关于解决requests请求可行，scrapy请求不行的问题

### 1 问题引出

在使用scrapy的过程中可能出现requests请求url可以得到正确的返回结果，而使用scrapy请求这会返回202,400等状态码的现象。

**出现该问题的前提条件是：不论是GET或POST请求，需求传递Header、Cookies参数。**

### 2 实例说明

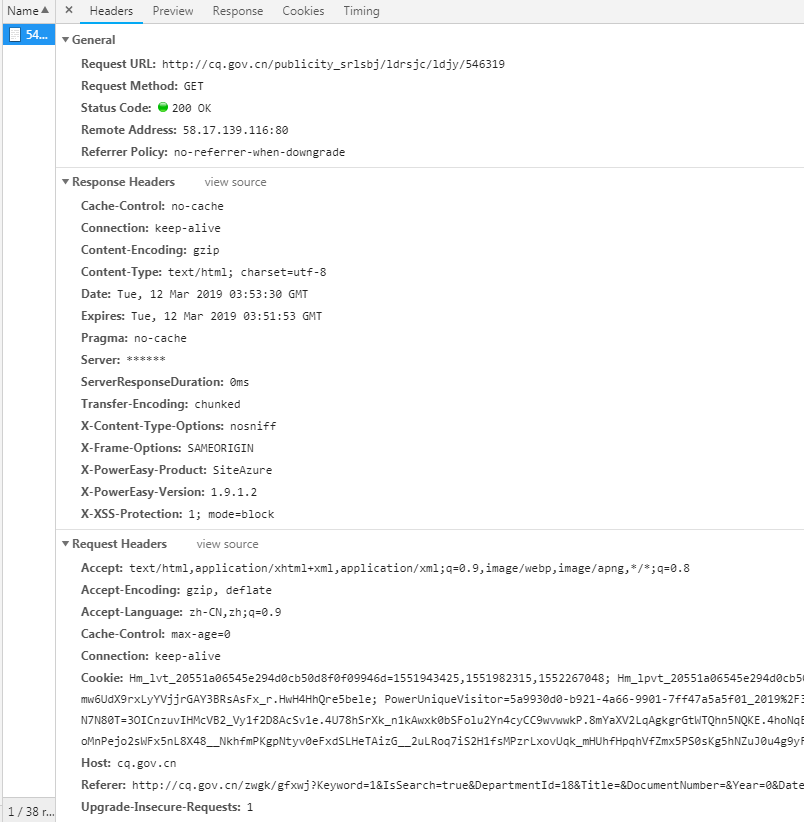
#### 2.1 Get请求实例说明

类似网站有：

<http://cq.gov.cn/zwgk/gfxwj?Keyword=1&IsSearch=true&DepartmentId=18&Title=&DocumentNumber=&Year=0&DateNumber=&PageId=1>

<http://cq.gov.cn/publicity_srlsbj/ldrsjc/ldjy/546319>

以重庆市人民政府的详情页为例：<http://cq.gov.cn/publicity_srlsbj/ldrsjc/ldjy/546319>



当爬虫代码为如下格式时：

import scrapy

class ExampleSpider(scrapy.Spider):

name = 'debug'

url = 'http://cq.gov.cn/publicity\_srlsbj/ldrsjc/shbz/546318'

header = {

'Accept': 'text/html,q=0.8',

'Accept-Encoding': 'gzip',

'Host': 'cq.gov.cn',

'Cookie': 'Hm\_lvt\_20551a06545e294d0cb50d8f0f09946d=1551943425,1551982315,1552267048; Hm\_lpvt\_20551a06545e294d0cb50d8f0f09946d=1552267048; FSSBBIl1UgzbN7N80S=G.jFzQSF.uARSxpCf0Krmw6UdX9rxLyYVjjrGAY3BRsAsFx\_r.HwH4HhQre5bele; PowerUniqueVisitor=5a9930d0-b921-4a66-9901-7ff47a5a5f01\_2019%2F3%2F12%200%3A00%3A00; PowerLeaveSitePrompts=NoShow; FSSBBIl1UgzbN7N80T=3OICnzuvIHMcVB2\_Vy1f2D8AcSv1e.4U78hSrXk\_n1kAwxk0bSFolu2Yn4cyCC9wvwwkP.8mYaXV2LqAgkgrGtWTQhn5NQKE.4hoNqE2.VRPONFMX6kRyVOgW1BG7DdYUrvrfhIPRYdjIVjHuGZ.qHyC6aCDpxfbeECYm4oMnPejo2sWFx5nL8X48\_\_NkhfmPKgpNtyv0eFxdSLHeTAizG\_\_2uLRoq7iS2H1fsMPzrLxovUqk\_mHUhfHpqhVfZmx5PS0sKg5hNZuJ0u4g9yFsm3qWQ8IVBnd.qwAYfk4Au3WbGh8pktIVKzI46ULJVOtM3EV',

