
常的旅客，交由卫健部门按规定处置；对不能提供48小时核酸检测阴性证明的旅客，一律劝返。
\n\n2.对出站旅客检码测温，检查48小时核酸检测阴性证明；对健康码为红黄码、体温异常的旅
客，交由属地卫健部门按规定处置；对健康码为绿码，但不能提供48小时核酸检测阴性证明的旅
客，由防疫人员引导至高铁站便民核酸采集点进行采样。对出站旅客动态语音播报，告知落实“3
+11”疫情防控要求。',
   'highway': '1.客运方面：客运站所有进出旅客须测温验码（苏康码、行程码），对有中高风险
地区所在设区市的低风险地区旅居史进入我市的人员额外查验48小时内核酸检测阴性证明。有本
土病例报告的城市来（返）宿人员，应在客运站出口主动申报，同时在抵宿后应12小时之内向所
在社区（村）和单位报备，须持48小时内核酸检测阴性证明方可有序流动。\n\n2.道路卡口方
面：各查验点对来自或途径连云港、苏州、上海、青岛等国内重点中高风险地区车辆和人员（查
看行程码）做到逢车必查、逢人必验；在查验过程中对健康码为绿码且持有48小时内核酸检测阴
性证明的，直接予以放行；未持有48小时内核酸检测阴性证明的，由查验点卫健工作人员提醒其
向目的地社区报备，按要求做核酸检测，登记车辆人员信息后予以放行；对健康码为黄码的宿迁
籍人员，由查验点卫健工作人员联系目的地疫情防控指挥部，落实集中隔离措施，外地人员实行
劝返；对健康码为红码的，禁止驾乘人员离开车辆，由公安、卫健部门将车辆人员引导至指定场
所，由当地卫健部门按规定实施集中隔离医学观察。对黄码和红码人员要做好被查车辆和人员信
息登记工作；对体温高于37.3度的人员，由查验点卫健工作人员联系120专用救护车送至医疗机
构发热门诊就诊。',
   'waterway': '进入我市辖区从事水路交通运输活动人员须测温、验码（健康码、行程码）、戴
口罩，从中高风险地区所在设区市进入我市的人员须额外提供48小时内核酸检测阴性证明。',
   'create_time': '2022年7月29日12时',
   'sendtime': 1656676800}],
 'leave_list': [],
 'come_list': [],
 'traffic_list': []},
 'code': 0,
 'msg': '查询成功'}
```

> 浏览器地址栏的 url 是 get 请求，但通过 F12 抓包的，请求方法并不一定绝对是 post，需要具体去看
>
> 
>
>
> interfaceJson
> interfaceJson
> interfaceJson
> interfaceJson
> interfaceJson
> interfaceJson
> interfaceJson
> interfaceJson
> element-icons.535877f5.woff
>
> ▼常规
> 请求网址: http://bmfw.www.gov.cn/bjww/interface/interfaceJson
> 请求方法: POST
> 状态代码: ● 200 OK
> 远程地址: 118.123.243.14:80
> 引荐来源网址政策: strict-origin-when-cross-origin
> ▶响应标头 (15)
> ▼请求标头
> Accept: application/json, text/javascript, */*; q=0.01
> Accept-Encoding: gzip, deflate

2.2.2 请求头 headers [...]

2.2.3 其他参数

2.2.3.1 params 参数

https://cn.bing.com/search?q=python (https://cn.bing.com/search?q=python)

In [21]:

```python
url = "https://cn.bing.com/search?q=python"
resp = requests.get(url=url, headers=headers)
# resp.text
resp.url
```

Out[21]:

```
'https://cn.bing.com/search?q=python'
```

In [23]:

```python
url = "https://cn.bing.com/search"
params = {
    "q": "python"
}
resp = requests.get(url=url, headers=headers, params=params)
# resp.text
resp.url
```

Out[23]:

```
'https://cn.bing.com/search?q=python'
```

2.2.3.2 post 请求的 json 参数

> requests.post(url, headers, pararms, json)

我们都知道，post 请求通常是需要提交数据给服务器的，这个时候，我们就可以这个样子（那个防疫政策的票子）

2.3 第二步：提取数据

2.3.1 resp.text

提取数据有很多库，比如说
bs4（库名叫 beautifulsoup4，熟称 美丽汤）
etree
re（著名的正则表达式）
这个地方就不讲了，建议去开头说的B站康康，正则表达式是最强的，正则yyds

2.3.2 resp.json 如果有

In [25]:

```
resp_json = '{"errorCode":0,"errorMsg":"","hasMore":true,"offset":"1651707295.141546","pageItems":[
print(resp_json)
print(type(resp_json))
```

