Linux下Scrapy的使用

ps：我最早在大二的时候其实使用过scrapy框架来获取数据，但是因为考研期间大半年没有触碰过编程相关，导致我的编程能力严重下降，这回正好做一个项目，因此将这份本领捡了回来。

以下是正文:

一、Scrapy的安装和配置

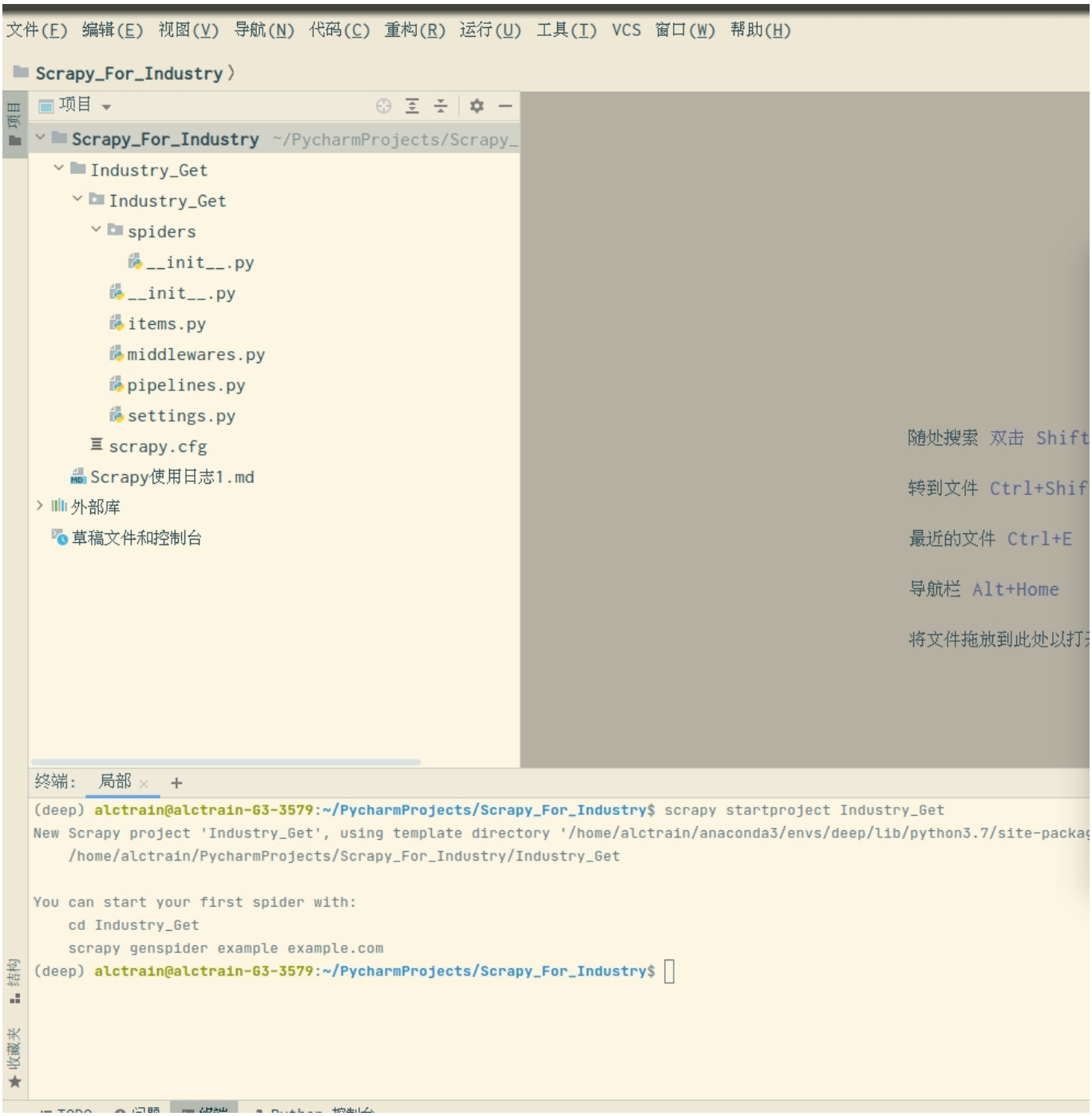
Scrapy的安装和配置一直是一个老大难问题，但是在Ubuntu系统上就不会出现Windows的各种报错，所以很简单常用的操作，打开终端：

```
source activate deep    #(激活anoconda安装的python环境)
pip intall scrapy       #安装Scrapy框架
```