'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/67.0.3396.99 Safari/537.36'

}

cookies = {

'Hm\_lvt\_20551a06545e294d0cb50d8f0f09946d': '1551943425,1551982315,1552267048',

'Hm\_lpvt\_20551a06545e294d0cb50d8f0f09946d': '1552267048',

'FSSBBIl1UgzbN7N80S': 'G.jFzQSF.uARSxpCf0Krmw6UdX9rxLyYVjjrGAY3BRsAsFx\_r.HwH4HhQre5bele',

'PowerUniqueVisitor': '541faedb-bc22-4989-a89e-91eee633bc3d\_2019%2F3%2F11%200%3A00%3A00',

'FSSBBIl1UgzbN7N80T': '3Kcfv3Xat\_DYrFgzrWHCaukoUl\_mL8bqy.KbVM8zvv8oEZ8OuloApDgcviIXOkytnLqV68k1qmukawwozC2KTdpPxu58iIoZcIAWDfRmFTk\_BY.M01STJX\_c8rK4RxBOApOTmgPwhvXNo6uReBu4BD5yKj3DojL\_xF86NVEE42LnHHyLpE4I1it6rve6Ox4ohrWIKDMCKehqGLW7JpxCDv9BVE\_yfAug1JwDMafabrxYJ2D2f0A1TH\_jIVQm2k3i0rqwDe.fH0KlYNzDLMcHtF5p9FWoUxSfgyMdreFpUypTYv1y4Cg9dFxf49UQOWemMjWE'

}

def start\_requests(self):

yield scrapy.Request(

url=self.url,

headers=self.header,

callback=self.parse\_ss,

dont\_filter=True,

)

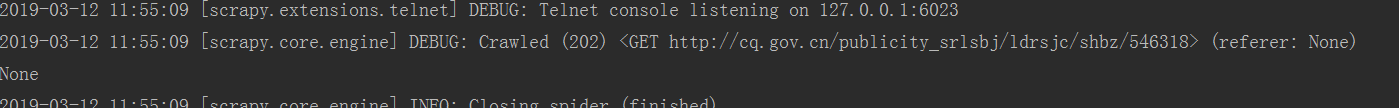
def parse\_ss(self, response):

xpath = '//td[contains(string(.), "名称")]/following-sibling::td/text()'

ss = response.xpath(xpath).extract\_first()

print(ss)

会出现以下格式错误：



**主要原因为：**

Cookies参数没有正确赋值，scrapy.Request中已经将Cookies独立出来作为一个参数，不能放在headers中，正确格式写法如下：

header = {

'Accept': 'text/html,q=0.8',

'Accept-Encoding': 'gzip',

'Host': 'cq.gov.cn',

}

cookies = {

'Hm\_lvt\_20551a06545e294d0cb50d8f0f09946d': '1551943425,1551982315,1552267048',

'Hm\_lpvt\_20551a06545e294d0cb50d8f0f09946d': '1552267048',

'FSSBBIl1UgzbN7N80S': 'G.jFzQSF.uARSxpCf0Krmw6UdX9rxLyYVjjrGAY3BRsAsFx\_r.HwH4HhQre5bele',

'PowerUniqueVisitor': '541faedb-bc22-4989-a89e-91eee633bc3d\_2019%2F3%2F11%200%3A00%3A00',

'FSSBBIl1UgzbN7N80T': '3Kcfv3Xat\_DYrFgzrWHCaukoUl\_mL8bqy.KbVM8zvv8oEZ8OuloApDgcviIXOkytnLqV68k1qmukawwozC2KTdpPxu58iIoZcIAWDfRmFTk\_BY.M01STJX\_c8rK4RxBOApOTmgPwhvXNo6uReBu4BD5yKj3DojL\_xF86NVEE42LnHHyLpE4I1it6rve6Ox4ohrWIKDMCKehqGLW7JpxCDv9BVE\_yfAug1JwDMafabrxYJ2D2f0A1TH\_jIVQm2k3i0rqwDe.fH0KlYNzDLMcHtF5p9FWoUxSfgyMdreFpUypTYv1y4Cg9dFxf49UQOWemMjWE'

