爬虫实战——re方法抓取ajax异步加载的工业信息化部政策

有时候html源代码不包含数据资料，这是因为开发者设置了异步加载数据的策略，本文以工业信息化部ajax抓取政策举例。

观察网址

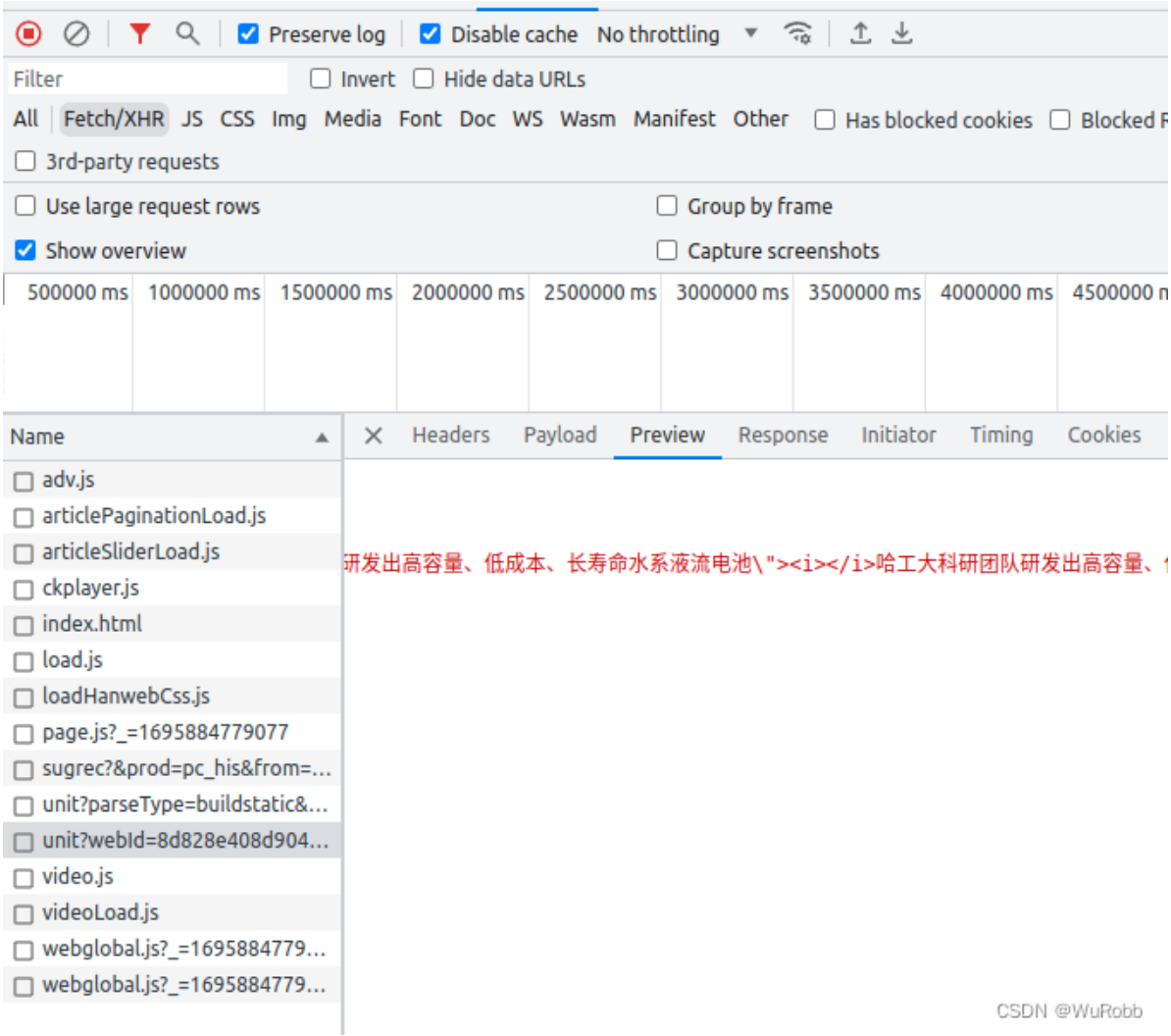
获取一下源代码

```
requests.get('https://wap.miit.gov.cn/xwdt/gxdt/bsdw/index.html',headers=head).
text
```