```
{"errorCode":0,"errorMsg":"","hasMore":true,"offset":"1651707295.141546","pageItems":[{"i
nfo":{"itemId":"100_36326333","type":100,"jumpUrl":"https://news.10jqka.com.cn/tapp/notic
e.html#seq=36326333","webrsid":"ann_36326333"},"combination":[{"topicLine":{"topic":{"ur
l":"https://basic.10jqka.com.cn/basicph/event.html?code=000004","name":"公司公告","topic
Webrsid":"ann_topic","iconInfo":{"urlForDay":"https://u.thsi.cn/imgsrc/flashcms/322848080_
0be2d48c31f4952176d8410d53332b09.png","urlForNight":"https://u.thsi.cn/imgsrc/flashcm
s/322848080_b05c3c8cda16771379d300065b9ce029.png"}}}},{"title":{"content":"ST国华：股
票交易异常波动公告"}},{"bottomBar":{"recommend":"","recommendType":0,"source":"深交所
股票","time":1652803396000,"clicks":-1}}]},{"info":{"itemId":"100_36370919","type":100,"jum
pUrl":"https://news.10jqka.com.cn/tapp/notice.html#seq=36370919","webrsid":"ann_363709
19"},"combination":[{"topicLine":{"topic":{"url":"https://basic.10jqka.com.cn/basicph/event.ht
ml?code=000004","name":"公司公告","topicWebrsid":"ann_topic","iconInfo":{"urlForDay":"htt
ps://u.thsi.cn/imgsrc/flashcms/322848080_0be2d48c31f4952176d8410d53332b09.png","urlF
orNight":"https://u.thsi.cn/imgsrc/flashcms/322848080_b05c3c8cda16771379d300065b9ce0
29.png"}}}},{"title":{"content":"国华网安:关于对深圳国华网安科技股份有限公司2021年年报问询
函的回复"}},{"bottomBar":{"recommend":"","recommendType":0,"source":"深交所","time":165
2803200000,"clicks":-1}}]},{"info":{"itemId":"1_639165834","type":1,"jumpUrl":"https://news.1
0jqka.com.cn/m639165834/","webrsid":"seq_639165834"},"combination":[{"title":{"conten
t":"【龙虎榜】 ST国华 05月17日成交明细"}},{"bottomBar":{"recommend":"","recommendTyp
e":0,"source":"同花顺资讯中心","time":1652775783000,"clicks":1360}}]},{"info":{"itemId":"1_6
39135965","type":1,"jumpUrl":"https://news.10jqka.com.cn/m639135965/","webrsid":"seq_6
39135965"},"combination":[{"title":{"content":"ST国华05月16日主力资金大幅流入"}},{"bottom
Bar":{"recommend":"","recommendType":0,"source":"同花顺AI资讯社","time":165268754800
0,"clicks":1583}}]}]}
```

```
<class 'str'>
```

In [26]:

```
import json

resp_json = json.loads(resp_json) # 只有对字符串需要这样转换为 json 格式，如果是 resp.json，那么拿到
print(resp_json)
print(type(resp_json))
```

```
{'errorCode': 0, 'errorMsg': '', 'hasMore': True, 'offset': '1651707295.141546', 'pageItems': [{'in
fo': {'itemId': '100_36326333', 'type': 100, 'jumpUrl': 'https://news.10jqka.com.cn/tapp/notice.
html#seq=36326333', 'webrsid': 'ann_36326333'}, 'combination': [{'topicLine': {'topic': {'url': 'h
ttps://basic.10jqka.com.cn/basicph/event.html?code=000004', 'name': '公司公告', 'topicWebrs
id': 'ann_topic', 'iconInfo': {'urlForDay': 'https://u.thsi.cn/imgsrc/flashcms/322848080_0be2d4
8c31f4952176d8410d53332b09.png', 'urlForNight': 'https://u.thsi.cn/imgsrc/flashcms/32284
8080_b05c3c8cda16771379d300065b9ce029.png'}}}}, {'title': {'content': 'ST国华：股票交易异
常波动公告'}}, {'bottomBar': {'recommend': '', 'recommendType': 0, 'source': '深交所股票', 'tim
e': 1652803396000, 'clicks': -1}}]}, {'info': {'itemId': '100_36370919', 'type': 100, 'jumpUrl': 'http
s://news.10jqka.com.cn/tapp/notice.html#seq=36370919', 'webrsid': 'ann_36370919'}, 'combi
nation': [{'topicLine': {'topic': {'url': 'https://basic.10jqka.com.cn/basicph/event.html?code=000
004', 'name': '公司公告', 'topicWebrsid': 'ann_topic', 'iconInfo': {'urlForDay': 'https://u.thsi.cn/i
mgsrc/flashcms/322848080_0be2d48c31f4952176d8410d53332b09.png', 'urlForNight': 'http
s://u.thsi.cn/imgsrc/flashcms/322848080_b05c3c8cda16771379d300065b9ce029.png'}}}}, {'tit
le': {'content': '国华网安:关于对深圳国华网安科技股份有限公司2021年年报问询函的回复'}}, {'bo
ttomBar': {'recommend': '', 'recommendType': 0, 'source': '深交所', 'time': 1652803200000, 'clic
ks': -1}}]}, {'info': {'itemId': '1_639165834', 'type': 1, 'jumpUrl': 'https://news.10jqka.com.cn/m6
39165834/', 'webrsid': 'seq_639165834'}, 'combination': [{'title': {'content': '【龙虎榜】 ST国华
05月17日成交明细'}}, {'bottomBar': {'recommend': '', 'recommendType': 0, 'source': '同花顺资
讯中心', 'time': 1652775783000, 'clicks': 1360}}]}, {'info': {'itemId': '1_639135965', 'type': 1, 'ju
mpUrl': 'https://news.10jqka.com.cn/m639135965/', 'webrsid': 'seq_639135965'}, 'combinatio
n': [{'title': {'content': 'ST国华05月16日主力资金大幅流入'}}, {'bottomBar': {'recommend': '', 'rec
ommendType': 0, 'source': '同花顺AI资讯社', 'time': 1652687548000, 'clicks': 1583}}]}]}
<class 'dict'>
```

In [36]:

```
li = []
for i in resp_json["pageItems"]:
    li.append(i["info"]["jumpUrl"])
li
```

Out[36]:

```
['https://news.10jqka.com.cn/tapp/notice.html#seq=36326333',
 'https://news.10jqka.com.cn/tapp/notice.html#seq=36370919',
 'https://news.10jqka.com.cn/m639165834/',
 'https://news.10jqka.com.cn/m639135965/']
```

2.3.3 resp.content

> 是在爬图片、视频、压缩包等文件时用
> 后面会举例子的

2.4 第三步：保存数据

2.4.1 文本数据

推荐使用 csv 储存

> 逗号分隔值（Comma-Separated Values，CSV，有时也称为字符分隔值，因为分隔字符也可以不是逗号），其文件以纯文本形式存储表格数据（数字和文本）。纯文本意味着该文件是一个字符序列，不含必须像二进制数字那样被解读的数据。CSV文件由任意数目的记录组成，记录间以某种换行符分隔；每条记录由字段组成，字段间的分隔符是其它字符或字符串，最常见的是逗号或制表符。

In [37]:

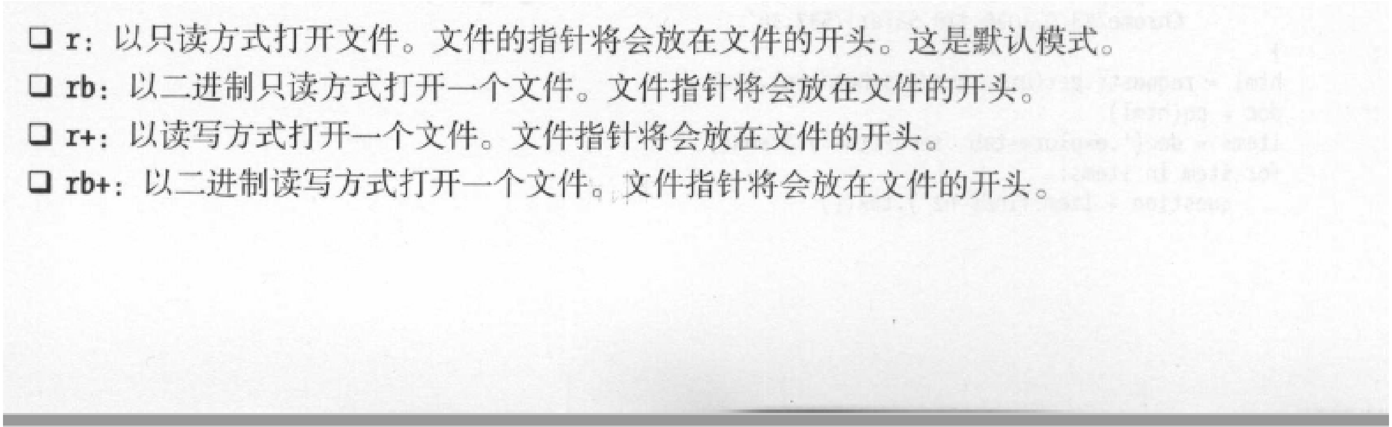
```
# 导库
import csv
```

2.4.1.1 写入数据的第一种方法 open

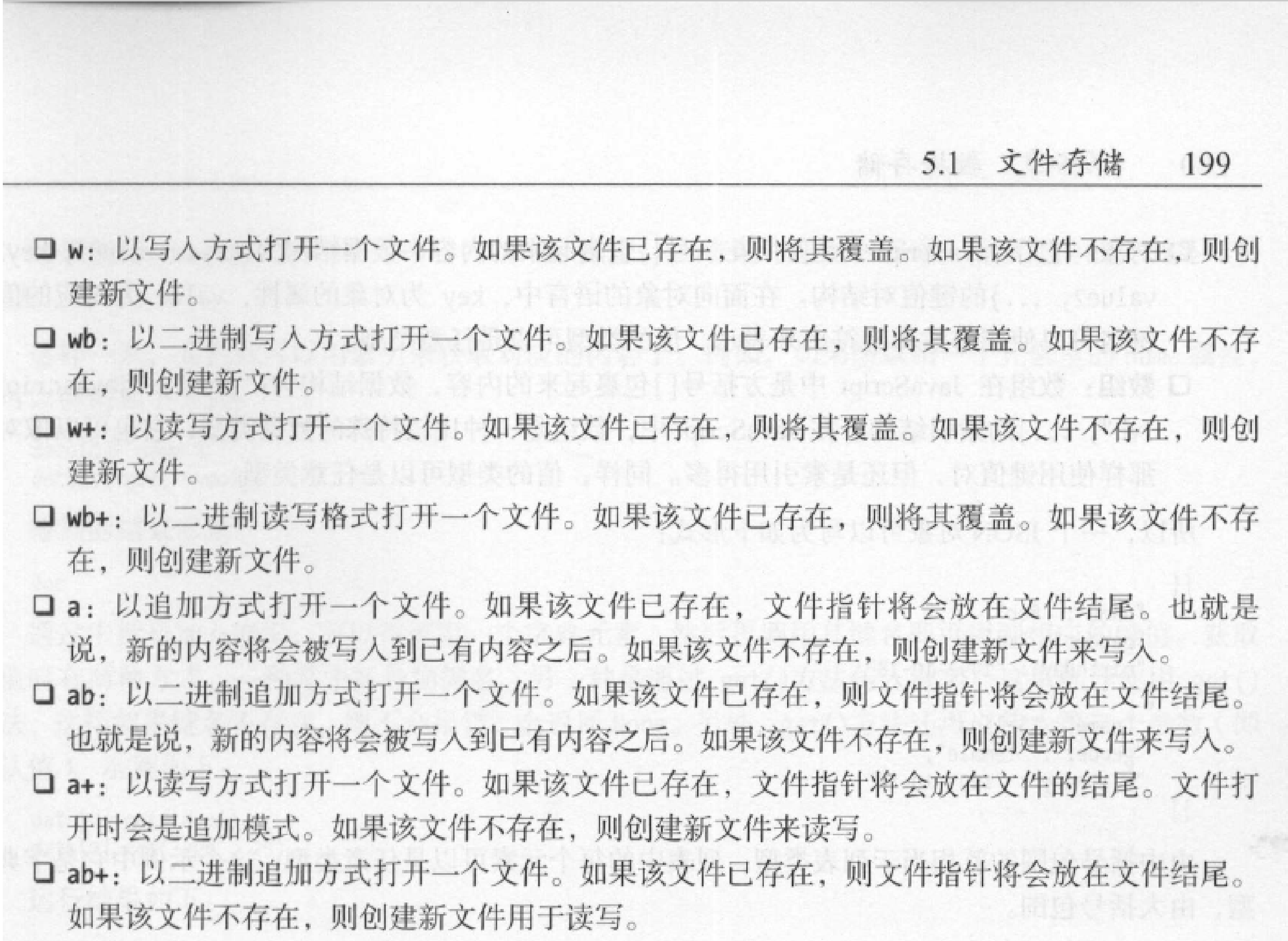
In [5]:

```
file = open(r"E:\JupyterNotebook\接单单\武侯区\data.csv", encoding="utf-8")
data = file.read()
# data
```

> open() 函数的参数
> open(file, mode='r', buffering=-1, encoding=None, errors=None, newline=None, closefd=True, opener=None)



- ❑ r：以只读方式打开文件。文件的指针将会放在文件的开头。这是默认模式。
- ❑ rb：以二进制只读方式打开一个文件。文件指针将会放在文件的开头。
- ❑ r+：以读写方式打开一个文件。文件指针将会放在文件的开头。
- ❑ rb+：以二进制读写方式打开一个文件。文件指针将会放在文件的开头。



5.1 文件存储 199

- ❑ w：以写入方式打开一个文件。如果该文件已存在，则将其覆盖。如果该文件不存在，则创建新文件。
- ❑ wb：以二进制写入方式打开一个文件。如果该文件已存在，则将其覆盖。如果该文件不存在，则创建新文件。
- ❑ w+：以读写方式打开一个文件。如果该文件已存在，则将其覆盖。如果该文件不存在，则创建新文件。
- ❑ wb+：以二进制读写格式打开一个文件。如果该文件已存在，则将其覆盖。如果该文件不存在，则创建新文件。
- ❑ a：以追加方式打开一个文件。如果该文件已存在，文件指针将会放在文件结尾。也就是说，新的内容将会被写入到已有内容之后。如果该文件不存在，则创建新文件来写入。
- ❑ ab：以二进制追加方式打开一个文件。如果该文件已存在，则文件指针将会放在文件结尾。也就是说，新的内容将会被写入到已有内容之后。如果该文件不存在，则创建新文件来写入。
- ❑ a+：以读写方式打开一个文件。如果该文件已存在，文件指针将会放在文件的结尾。文件打开时会是追加模式。如果该文件不存在，则创建新文件来读写。
- ❑ ab+：以二进制追加方式打开一个文件。如果该文件已存在，则文件指针将会放在文件结尾。如果该文件不存在，则创建新文件用于读写。

其实就关注：
r：读
w：写
a：追加
rb、wb、ab：二进制（在图片、视频、文件才会用到）

In [41]:

```
data_list = [1, 2, 3, 4]
data_lists = [
    [7, 8, 9, 10],
    [11, 18, 0, 0]
        ]
```

In [43]:

```
f = open("data1.csv", encoding="utf-8", mode="w", newline="")
dataWriter = csv.writer(f)
dataWriter.writerow(data_list)
dataWriter.writerows(data_lists)
f.close()
```

2.4.1.2 写入数据的第二种方法 with open

In [46]:

```
with open("data2.csv", encoding="utf-8", mode="w", newline="") as f:
    dataWriter1 = csv.writer(f)
    dataWriter1.writerows(data_lists)
    dataWriter1.writerow(data_list)
```

2.4.1.3 视频或者图片或者压缩包

http://www.rsscc.cn/index.html#first (http://www.rsscc.cn/index.html#first)

In [4]:

```
resp = requests.get("https://cdn.133.cn/ticket/web/official-website/video/index-header.mp4")
resp_content = resp.content  # 二进制数据
# resp_content
```

In [48]:

```
with open("装逼的视频.mp4", mode="wb") as f:
    f.write(resp_content)
```

3 反爬之一：需要登录

3.1 方法一：session 使用会话

In [4]:

```
session = requests.Session()

resp = session.post(url="https://www.bilibli.com/login", json={"zhanghao":"00", "pwd":"00"})

resp = session.get()
resp.json

session.get()
```

3.2 方法二：携带表明你身份的东西：例如cookie

下面这个案例就是介个方法

4 手搓一个爬虫

4.1 第一只爬虫

In [3]:

```
import pandas as pd
import tqdm
import time
from concurrent import futures
import random


def get_data(air, da_start, da_end):
    params = {
        "airport": air,
        "route_type": "all",
        "start_time": da_start,
        "end_time": da_end
    }
    headers = {
        "Authorization": 填你自己的,
        "Origin": "https://dast.133.cn",
        "Referer": "https://dast.133.cn/",
        "user-agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
        "client-id": 填你自己的,
        "Content-Type": "application/x-www-form-urlencoded",
        "sec-ch-ua-mobile": "?0",
        "sec-ch-ua-platform": "Windows",
        "Sec-Fetch-Dest": "empty",
        "Sec-Fetch-Mode": "cors",
        "Sec-Fetch-Site": "same-site"
    }
    url = " https://data-api.133.cn/api/v1/airport/statistics"
    resp = requests.get(url=url, params=params, headers=headers)
    resp_json = resp.json()

    arr_plan = resp_json["data"]["fluctuation"]["arr_plan"]  # 计划进港
    arr_real = resp_json["data"]["fluctuation"]["arr_real"]  # 实际进港
    dep_plan = resp_json["data"]["fluctuation"]["dep_plan"]  # 计划出港
    dep_real = resp_json["data"]["fluctuation"]["dep_real"]  # 实际出港

    data_list = arr_plan + arr_real + dep_plan + dep_real
    return data_list


def main(da):
    data = get_data("PEK", da_start=da, da_end=da)
    dataWriter.writerow([da] + data)


if __name__ == '__main__':
    with open("CTU_data_twoYears.csv", mode="a", encoding="utf-8", newline="") as f:
        dataWriter = csv.writer(f)