二、创建一个工程和基本配置

###（1）Scrapy安装好后，已经可以在终端调用，所以想要进行工程创建，只需要打开终端：

```
cd 到想要的路径      #进入指定文件夹
scrapy startproject Industry_Get    #(创建工程，工程名可以任意起)
```



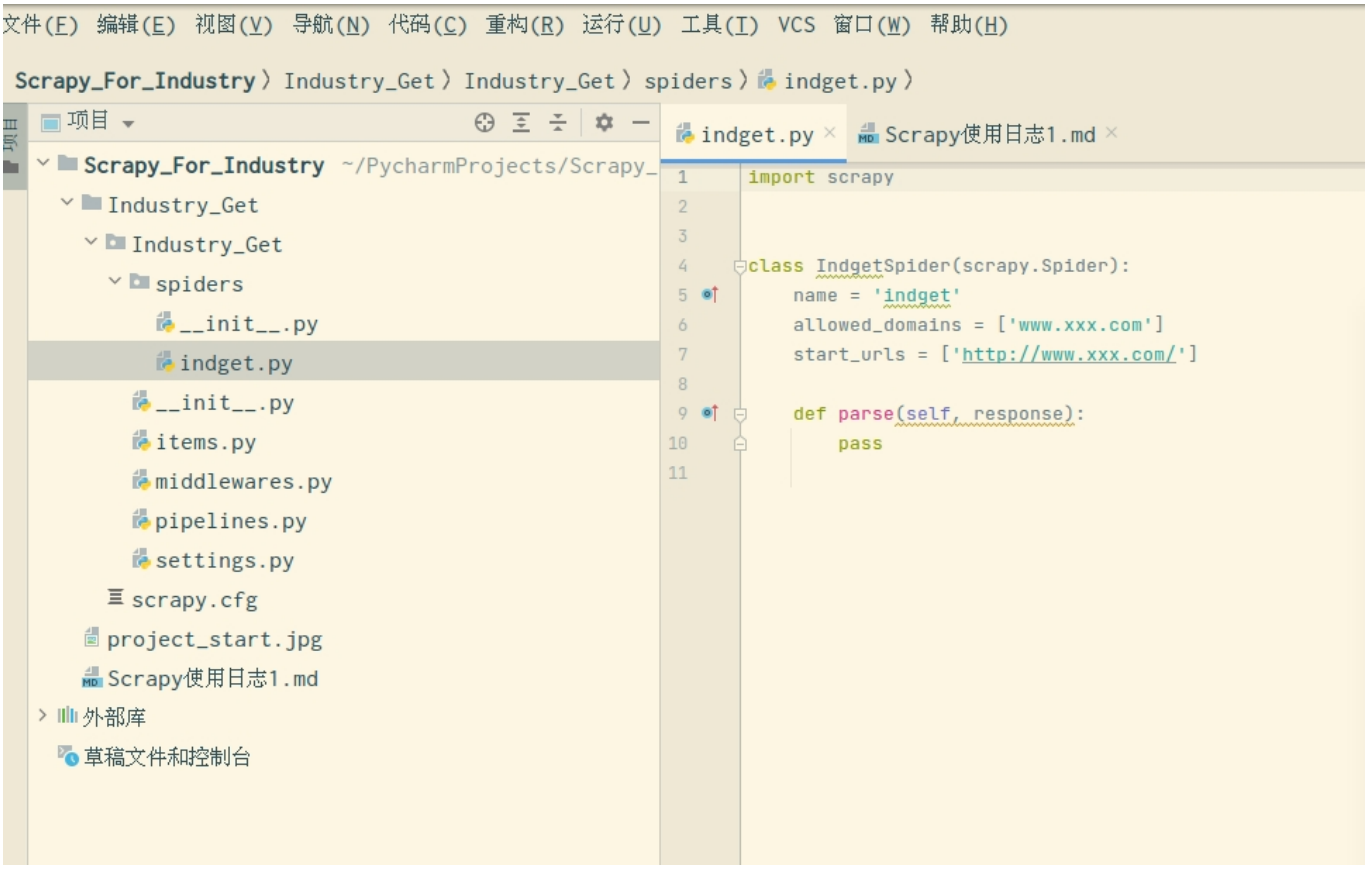
如图所示，工程已经创建完毕。 需要说明的是，笔者一般都是直接通过Pychram打开终端，这是已经在指定文件夹下，所以不用cd操作即可直接进行工程创建命令。