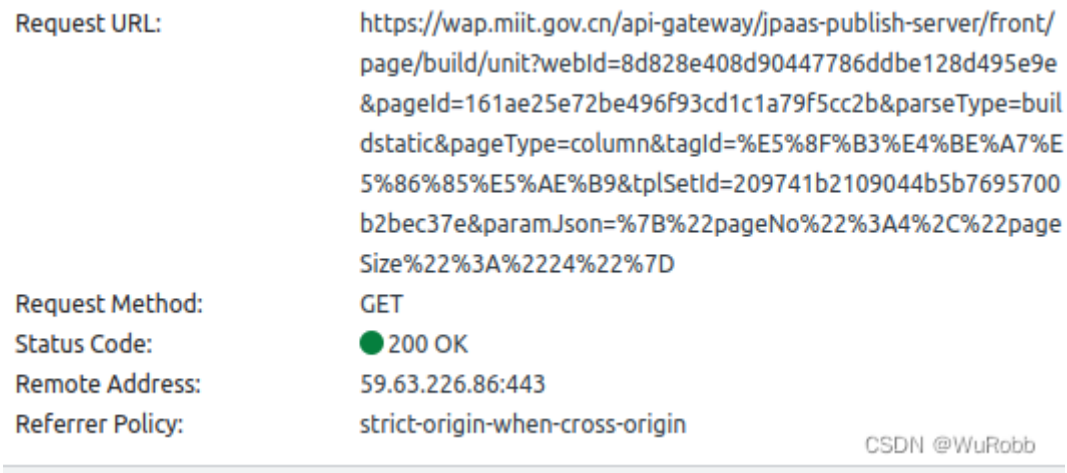

```
'<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
tional.dtd">\n<html xmlns="http://www.w3.org/1999/xhtml">\n<head>\n    <meta http-equiv="Content-Type" content
="text/html; charset=utf-8" />\n    <meta http-equiv="X-UA-Compatible" content="IE=edge">\n    <meta http-equi
v="X-UA-Compatible" content="IE=8">\n    \n\t<META HTTP-EQUIV="pragma" CONTENT="no-cache"><META HTTP-EQUIV="Ca
che-Control" CONTENT="no-store, must-revalidate">\n</META></META></meta></meta><script src="/cms_files/defaul
t/script/jslib/jquery.js"></script>\r\n <script src="/cms_files/default/script/jslib/jquery.cookie.js"></scrip
t>\r\n <script src="/cms_files/default/script/layui/layui.all.js"></script>\r\n <script src="/cms_files/defaul
t/script/layui/layer/layer.js"></script>\r\n <script src="/script/webglobal.js"></script>\r\n <script src="/pl
ugins/libs/require.js"></script>\r\n <script src="/plugins/libs/load.js"></script>\r\n <script src="/script/we
bgray.js"></script>\r\n <link rel="stylesheet" href="/cms_files/default/css/iconfont.css" />\r\n<link rel="sty
lesheet" href="/cms_files/default/script/layui/css/layui.css" />\r\n<title>部属单位</title>\n<meta name="MakeTi
e" content="2023-09-26 14:11:37">\n<meta name="SiteName" content="中华人民共和国工业和信息化部">\n<meta name="SiteD
omain" content="www.miit.gov.cn">\n<meta name="SiteIDCode" content="bm07000001">\n<meta name="ColumnName" cont
ent="部属单位">\n<meta name="ColumnType" content="部属单位">\n<meta name="ColumnDescription" content="部属单位">\r
<meta name="ColumnKeywords" content="部属单位">\n<meta id="pagetype" content="2">\n</meta></meta></meta></meta>
</meta></meta></meta></meta></meta></head>\n<body id="page_type" alt="列表页" class="lmy-wrapper ty_lmy">\n<lin
k type="text/css" href="/cms_files/webmiit/tplobject/gxbsy/811d480f91e84255b1a89e4b276efc0e/images/layout.css"
rel="stylesheet" />\n<link href="/cms_files/webmiit/tplobject/gxbsy/811d480f91e84255b1a89e4b276efc0e/images/ma
in.css" type="text/css" rel="stylesheet">\n<script type="text/javascript" src="/cms_files/webmiit/tplobject/gx
bsy/811d480f91e84255b1a89e4b276efc0e/images/pdy_tyyr.js"></script>\n\n<div class="w1100 head clearfix"><script
src="/cms_files/filemanager/script/201911/6b79016112cc4d66aeb22aa6d9d999dc.js" type="text/javascript" charset
="utf-8"></script>\n</div>\n<script src="/cms_files/filemanager/script/20201/f77a47db235946388a3ee648b566c5ac.
js" type="text/javascript" charset="utf-8"></script>\n<div style="background:url(/cms_files/webmiit/tplobject/