        for date in tqdm.tqdm(pd.date_range("2020-06-25", "2022-07-27")):
            date_str = str(date.date())
            main(date_str)
            time.sleep(random.random() + 2)
```

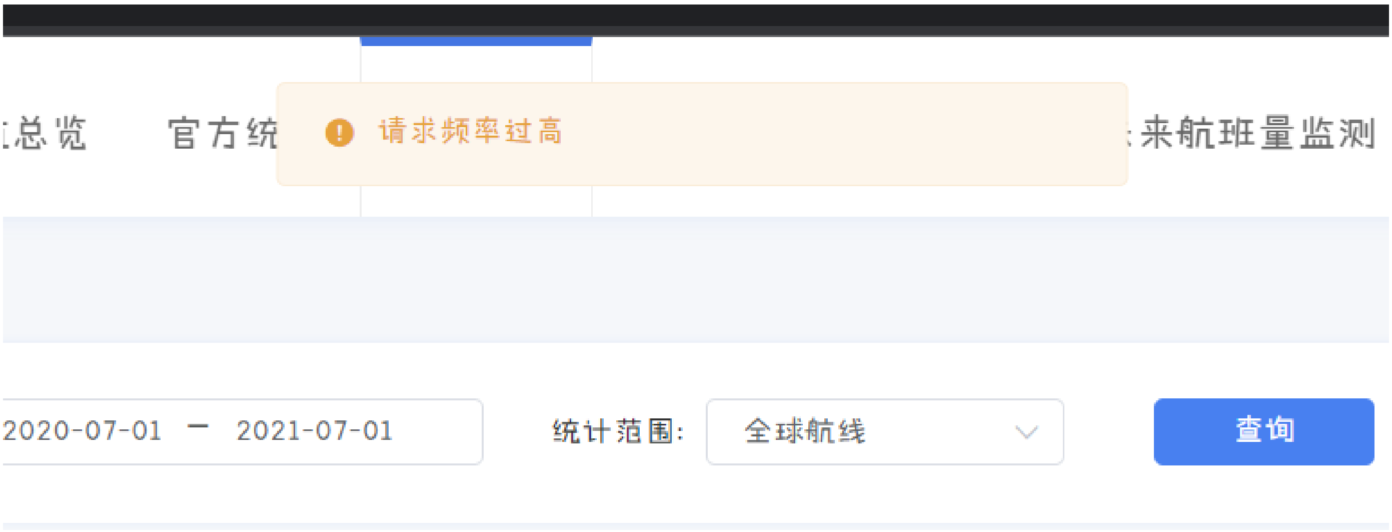
```
  0%|▌                                                  | 2/763 [00:08<53:58,  4.2
6s/it]

---------------------------------------------------------------------------
```

```
KeyboardInterrupt                         Traceback (most recent call last)
Input In [3], in <cell line: 46>()
     51 date_str = str(date.date())
     52 main(date_str)
---> 53 time.sleep(random.random() + 2)

KeyboardInterrupt: 
```

4.2 飞起来爬

In [ ]:

```python
import pandas as pd
import tqdm
import time
from concurrent import futures
import random


def get_data(air, da_start, da_end):
    params = {
        "airport": air,
        "route_type": "all",
        "start_time": da_start,
        "end_time": da_end
    }
    headers = {
        "Authorization": 填你自己的,
        "Origin": "https://dast.133.cn",
        "Referer": "https://dast.133.cn/",
        "user-agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
        "client-id": 填你自己的,
        "Content-Type": "application/x-www-form-urlencoded",
        "sec-ch-ua-mobile": "?0",
        "sec-ch-ua-platform": "Windows",
        "Sec-Fetch-Dest": "empty",
        "Sec-Fetch-Mode": "cors",
        "Sec-Fetch-Site": "same-site"
    }
    url = " https://data-api.133.cn/api/v1/airport/statistics"
    resp = requests.get(url=url, params=params, headers=headers)
    resp_json = resp.json()

    arr_plan = resp_json["data"]["fluctuation"]["arr_plan"]  # 计划进港
    arr_real = resp_json["data"]["fluctuation"]["arr_real"]  # 实际进港
    dep_plan = resp_json["data"]["fluctuation"]["dep_plan"]  # 计划出港
    dep_real = resp_json["data"]["fluctuation"]["dep_real"]  # 实际出港

    data_list = arr_plan + arr_real + dep_plan + dep_real
    return data_list


def main(da):
    data = get_data("PEK", da_start=da, da_end=da)
    dataWriter.writerow([da] + data)


if __name__ == '__main__':
    with open("CTU_data_twoYears.csv", mode="a", encoding="utf-8", newline="") as f:
        dataWriter = csv.writer(f)