(2)创建一个爬虫文件：

在创建完工程后，我们需要创建一个爬虫文件，爬虫文件需要建立在spiders子目录里，但是只需要cd入工程目录即可，在此目录下运行终端命令，终端命令会自动找到spiders子目录并且创建爬虫文件。

```
cd Industry_Get      #进入Scrapy工程目录
scrapy genspider indget www.xxx.com     #scrapy genspider 爬虫名 起始爬取网址（www.xxx.com），在这一步后我们就已经成功创建爬虫文件
```

在这一步后，我们就已经创建好了爬虫文件：



(3)进行某些设置：

打开settings.py文件，对爬虫进行配置：
首先找到ROBOTSTXT_OBEY，把他的值改成False。意思是我们的爬虫不再遵守机器协议：

```
ROBOTSTXT_OBEY=False
```

然后在本文件的任意位置，添加一行代码：

```
LOG_LEVEL='ERROR'
```

本代码的作用是在程序没有产生错误的时候，不得输出日志，以免让日志淹没我们的爬虫信息。
然后要修改user-agent:

```
USER_AGENT = 'Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:88.0)
Gecko/20100101 Firefox/88.0'
```

user-agent不一定要用我这个，可以使用自己的，至于怎么找自己的，建议网上搜。

三、进行数据爬取

这一步是最关键的，首先看一下我的爬取任务：

(1)任务综述和初步思路

网址：
https://www.chyxx.com/。
任务：爬取这个网站中所有行业的报告，包括报告的详细信息。 以能源动力行业为例子：
https://www.chyxx.com/industry/nengyuan/

观察得出，每一条行业研究报告的网址没有规律（此网站应该是每天固定更新所有行业的报告，并且统一排序后才进行分行业）。
所以我们应该首先想办法拿到每个行业的网址，从而进行抓爬。