gxbsy/811d480f91e84255b1a89e4b276efc0e/images/7037090.png) repeat-x; width:100%; height:10px;"></div>\n\n<div
class="w980 center mnav">\n\t<span>\n\t\t\t<a href="/index.html">首页</a>  >  \n\t\t\t<a href="/xwdt/
index.html">新闻动态</a>  >  \n\t\t\t<a href="/xwdt/gxdt/index.html">工信动态</a>  >  \n\t\t
\t<a href="/xwdt/gxdt/bsdw/index.html">部属单位</a>\n\t</span>\n</div>\n\n<div class="clist w1100 center">\n
<table width="100%" cellpadding="0" cellspacing="0">\n        <tr>\n            <td width="249" bgcolor="" val
ign="top">\n\n                <div class="ctree">\n                    <div class="ctree_title">\n
<h1>工信动态\n</h1>                    </div>\n\n                    <dl>\n<dt><a href="/xwdt/szyw/index.html">
时政要闻</a></dt>\n<dt><a href="/xwdt/gxdt/index.html">工信动态</a></dt>\n<dt><a href="/xwdt/dwjl/index.html">对
交流</a></dt>\n<dt><a href="/xwdt/spxw/index.html">视频新闻</a></dt>\n<dt><a href="/xwdt/tpxwa22/index.html">图
新闻</a></dt>\n                    </dl>\n\t\t\t\t\t<div class="ctree-3">\n<dd>\t<a href="/xwdt/gxdt/ldhd/inde
x.html">» 领导活动</a>\n\t<a href="/xwdt/gxdt/sjdt/index.html">» 司局动态</a>\n\t<a href="/xwdt/gxdt/dfgz/index.
tml">» 地方工作</a>\n\t<a href="/xwdt/gxdt/bsdw/index.html">» 部属单位</a>\n</dd>\t\t\t\t\t\t</div>\n\t\t\t\t\t\t<di
class="ctree-4">\n\t\t\t\t\t\t</div>\n\n                    <div style="height:30px;"> </div>\n
</div>\n     \n            </td>\n            <td width="61"> </td>\n            <td valign="top">\n
<div class="clist_r_title">\n                    <span><dt>部属单位</dt>\n</span>                </div>\n\n
<span id="comp_5790085">\n                    <div class="clist_con" style="min-height:600px; _height:600px;">
\n<script>\r\nvar authorizedReadUnitId = "d806176763b84a66aa561f71020025ff";\r\n</script><script id="d8061767
3b84a66aa561f71020025ff" src="/cms_files/default/script/AuthorizedRead/unitbuild.js?v=3.1.3-GXB" webId="8d828e
408d90447786ddbe128d495e9e" parseType="buildstatic" url="/api-gateway/jpaas-publish-server/front/page/build/un
it" queryData="{\'parseType\':\'buildstatic\',\'webId\':\'8d828e408d90447786ddbe128d495e9e\',\'tplSetId\':\'20
9741b2109044b5b7695700b2bec37e\',\'pageType\':\'column\',\'tagId\':\'右侧内容\',\'editType\':\'null\',\'pageId
\':\'161ae25e72be496f93cd1c1a79f5cc2b\'}"></script>\r\n                    </div>\n                </span>\n
<div style="height:30px;"> </div>\n            </td>\n        </tr>\n    </table>\n</div>\n\n<script src
="/cms_files/filemanager/script/20201/840397482ca5455aa2209d4942e43e72.js" type="text/javascript" charset="utf
-8"></script></link></body>\n</html>'
```