        tasks = []
        with futures.ThreadPoolExecutor(20) as t:
            for date in pd.date_range("2020-01-01", "2022-07-27"):
                tasks.append(t.submit(main, date))
            print("爬")
            for task in tqdm.tqdm(futures.as_completed(tasks), total=len(tasks)):
                task.result()
```



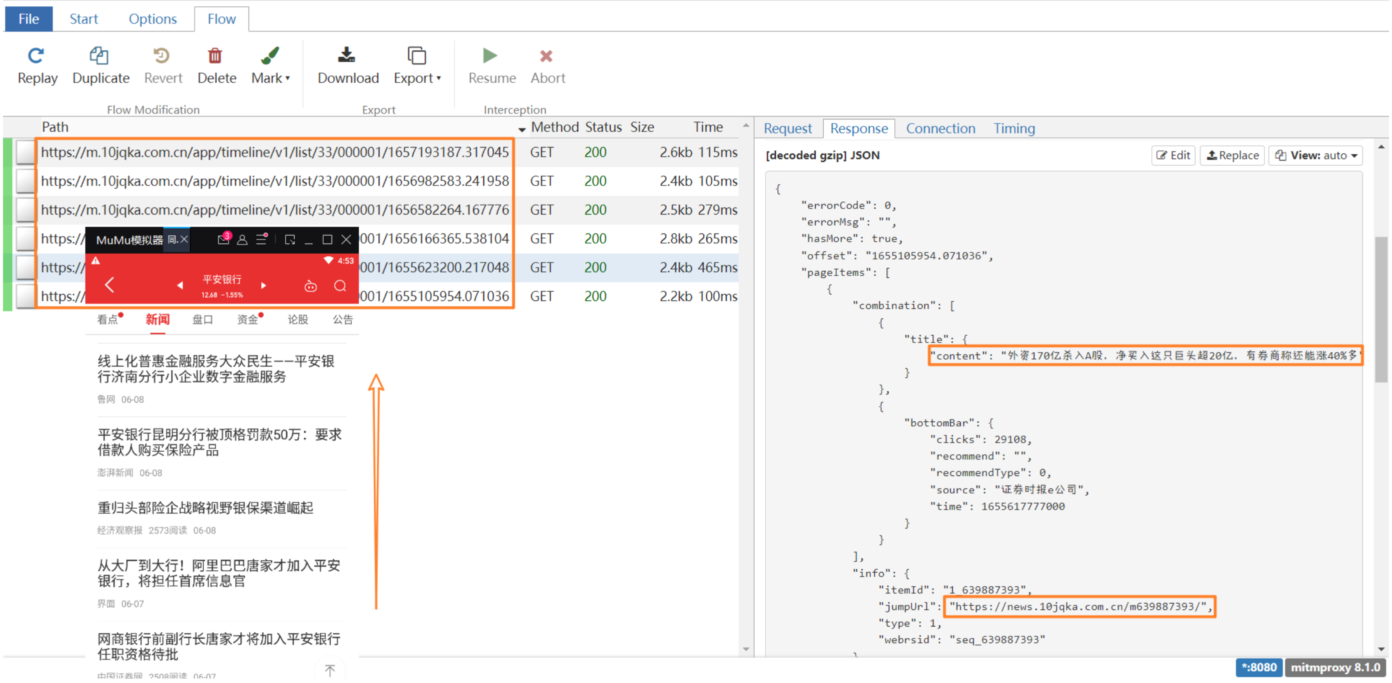
2020-07-01 — 2021-07-01

统计范围： 全球航线

反反爬方法一： 买账号
反反爬方法二： 买代理 ip（hui产）
反反爬方法三： selenium解君愁

5 app 爬虫

> 模拟器： 一定要 mumu 模拟器的 安卓6，如果安装不起64位的可以安装32位
> 抓包软件： fiddler charles mitm



Path	Method	Status	Size	Time
https://m.10jqka.com.cn/app/timeline/v1/list/33/000001/1657193187.317045	GET	200	2.6kb	115ms
https://m.10jqka.com.cn/app/timeline/v1/list/33/000001/1656982583.241958	GET	200	2.4kb	105ms
https://m.10jqka.com.cn/app/timeline/v1/list/33/000001/1656582264.167776	GET	200	2.5kb	279ms
https:/// … 001/1656166365.538104	GET	200	2.8kb	265ms
https:/// … 001/1655623200.217048	GET	200	2.4kb	465ms
https:/// … 001/1655105954.071036	GET	200	2.2kb	100ms

线上化普惠金融服务大众民生——平安银行济南分行小企业数字金融服务
鲁网 06-08

平安银行昆明分行被顶格罚款50万：要求借款人购买保险产品
澎湃新闻 06-08

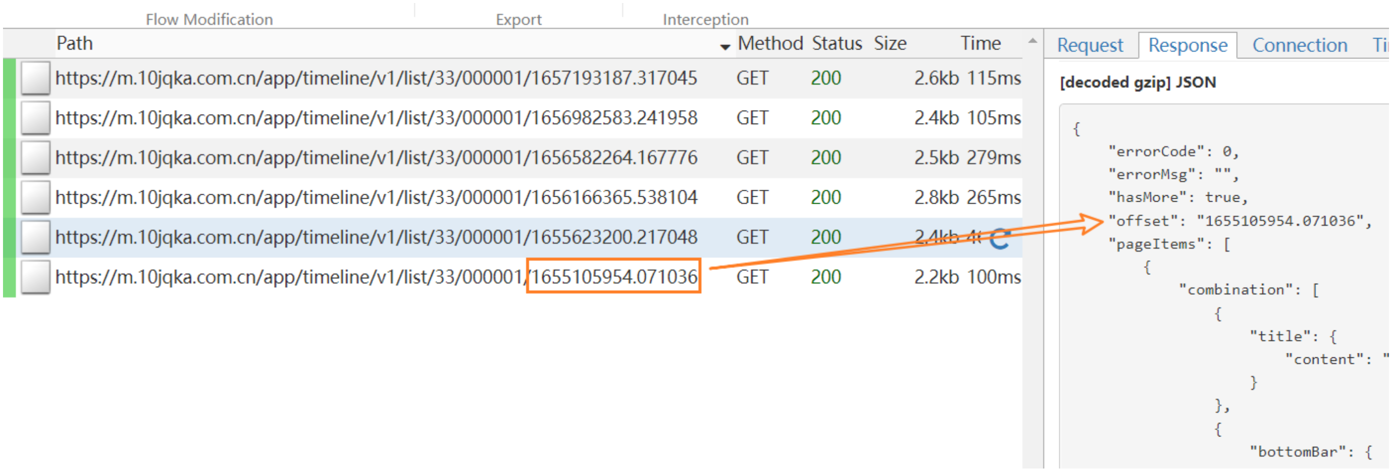
重归头部险企战略视野银保渠道崛起
经济观察报 2573阅读 06-08

从大厂到大行！阿里巴巴唐家才加入平安银行，将担任首席信息官
界面 06-07

网商银行前副行长唐家才将加入平安银行任职资格待批
中国证券网 2508阅读 06-07

[decoded gzip] JSON

```
{
    "errorCode": 0,
    "errorMsg": "",
    "hasMore": true,
    "offset": "1655105954.071036",
    "pageItems": [
        {
            "combination": [
                {
                    "title": {
                        "content": "外资170亿杀入A股，净买入这只巨头超20亿，有券商称还能涨40%多
                    }
                },
                {
                    "bottomBar": {
                        "clicks": 29108,
                        "recommend": "",
                        "recommendType": 0,
                        "source": "证券时报e公司",
                        "time": 1655617777000
                    }
                }
            ],
            "info": {
                "itemId": "1_639887393",
                "jumpUrl": "https://news.10jqka.com.cn/m639887393/",
                "type": 1,
                "webrsid": "seq_639887393"
```



Path	Method	Status	Size	Time
https://m.10jqka.com.cn/app/timeline/v1/list/33/000001/1657193187.317045	GET	200	2.6kb	115ms
https://m.10jqka.com.cn/app/timeline/v1/list/33/000001/1656982583.241958	GET	200	2.4kb	105ms
https://m.10jqka.com.cn/app/timeline/v1/list/33/000001/1656582264.167776	GET	200	2.5kb	279ms
https://m.10jqka.com.cn/app/timeline/v1/list/33/000001/1656166365.538104	GET	200	2.8kb	265ms
https://m.10jqka.com.cn/app/timeline/v1/list/33/000001/1655623200.217048	GET	200	2.4kb	4
https://m.10jqka.com.cn/app/timeline/v1/list/33/000001/1655105954.071036	GET	200	2.2kb	100ms

[decoded gzip] JSON