(2)编辑爬虫文件：

这一步是最重要的，在之前已经创建了indget.py文件（位于spiders文件夹下），编辑此文件，设置爬虫。
爬虫文件的默认的程序为：

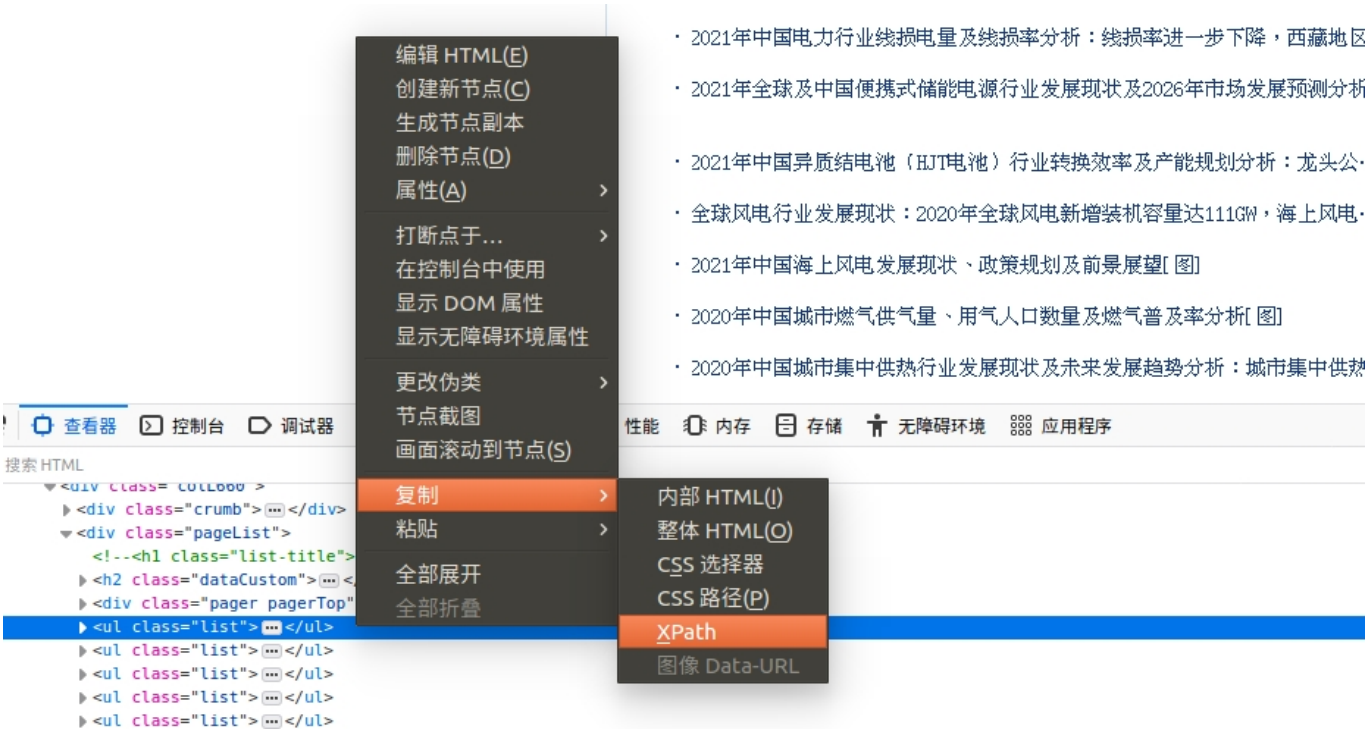
```
import scrapy   #导入包


class IndgetSpider(scrapy.Spider):
    name = 'indget'      #爬虫名，作为本爬虫文件的唯一标示，不可更改。
    allowed_domains = ['www.xxx.com']    #限定网页，一般可以注释掉，用来限制
start_urls里哪些网页可以爬，哪些不可以。
    start_urls = ['http://www.xxx.com/']     #要爬取的网页列表
```

```
    def parse(self, response):  #爬虫每爬取一个url，都会把网页的信息作为response
返回给parse函数，所以应当针对parse函数进行修改以便解析网页信息。
        pass
```

首先观察一下起始url所对应的页面，推荐使用Xpath进行解析，（因为火狐浏览器可以通过检查选项返回指定板块的xpath路径解析式，不再需要我自己行再寻找）。

如图：



此页面中所有的行业采用ul标签中嵌套五个li标签的形式，li标签中包含了需要的子url信息以及日期和行研标题信息。 在此需要得到所有行研报告的标题，指向url，以及日期。

且不管后续的页面，在此只先处理一下第一个页面的信息，得到研报发布的时间，标题，以及子链接。

改写过后的parse函数如下：

```
    def parse(self, response):
        ul_list=response.xpath('/html/body/div[4]/div/div[1]/div[2]/ul')
#获取当前页面中所有的ul标签，并且制成一个列表存储起来。
        for ul in ul_list:
            li_list=ul.xpath('./li')
            for li in li_list:
                href=li.xpath('./a/@href').get()    #获取a标签中href属性中蕴含
的子链接信息，并且把其转化为字符串。
                title=li.xpath('./a/text()').extract()[0]   #获取a标签的文本
信息，即标题。
                date = li.xpath('./span/text()').extract()[0]   #或许span标
签的文本信息，即发布时间。
                print(href,title,date)
                print(type(href))
            break
```