发现并没有 li 标签。

在看一下浏览器NETWORK中发现 li 标签在一个 XMLHttpRequest 网络中



查看一下这个请求的请求头，是一个 get 请求，观察一下 url 和 payload 发现原网址为 https://wap.miit.gov.cn/api-gateway/jpaas-publish-server/front/page/build/unit，？后面为patload的参数。



至此，目标 url 就解析完毕了。

获取数据 li 标签所在的 html

获得目标url后，直接对该网址请求

```
url = 'https://wap.miit.gov.cn/api-gateway/jpaas-publish-server/front/page/build/unit'
param_dict = {
        "webId": "8d828e408d90447786ddbe128d495e9e",
        "pageId": "161ae25e72be496f93cd1c1a79f5cc2b",
```

```
        "parseType": "buildstatic",
        "pageType": "column",
        "tagId": "右侧内容",
        "tplSetId": "209741b2109044b5b7695700b2bec37e",
        "paramJson": '{{"pageNo": {}, "pageSize": "24"}}'.format(1),
        # 'editType': 'null'
    }
    response = requests.get(url, headers=head, params=param_dict)
    response.encoding = "utf-8"
    print(response.text)
```


```
{"roles":null,"permissions":null,"success":true,"code":"200","data":{"html":"\t\n<div id=\"右侧内容\">\n  <div c
lass=\"page-content\">\n\t<ul>\n\t\t\t<li class=\"cf\">\n\t\t\t\t<a class=\"fl\" href=\"/xwdt/gxdt/bsdw/art/20
23/art_4ee5ac5cd4e7430b809eaac04509ecf5.html\" target=\"_blank\" title=\"哈工大揭示疟原虫多药耐药蛋白结构和调节机制
\"><i></i>哈工大揭示疟原虫多药耐药蛋白结构和调节机制</a>\n\t\t\t\t<span class=\"fr\">2023-09-25</span>\n\t\t\t</li>\n
\t\t\t<li class=\"cf\">\n\t\t\t\t<a class=\"fl\" href=\"/xwdt/gxdt/bsdw/art/2023/art_db58004d58494535ab702c577
bcb0eb4.html\" target=\"_blank\" title=\"哈工大主办第三届全国搅拌摩擦焊接与加工学术会议\"><i></i>哈工大主办第三届全国搅拌
摩擦焊接与加工学术会议</a>\n\t\t\t\t<span class=\"fr\">2023-09-15</span>\n\t\t\t</li>\n\t\t\t<li class=\"cf\">\n\t
\t\t\t<a class=\"fl\" href=\"/xwdt/gxdt/bsdw/art/2023/art_1926baa204a04be2b517d62dee6da816.html\" target=\"_bl
ank\" title=\"哈工程成功举办70周年校庆大学校长论坛\"><i></i>哈工程成功举办70周年校庆大学校长论坛</a>\n\t\t\t\t<span class=
\"fr\">2023-09-12</span>\n\t\t\t</li>\n\t\t\t<li class=\"cf\">\n\t\t\t\t<a class=\"fl\" href=\"/xwdt/gxdt/bsd
w/art/2023/art_3892dffa484c4bc689684b844fab2d9e.html\" target=\"_blank\" title=\"哈工大承办微生物生态专委会2023年学
术年会暨全球华人学者环境科技前沿论坛\"><i></i>哈工大承办微生物生态专委会2023年学术年会暨全球华人学者环境科技前沿论坛</a>\n\t\t
\t\t<span class=\"fr\">2023-09-07</span>\n\t\t\t</li>\n\t\t\t<li class=\"cf\">\n\t\t\t\t<a class=\"fl\" href=
\"/xwdt/gxdt/bsdw/art/2023/art_0bb0dda0a13d49a3940a3740acc08ad4.html\" target=\"_blank\" title=\"哈工程成功举办20
23世界机电一体化与自动化国际会议\"><i></i>哈工程成功举办2023世界机电一体化与自动化国际会议</a>\n\t\t\t\t<span class=\"fr\">
2023-09-04</span>\n\t\t\t</li>\n\t\t\t<li class=\"cf border-line\">\n\t\t\t\t<a class=\"fl\" href=\"/xwdt/gxd
t/bsdw/art/2023/art_9bda7d2b1d3a47db8ae3b988500a3209.html\" target=\"_blank\" title=\"哈工程成功举办中国惯性技术学
会2023年科技工作者研讨会\"><i></i>哈工程成功举办中国惯性技术学会2023年科技工作者研讨会</a>\n\t\t\t\t<span class=\"fr\">202
3-08-31</span>\n\t\t\t</li>\n\t\t\t<li class=\"cf\">\n\t\t\t\t<a class=\"fl\" href=\"/xwdt/gxdt/bsdw/art/2023/
art_0a114cd0ba764cbbb64f95ba7baeafdb.html\" target=\"_blank\" title=\"西工大团队在《自然·材料》发表原创性研究成果提出
```