```
{
    "errorCode": 0,
    "errorMsg": "",
    "hasMore": true,
    "offset": "1655105954.071036",
    "pageItems": [
        {
            "combination": [
                {
                    "title": {
                        "content": "
                    }
                },
                {
                    "bottomBar": {
```

In [2]:

```
# 直接用递归算法好像要画好久时间跳出递归, 所以爬完一个要 等很久很久, 所以才用的 while True
# 这个代码写的很丑, 呜呜呜

import requests
import csv
import os
import re
import pandas as pd
import numpy as np
import time
import fake_useragent


# def save_data(one_data, code, name):
#     # 存文件
#     road = re.sub(r'[\\:*?"<>|]', " ", r"./同花顺APP新闻数据/%s-%s" % (code, name))

#     if not os.path.exists(road):
#         os.mkdir(road)
#     file = open(r"%s/%s.txt" % (road, str(time.time())), mode="w", encoding="utf-8")
#     file.write(one_data)

#     # 提取数据
#     data_list = re.findall(
#         r'"info":.*?"jumpUrl":"(.*?)",".*?"title":\{"content":"(.*?)"}},.*?"recommend":"(.*?)",.*?"source":"(.*?)","ti
#         one_data,
#         re.S)
#     data_list_save = [[code, name] + list(i) for i in data_list]
#     dataWriter.writerows(data_list_save)


def get_data_url(code, offset, name):
    url_api = f"https://m.10jqka.com.cn/app/timeline/v1/list/33/{code}/{offset}"
    resp = requests.get(url_api, headers=headers)
    resp_text = resp.text
    print(resp_text)
    offset_list = re.findall(r'"offset":"(\d+\.\d+)"', resp_text, re.S)
    try:
        hasMore = re.search(r'"hasMore":(.*?),"', resp_text, re.S).group(1)
    except:
        print(url_api)
        hasMore = True
    return offset_list, hasMore


ua = fake_useragent.UserAgent()
headers = {
    "user-agent": ua.random,
    "Host": "m.10jqka.com.cn",
    "Connection": "Keep-Alive",
    "Accept-Encoding": "gzip",
}



for one_code, one_name in zip(["00001"], ["平安银行"]): # 直接拿 生产环境 的代码改过来的, 所以这里
    url = f"https://m.10jqka.com.cn/app/timeline/v1/list/33/{one_code}/first"
    resp = requests.get(url, headers=headers)
    resp_text = resp.text
```

```python
# todo: 因为：{"errorCode":0,"errorMsg":"","hasMore":false,"pageItems":[{"info":{"itemId":"1_640405988
try:
  offset = re.findall(r'"offset":"(\d+\.\d+)"', resp_text, re.S)[0]
except:
  offset = "first"
while True:
  off_list, check = get_data_url(code=one_code, offset=offset, name=one_name)
  if check == "false":
    break
  else:
    offset = off_list[0]
```

{"errorCode":0,"errorMsg":"","hasMore":false,"pageItems":[{"info":{"itemId":"1_64080385
5","type":1,"jumpUrl":"https://news.10jqka.com.cn/m640803855/","webrsid":"seq_640803
855"},"combination":[{"title":{"content":"8月2日至8月4日，上海相关区域将开展“3天2
检”"}},{"bottomBar":{"recommend":"置顶","recommendType":2,"source":"同花顺7x24快
讯","time":1659263938000,"clicks":31902}}]},{"info":{"itemId":"1_640803687","type":1,"ju
mpUrl":"https://news.10jqka.com.cn/m640803687/","webrsid":"seq_640803687"},"combin
ation":[{"title":{"content":"监管出手“迷你基金”最新要求来了"}},{"bottomBar":{"recommen
d":"置顶","recommendType":2,"source":"同花顺7x24快讯","time":1659259856000,"clicks":
14480}}]},{"info":{"itemId":"1_640803867","type":1,"jumpUrl":"https://news.10jqka.com.c
n/m640803867/","webrsid":"seq_640803867"},"combination":[{"title":{"content":"2连板宝
馨科技：新能源高端智能制造项目近期完成土地摘牌"}},{"bottomBar":{"recommend":"","rec
ommendType":0,"source":"同花顺7x24快讯","time":1659264759000,"clicks":26064}}]},{"inf
o":{"itemId":"1_640803669","type":1,"jumpUrl":"https://news.10jqka.com.cn/m64080366
9/","webrsid":"seq_640803669"},"combination":[{"title":{"content":"四川路桥：与比亚迪等
签署补充协议 计划委托比亚迪对蜀能矿产磷酸铁锂项目进行管理"}},{"bottomBar":{"recomm
end":"热","recommendType":7,"source":"同花顺7x24快讯","time":1659259458000,"clicks":
28385}}]},{"info":{"itemId":"1_640803585","type":1,"jumpUrl":"https://news.10jqka.com.c
n/m640803585/","webrsid":"seq_640803585"},"combination":[{"title":{"content":"榨菜里
吃出脚指甲？涪陵榨菜回应：异物被丢弃已无法检测和查证，研判为青菜头根茎"}},{"bottom
Bar":{"recommend":"热","recommendType":7,"source":"同花顺7x24快讯","time":16592581
01000,"clicks":5961}}]},{"info":{"itemId":"1_640803530","type":1,"jumpUrl":"https://news.1
0jqka.com.cn/m640803530/","webrsid":"seq_640803530"},"combination":[{"title":{"conten
t":"绿康生化：拟收购江西纬科100%股权"}},{"bottomBar":{"recommend":"热","recommend
Type":7,"source":"同花顺7x24快讯","time":1659256728000,"clicks":11813}}]},{"info":{"item
Id":"1_640803481","type":1,"jumpUrl":"https://news.10jqka.com.cn/m640803481/","webr
sid":"seq_640803481"},"combination":[{"title":{"content":"中伟股份：将向特斯拉供应电池
材料三元前驱体产品"}},{"bottomBar":{"recommend":"热","recommendType":7,"source":"同
花顺7x24快讯","time":1659255902000,"clicks":18891}}]},{"info":{"itemId":"1_64080346
9","type":1,"jumpUrl":"https://news.10jqka.com.cn/m640803469/","webrsid":"seq_640803
469"},"combination":[{"title":{"content":"九安医疗：美国子公司日常经营重大合同履行完
毕"}},{"bottomBar":{"recommend":"热","recommendType":7,"source":"同花顺7x24快讯","ti
me":1659255671000,"clicks":64095}}]},{"info":{"itemId":"1_640803429","type":1,"jumpUr
l":"https://news.10jqka.com.cn/m640803429/","webrsid":"seq_640803429"},"combinatio
n":[{"title":{"content":"四连板禾盛新材：实控人被判没收个人全部财产 公司存在实际控制权
变动的风险"}},{"bottomBar":{"recommend":"","recommendType":0,"source":"同花顺7x24
快讯","time":1659254846000,"clicks":12407}}]},{"info":{"itemId":"1_640803394","type":1,"j
umpUrl":"https://news.10jqka.com.cn/m640803394/","webrsid":"seq_640803394"},"combi
nation":[{"title":{"content":"艾比森：上半年净利8142万元 中报拟10派15元"}},{"bottomBa
r":{"recommend":"热","recommendType":7,"source":"同花顺7x24快讯","time":1659253993
000,"clicks":7981}}]},{"info":{"itemId":"1_640803339","type":1,"jumpUrl":"https://news.10j
qka.com.cn/m640803339/","webrsid":"seq_640803339"},"combination":[{"title":{"conten
t":"中国神华：王祥喜请辞董事长等职务"}},{"bottomBar":{"recommend":"热","recommendT
ype":7,"source":"同花顺7x24快讯","time":1659253078000,"clicks":6299}}]},{"info":{"itemI
d":"1_640803301","type":1,"jumpUrl":"https://news.10jqka.com.cn/m640803301/","webrs
id":"seq_640803301"},"combination":[{"title":{"content":"周末要闻回顾：中国7月官方制造
业PMI为49%"}},{"bottomBar":{"recommend":"","recommendType":0,"source":"同花顺7x24
快讯","time":1659252320000,"clicks":25359}}]},{"info":{"itemId":"1_640803131","type":1,"j
umpUrl":"https://news.10jqka.com.cn/m640803131/","webrsid":"seq_640803131"},"combi