在此步骤之后，可以得到每个研报的所在链接，在此需要对每个研报所在的网页进行抓取和解析，在此步骤中，需要执行请求传参和递归。解析完后使用pipeline进行保存文件。
爬虫文件的解析方法：

```
    def parse(self, response):
        time.sleep(1)
        ul_list=response.xpath('/html/body/div[4]/div/div[1]/div[2]/ul')
#获取当前页面中所有的ul标签，并且制成一个列表存储起来。
        if self.page_num==2:
            '''
            获取当前行业一共有多少页新闻，并且把值赋给all_page_count变量
            '''

self.all_page_count=int(response.xpath('/html/body/div[4]/div/div[1]/div[2]
/div[2]/a')[-2].xpath('./text()').extract()[0])
        for ul in ul_list:
            '''这个循环可以获取当前页面下左右新闻的标题，发布日期，链接地址'''
            li_list=ul.xpath('./li')
            for li in li_list:
                item=IndustryGetItem()
                href=li.xpath('./a/@href').get()    #获取a标签中href属性中蕴含
的子链接信息，并且把其转化为字符串。
                title=li.xpath('./a/text()').extract()[0]   #获取a标签的文本
信息，即标题。
                date = li.xpath('./span/text()').extract()[0]   #或许span标
签的文本信息，即发布时间。
                detail_href=self.yanbao_all_url+href
                item['information_title']=title
                item['information_date']=date
                #对详情页发请求，获取行研的源码数据。
                #使用手动请求发送即可
                yield
scrapy.Request(detail_href,callback=self.parse_detail,meta={'item':item})
#请求传参，通过meta带一个字典，就可以把meta字典传递给请求对应的回调函数。
```

```
        if self.page_num <= self.all_page_count:
            next_url=format(self.all_use_url%self.page_num)     #更换为新的
url
            print("当前为第"+str(self.page_num)+"页")
            self.page_num +=1       #page_num加1，切换为下一页的url
            yield scrapy.Request(url=next_url,callback=self.parse)  #递归调
用parse方法
```

(3)使用管道进行保存

(1)进行数据解析。
(2)在item类中定义相关的属性。
(3)将解析的数据存储到item类型的对象中(参考items.py)，给QiubaiproItem类添加属性，以方便实例化之后进行存储:
item=QiubaiproItem() item['author']=author item['content']=content
yield item # 将item提交给了管道 (4)将item类型的对象提交给管道(pipeline.py)进行持久化存储操作，
process_item方法是用来处理item对象（此时数据已经存到了item对象中）的方法，在这个方法中要把接收到的item对象中存储的数据进行持久化存储操作。
该方法可以接收爬虫文件提交过来的item对象，该方法每接收到一次item对象就会被调用一次。
(5)如果想使用开启一次的方法，可以考虑重写父类方法的open_spider()和close_spider()。
(6)在配置文件中开启管道(setting.py):
ITEM_PIPELINES = { 'qiubaipro.pipelines.QiubaiproPipeline': 300, }
300表示的是优先级，数值越小，优先级越高。