返回的是一个json文件，用 json 读取 html 文本

```
data = json.loads(response.text)["data"]
html = data["html"]
html
```

至此 li 标签所在的html就解析完毕

re获取li标签信息

```
# .不匹配换行符，所以去掉换行符
    # \t缩进符号不需要可替换
html = re.sub(r"\n|\t", "", html)

# 获取li列表
re_compile = re.compile(r'<li class="cf">(.*?)</li>')
li_list = re_compile.findall(html)

# 获取li列表信息
li_list = [
    [  # a_href
        "https://wap.miit.gov.cn" + re.compile(r'href="(.*?)"').findall(li)[0],
        # a_title
        re.compile(r'title="(.*?)"').findall(li)[0],
        # a_time
        re.compile(r'<span class="fr">(.*?)</span>').findall(li)[0],
    ]
    for li in li_list
]
```

```
[['https://wap.miit.gov.cn/xwdt/gxdt/bsdw/art/2023/art_4ee5ac5cd4e7430b809eaac04509ecf5.html',
  '哈工大揭示疟原虫多药耐药蛋白结构和调节机制',
  '2023-09-25'],
 ['https://wap.miit.gov.cn/xwdt/gxdt/bsdw/art/2023/art_db58004d58494535ab702c577bcb0eb4.html',
  '哈工大主办第三届全国搅拌摩擦焊接与加工学术会议',
  '2023-09-15'],
 ['https://wap.miit.gov.cn/xwdt/gxdt/bsdw/art/2023/art_1926baa204a04be2b517d62dee6da816.html',
  '哈工程成功举办70周年校庆大学校长论坛',
  '2023-09-12'],
 ['https://wap.miit.gov.cn/xwdt/gxdt/bsdw/art/2023/art_3892dffa484c4bc689684b844fab2d9e.html',
  '哈工大承办微生物生态专委会2023年学术年会暨全球华人学者环境科技前沿论坛',
  '2023-09-07'],
 ['https://wap.miit.gov.cn/xwdt/gxdt/bsdw/art/2023/art_0bb0dda0a13d49a3940a3740acc08ad4.html',
  '哈工程成功举办2023世界机电一体化与自动化国际会议',
  '2023-09-04'],
 ['https://wap.miit.gov.cn/xwdt/gxdt/bsdw/art/2023/art_0a114cd0ba764cbbb64f95ba7baeafdb.html',
  '西工大团队在《自然·材料》发表原创性研究成果提出低维量子结构设计新机制',
  '2023-08-30'],
 ['https://wap.miit.gov.cn/xwdt/gxdt/bsdw/art/2023/art_a2d08d8da0dd42d2879d11cfa038487e.html',
  '西工大团队在立体发散式合成手性2-/3-烷基取代吡咯烷的领域取得重要进展',
  '2023-08-28'],
```