```
nation":[{"title":{"content":"佩洛西亚洲行行程公布，未提及台湾地区"}},{"bottomBar":{"recommend":"热","recommendType":7,"source":"同花顺7x24快讯","time":1659246485000,"clicks":359170}}]},{"info":{"itemId":"1_640803073","type":1,"jumpUrl":"https://news.10jqka.com.cn/m640803073/","webrsid":"seq_640803073"},"combination":[{"title":{"content":"人民网发布《2021-2022游戏企业社会责任报告》"}},{"bottomBar":{"recommend":"热","recommendType":7,"source":"同花顺7x24快讯","time":1659244166000,"clicks":9181}}]},{"info":{"itemId":"1_640803039","type":1,"jumpUrl":"https://news.10jqka.com.cn/m640803039/","webrsid":"seq_640803039"},"combination":[{"title":{"content":"中信建投：7月市场靴子落地，8月行情将产生新的定价逻辑"}},{"bottomBar":{"recommend":"热","recommendType":7,"source":"同花顺7x24快讯","time":1659242925000,"clicks":4232}}]},{"info":{"itemId":"1_640803035","type":1,"jumpUrl":"https://news.10jqka.com.cn/m640803035/","webrsid":"seq_640803035"},"combination":[{"title":{"content":"“二舅币”发行人疑似诈骗130万美金后跑路"}},{"bottomBar":{"recommend":"热","recommendType":7,"source":"证券时报e公司","time":1659242485000,"clicks":45564}}]},{"info":{"itemId":"1_640803012","type":1,"jumpUrl":"https://news.10jqka.com.cn/m640803012/","webrsid":"seq_640803012"},"combination":[{"title":{"content":"空军发言人：空军多型战机绕飞祖国宝岛 坚定不移捍卫国家主权和领土完整"}},{"bottomBar":{"recommend":"热","recommendType":7,"source":"同花顺7x24快讯","time":1659241528000,"clicks":476862}}]},{"info":{"itemId":"1_640802877","type":1,"jumpUrl":"https://news.10jqka.com.cn/m640802877/","webrsid":"seq_640802877"},"combination":[{"title":{"content":"运油-20投入练兵备战 和歼-16开展海上空中加油训练"}},{"bottomBar":{"recommend":"热","recommendType":7,"source":"同花顺7x24快讯","time":1659237262000,"clicks":4628}}]},{"info":{"itemId":"1_640802843","type":1,"jumpUrl":"https://news.10jqka.com.cn/m640802843/","webrsid":"seq_640802843"},"combination":[{"title":{"content":"恒驰：恒驰5 8月1日零点开启大定"}},{"bottomBar":{"recommend":"热","recommendType":7,"source":"同花顺7x24快讯","time":1659236221000,"clicks":16633}}]},{"info":{"itemId":"1_640802704","type":1,"jumpUrl":"https://news.10jqka.com.cn/m640802704/","webrsid":"seq_640802704"},"combination":[{"title":{"content":"国家统计局：7月建筑业商务活动指数为59.2%，生产活动有所加快"}},{"bottomBar":{"recommend":"热","recommendType":7,"source":"同花顺7x24快讯","time":1659232138000,"clicks":628}}]}]}
```