}

yield scrapy.Request(

url=self.url,

headers=self.header,

cookies=self.cookies,

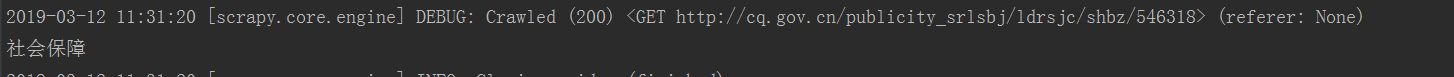
callback=self.parse\_ss,

dont\_filter=True,

)

注意：Cookies参数是以key，value存在，与网友复制下来的key=value不一样

修改后返回正确结果



#### 2.2 POST请求实例说明

实例对象为重庆市发改委信息分类目录中的计划规划：

网页首页为：

<http://www.cq.gov.cn/publicinfo/web/views/Show!index.action?orgSid=76>

请求网页为：

<http://www.cq.gov.cn/publicinfo/web/views/Show!query.action?MmEwMD=3s4X30NmPSIN4Em9gkWr0YWnvaTa1CMwW8yInIJ1Hw2ouzYA4OainQUtrA9SaJgvYh.uQKbA.edvO1MssPMalvn5oLAsTOIKxKMaV9yFrU.X8wA8SKXlpZH0pDy3IRh.15EC35pCFs4NfsL9s6V8i_f1uaY_ZzyxO3nKG9MLVBdJKObfRVWJFnWeWFSzdyBYmE1TmMtavG5SdzeO6sujRLk4l5c9NsZZIwFjmJwfOjZWMsQHxkDKvRXo00xggbFpwxiR6Nt1eh9fIlrmICUaAz5U4AOhxTqBv.9pW5AEeuee.AzwvTlyyLHMJmiT9IJy.3KqG4u7gxc8SZo3kkyha6fadILf0E1.GpJmOqr.wBqfned._y55xsATN4DQC401ZdTT>

（MmEwMD为加密参数，直接复制即可）

**请求函数建议使用：scrapy.FormRequest，原因：scrapy.Request函数虽然支持发起POST请求，但是不能提交表单参数，scrapy.FormRequest可行，同时scrapy.FormRequest继承了scrapy.Request所有方法，因为它们都继承Request类。**

**关于cookies的设置与上述的GET请求方式设置一样，独立设置。**

正确代码如下：

url = 'http://www.cq.gov.cn/publicinfo/web/views/' \

'Show!query.action?MmEwMD=3s4X30NmPSIN4Em9gkWr0YWnvaTa1CMwW8yInIJ1Hw2ouzYA4OainQUtrA9SaJgvYh.' \

'uQKbA.edvO1MssPMalvn5oLAsTOIKxKMaV9yFrU.X8wA8SKXlpZH0pDy3IRh.' \

'15EC35pCFs4NfsL9s6V8i\_f1uaY\_ZzyxO3nKG9MLVBdJKObfRVWJFnWeWFSzdyBYmE1TmMtavG5SdzeO6sujRLk' \

'4l5c9NsZZIwFjmJwfOjZWMsQHxkDKvRXo00xggbFpwxiR6Nt1eh9fIlrmICUaAz5U4AOhxTqBv.9pW5AEeuee.AzwvTl'\ 'yyLHMJmiT9IJy.3KqG4u7gxc8SZo3kkyha6fadILf0E1.GpJmOqr.wBqfned.\_y55xsATN4DQC401ZdTT'

header = {

'Accept': '\*/\*',

'Accept-Encoding': 'gzip, deflate',

'Accept-Language': 'zh-CN,zh;q=0.9',

'Connection': 'keep-alive',

'Content-Type': 'application/x-www-form-urlencoded; charset=UTF-8',

'Host':'www.cq.gov.cn',

'Origin': 'http://www.cq.gov.cn',

'Referer': 'http://www.cq.gov.cn/publicinfo/web/views/Show!index.action?orgSid=76',

'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/67.0.3396.99 Safari/537.36',

'X-Requested-With': 'XMLHttpRequest',

}

data = {

'datime':'1552368241783',

'pageNo':'1',

'pageSize':'13',

'category':'19',

'orgSid':'76',

'searchCont':'',

'searchType':'0',

'searchYear':'0',

'querytype':'1',

}

cookies = {

'JSESSIONID':'370161D7D3D63D737A0D6898252D6D44',

'FSSBBIl1UgzbN7N80S':'kWQIfQ\_b1yjJC7UzRfw23Vo0z5Lxyi\_4i0Gi8Mdh27b\_zgEJlgjUGqlqe01U4RXi',