这时候各文本的url、标题、发布时间就得到了

对每一个网站进行抓取

用正则表达式对需要信息进行匹配，不赘述

```
def crawl_li_list(li_list, db_insert_data):
    for li in li_list:
        publish_time = datetime.strptime(li[-1], "%Y-%m-%d")
        if end_time <= publish_time < begin_time:
            a_url = li[0]
            a_request = requests.get(a_url, headers=head)
            a_request.encoding = "utf-8"
            ## 抓取分类
            html = re.sub(r"\n|\t", "", a_request.text)
            a_compile = re.compile('<div class="w980 center mnav">.*?</div>')
            # .不匹配换行符，所以去掉换行符
            # \t缩进符号不需要可替换
            a_html = a_compile.findall(html)
            file_path = []
            db_insert_data.append(
                [
                    # crawl_time
                    datetime.now(),
                    # publish_time
                    publish_time,
                    # 原始网址
                    a_url,
                    # 网站模块
                    re.compile('<a href="(.*?)">(.*?)</a>').findall(
                        re.compile('<div class="w980 center mnav">.*?
</div>').findall(html)[
                            0
                        ]
                    )[-1][-1],
                    # 标题
                    li[1],
                    # 作者或来源
                    "中华人民共和国工业和信息化部",
                    # 文章内容
                    '\n'.join(re.compile('>(.*?)</').findall(
```

```
    re.compile('<div class="ccontent center" id="con_con" (.*?)">(.*?)
</div>').findall(html)[0][-1]
    )),
                    # 附件 存储附件地址 media/app03/data/
                    "\n".join(file_path),
                ]
            )

    return db_insert_data
```

看一下效果

```
[[datetime.datetime(2023, 9, 26, 17, 40, 50, 178221),
  datetime.datetime(2023, 9, 25, 0, 0),
  'https://wap.miit.gov.cn/xwdt/gxdt/bsdw/art/2023/art_4ee5ac5cd4e7430b809eaac04509ecf5.html',
  '部属单位',
  '哈工大揭示疟原虫多药耐药蛋白结构和调节机制',
  '中华人民共和国工业和信息化部',
  '近日，哈尔滨工业大学生命科学中心李明晖课题组在《美国科学院院刊》（PNAS）上在线发表题为《恶性疟原虫多药耐药蛋白1
（PfMDR1）的结构揭示N端调节结构域》（The structure of Plasmodium falciparum multidrug resistance prote
in 1 reveals an N-terminal regulatory domain）的研究论文，揭示疟原虫多药耐药蛋白结构和调节机制，为抗疟药物
的理性设计和优化提供结构基础和理论依据。\n<p><br />',
  ''],
```