'ASP.NET\_SessionId':'vltkelfpgrjwcojenoiftn55',

'Hm\_lvt\_20551a06545e294d0cb50d8f0f09946d':'1551943425,1551982315,1552267048',

'Hm\_lpvt\_20551a06545e294d0cb50d8f0f09946d':'1552368224',

'FSSBBIl1UgzbN7N80T':'3PvywmT1vqk55yRuenAKS9A2PrNvUootxB2E6Xpmt7yACQxo517R6zhwlfgbTZ9aA\_69Rv\_oXWOabOolKsovrKXNc3vS1gx3uUrYvqlfLvBUz6QnvEx7UmDXHIkpSlW2sYEQhgW8.nnNfWBQy26tkid5sn0eVK7mHRJasw7u2CTODJGcP57X3xT79a40Tt8FRanHWUjELzpjl1f1P35IvvEXcvFDj13Bgjep.t7Vl4b4QW8gvwyzGbAAug3yx0Rm1s9sCwuda5Pt00U7JFaYEjZg8lbdM2NO1MIBy3IzmoRH1yuYms8ZfaoLTgaUKypeVqM7'

}

def start\_requests(self):

yield scrapy.FormRequest(

url=self.url,

headers=self.header,

cookies=self.cookies,

formdata=self.data,

callback=self.parse\_ss,

dont\_filter=True,

)

#### 2.3 Header参数特别说明

**请求头中的Content-Length不要填入，否则同样会返回202或400。**

### Scrapy中cookies设置

参考链接：

<https://note.youdao.com/ynoteshare1/index.html?id=244294a2004c3fa3553229617f89a07d&type=note#/>

1. 首先，网上常见的链接关于COOKIES\_ENABLED的设置，其实并没有作用。

如果需要将cookie写在header中，需要在settings中设置，表示不使用scrapy默认的cookies，如果使用的话scrapy默认的cookies会将header中的cookies覆盖。 Cookies优先级可以查看上述参考链接。

DOWNLOADER\_MIDDLEWARES = {

'scrapy.downloadermiddlewares.cookies.CookiesMiddleware': None,

}

同时，如果将cookies写在header中，格式如下：

header = {

'Accept': '\*/\*',

'Accept-Encoding': 'gzip, deflate',

'Accept-Language': 'zh-CN,zh;q=0.9',

'Connection': 'keep-alive',

'Cookie':'JSESSIONID=370161D7D3D63D737A0D6898252D6D44;FSSBBIl1UgzbN7N80S=kWQIfQ\_b1yjJC7UzRfw23Vo0z5Lxyi\_4i0Gi8Mdh27b\_zgEJlgjUGqlqe01U4RXi;ASP.NET\_SessionId=vltkelfpgrjwcojenoiftn55;Hm\_lvt\_20551a06545e294d0cb50d8f0f09946d=1551943425,1551982315,1552267048;Hm\_lpvt\_20551a06545e294d0cb50d8f0f09946d=1552368224;FSSBBIl1UgzbN7N80T=3PvywmT1vqk55yRuenAKS9A2PrNvUootxB2E6Xpmt7yACQxo517R6zhwlfgbTZ9aA\_69Rv\_oXWOabOolKsovrKXNc3vS1gx3uUrYvqlfLvBUz6QnvEx7UmDXHIkpSlW2sYEQhgW8.nnNfWBQy26tkid5sn0eVK7mHRJasw7u2CTODJGcP57X3xT79a40Tt8FRanHWUjELzpjl1f1P35IvvEXcvFDj13Bgjep.t7Vl4b4QW8gvwyzGbAAug3yx0Rm1s9sCwuda5Pt00U7JFaYEjZg8lbdM2NO1MIBy3IzmoRH1yuYms8ZfaoLTgaUKypeVqM7',

'Content-Type': 'application/x-www-form-urlencoded; charset=UTF-8',

'Host':'www.cq.gov.cn',

'Origin': 'http://www.cq.gov.cn',

'Referer':'http://www.cq.gov.cn/publicinfo/web/views/Show!index.action?orgSid=76',

'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/67.0.3396.99 Safari/537.36',

'X-Requested-With': 'XMLHttpRequest',

}

1. 如果将cookie参数从header中抽取出来，者按照GET方法的格式编写，但是得在settings设置cookies，注释掉'scrapy.downloadermiddlewares.cookies.CookiesMiddleware': None,设置如下

DOWNLOADER\_MIDDLEWARES = {

# 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware': None,

}