完整代码

最后附上完整代码

```
import json
import os
import re
from datetime import datetime, timedelta

import numpy as np
import pandas as pd
import requests

def get_today_and_lastday():
    t = datetime.now()
    oneday = timedelta(days=1)
    yesterday = t - oneday
    return datetime(year=t.year, month=t.month, day=t.day), datetime(
        year=yesterday.year, month=yesterday.month, day=yesterday.day
    )


def str_to_datetime(s):
    return datetime.strptime(s, "%Y-%m-%d %H:%M:%S")


def get_l_list(url, end_time, page=1):
    # get 请求所需参数
    param_dict = {
        "webId": "8d828e408d90447786ddbe128d495e9e",
        "pageId": "161ae25e72be496f93cd1c1a79f5cc2b",
        "parseType": "buildstatic",
        "pageType": "column",
        "tagId": "右侧内容",
        "tplSetId": "209741b2109044b5b7695700b2bec37e",
        "paramJson": '{{"pageNo": {}, "pageSize": "24"}}'.format(page),
        # 'editType': 'null'
```

```
    }
    response = requests.get(url, headers=head, params=param_dict)
    response.encoding = "utf-8"
    data = json.loads(response.text)["data"]
    html = data["html"]
    # .不匹配换行符，所以去掉换行符
    # \t缩进符号不需要可替换
    html = re.sub(r"\n|\t", "", html)

    # 获取li列表
    re_compile = re.compile(r'<li class="cf">(.*?)</li>')
    li_list = re_compile.findall(html)

    # 获取li列表信息
    li_list = [
        [  # a_href
            "https://wap.miit.gov.cn" + re.compile(r'href="(.*?)"').findall(li)
[0],
            # a_title
            re.compile(r'title="(.*?)"').findall(li)[0],
            # a_time
            re.compile(r'<span class="fr">(.*?)</span>').findall(li)[0],
        ]
        for li in li_list
    ]

    # 如果li列表时间未到设置停止时间继续爬取
    publish_time = datetime.strptime(li_list[-1][-1], "%Y-%m-%d")
    if publish_time < end_time:
        return li_list
    page += 1
    li_list += get_l_list(url, end_time, page)
    return li_list


def crawl_li_list(li_list, db_insert_data):
    for li in li_list:
        publish_time = datetime.strptime(li[-1], "%Y-%m-%d")
        if end_time <= publish_time < begin_time:
            a_url = li[0]
            a_request = requests.get(a_url, headers=head)
            a_request.encoding = "utf-8"
            ## 抓取分类
            html = re.sub(r"\n|\t", "", a_request.text)
            a_compile = re.compile('<div class="w980 center mnav">.*?</div>')
            # .不匹配换行符，所以去掉换行符
            # \t缩进符号不需要可替换
            a_html = a_compile.findall(html)
            file_path = []
            db_insert_data.append(
                [
                    # crawl_time
                    datetime.now(),
                    # publish_time
                    publish_time,
                    # 原始网址
                    a_url,
                    # 网站模块
```

```
                    re.compile('<a href="(.*?)">(.*?)</a>').findall(
                        re.compile('<div class="w980 center mnav">.*?
</div>').findall(
                            html
                        )[0]
                    )[-1][-1],
                    # 标题
                    li[1],
                    # 作者或来源
                    "中华人民共和国工业和信息化部",
                    # 文章内容
                    "\n".join(
                        re.compile(">(.*?)</").findall(
                            re.compile(
                                '<div class="ccontent center" id="con_con"
(.*?)">(.*?)</div>'
                            ).findall(html)[0][-1]
                        )
                    ),
                    # 附件 存储附件地址 media/app03/data/
                    "\n".join(file_path),
                ]
            )

    return db_insert_data


begin_time = get_today_and_lastday()[0]
end_time = datetime(2023, 1, 1)

head = {
    "User-Agent": np.random.choice(
        [
            "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like
Gecko) Chrome/41.0.2228.0 Safari/537.36",
            "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_10_1) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/41.0.2227.1 Safari/537.36",
            "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like
Gecko) Chrome/41.0.2227.0 Safari/537.36",
            "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/41.0.2227.0 Safari/537.36",
            "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/41.0.2226.0 Safari/537.36",
            "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0;
chromeframe/13.0.782.215)",
            "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0;
chromeframe/11.0.696.57)",
            "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0)
chromeframe/10.0.648.205",
            "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/4.0;
GTB7.4; InfoPath.1; SV1; .NET CLR 2.8.52393; WOW64; en-US)",
            "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.0; Trident/5.0;
chromeframe/11.0.696.57)",
            "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.0; Trident/4.0;
GTB7.4; InfoPath.3; SV1; .NET CLR 3.1.76908; WOW64; en-US)",
            "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_3)
AppleWebKit/537.75.14 (KHTML, like Gecko) Version/7.0.3 Safari/7046A194A",
```

```
            "Mozilla/5.0 (iPad; CPU OS 6_0 like Mac OS X) AppleWebKit/536.26
(KHTML, like Gecko) Version/6.0 Mobile/10A5355d Safari/8536.25",
            "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/537.13+
(KHTML, like Gecko) Version/5.1.7 Safari/534.57.2",
            "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_7_3)
AppleWebKit/534.55.3 (KHTML, like Gecko) Version/5.1.3 Safari/534.53.10",
            "Mozilla/5.0 (iPad; CPU OS 5_1 like Mac OS X) AppleWebKit/534.46
(KHTML, like Gecko ) Version/5.1 Mobile/9B176 Safari/7534.48.3",
        ],
    ),
    "Accept-Language": "en-US,en;q=0.9,zh-CN;q=0.8,zh;q=0.7",
}

init_url = "http://js.shaanxi.gov.cn"
url_list = [
    "http://js.shaanxi.gov.cn/zixun/list2006",
    "http://js.shaanxi.gov.cn/zixun/list2080",
    "http://js.shaanxi.gov.cn/zixun/list2077",
    "http://js.shaanxi.gov.cn/zixun/list2010",
]

url = "https://wap.miit.gov.cn/api-gateway/jpaas-publish-
server/front/page/build/unit"
li_list = get_l_list(url, end_time, page=1)
db_insert_data = []
li_list = get_l_list(url, end_time)
crawl_li_list(li_list, db_insert_data)
```

参考：https://www.bilibili.com/video/BV1NX4y1X7AE?p=9&spm_id_from=pageDriver&vd_source=e78d869c28d2119248eff5d85f195ece