四、ip掩盖

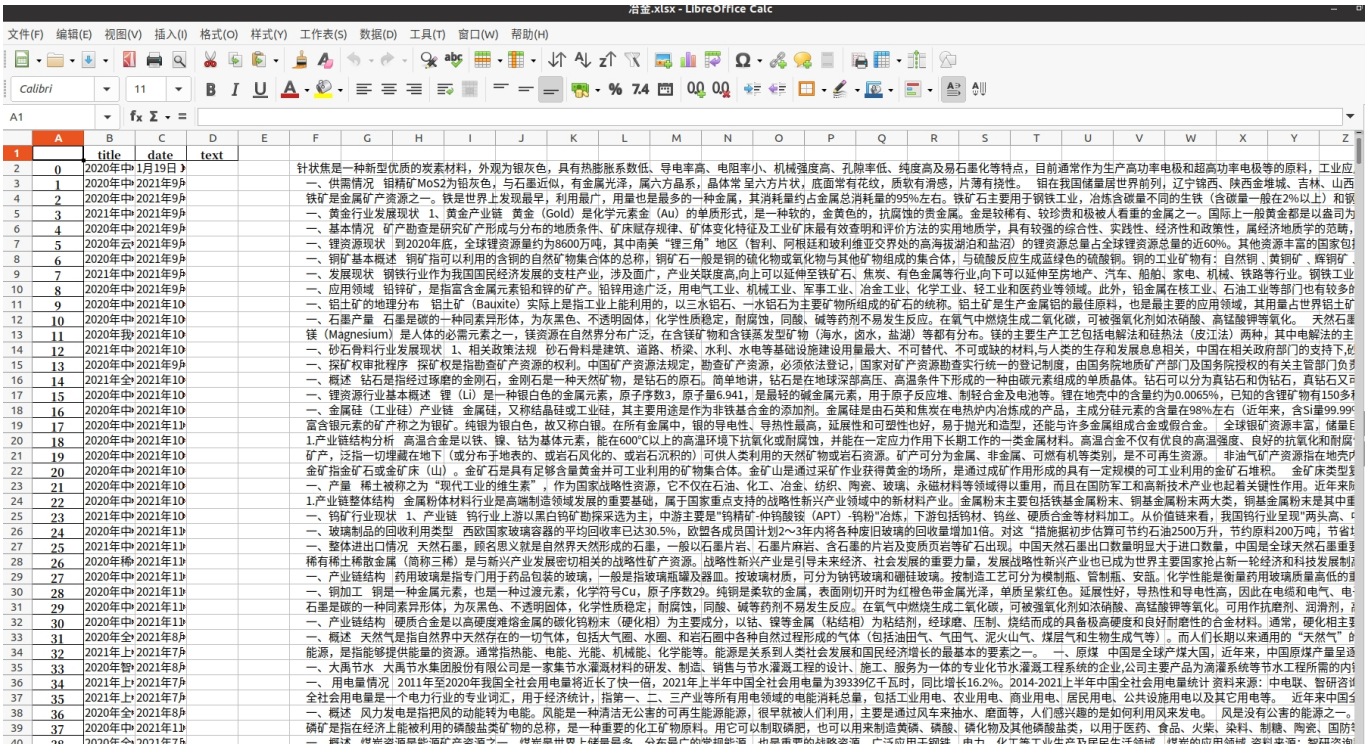
最后是要进行挂ip池，如果不挂ip池，大概率因为段时间内爬取过多被封禁。
掩盖ip需要购买ip，然后进行中间件操作：
middlewares.py即是我们的中间件文件。
(1)删去爬虫中间件,SpiderMiddleware类)，重点在于下载中间件(DownloaderMiddleware)类。
(2)重点看其中的三个方法：
删去from_crawler，spider_opened方法。
process_request:拦截并且处理请求的方法。
process_response：拦截并且处理所有的响应的方法。
process_exception：拦截所有的异常。
进行UA伪装：
进行ip代理：
在setting.py里开启中间件：

```
# Enable or disable downloader middlewares
# See https://docs.scrapy.org/en/latest/topics/downloader-middleware.html
DOWNLOADER_MIDDLEWARES = {
   'Ip_Forget.middlewares.IpForgetDownloaderMiddleware': 543,
}
```

在这里因为ip池收费的原因，所以只提供一个UA池：

```
    user_agents_list = [
        'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/39.0.2171.95 Safari/537.36 OPR/26.0.1656.60',
        'Opera/8.0 (Windows NT 5.1; U; en)',
        'Mozilla/5.0 (Windows NT 5.1; U; en; rv:1.8.1) Gecko/20061208 Firefox/2.0.0 Opera 9.50',
        'Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; en) Opera 9.50',
        'Mozilla/5.0 (Windows NT 6.1; WOW64; rv:34.0) Gecko/20100101 Firefox/34.0',
        'Mozilla/5.0 (X11; U; Linux x86_64; zh-CN; rv:1.9.2.10) Gecko/20100922 Ubuntu/10.10 (maverick) Firefox/3.6.10',
        'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/534.57.2 (KHTML, like Gecko) Version/5.1.7 Safari/534.57.2 ',
        'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/39.0.2171.71 Safari/537.36',
        'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.11 (KHTML, like Gecko) Chrome/23.0.1271.64 Safari/537.11',
        'Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US) AppleWebKit/534.16 (KHTML, like Gecko) Chrome/10.0.648.133 Safari/534.16',
        'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599.101 Safari/537.36',
        'Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko',
        'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/536.11 (KHTML, like Gecko) Chrome/20.0.1132.11 TaoBrowser/2.0 Safari/536.11',
        'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.1 (KHTML, like Gecko) Chrome/21.0.1180.71 Safari/537.1 LBBROWSER',
        'Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; QQDownload 732; .NET4.0C; .NET4.0E)',
        'Mozilla/5.0 (Windows NT 5.1) AppleWebKit/535.11 (KHTML, like Gecko) Chrome/17.0.963.84 Safari/535.11 SE 2.X MetaSr 1.0',
        'Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0; SV1; QQDownload 732; .NET4.0C; .NET4.0E; SE 2.X MetaSr 1.0) ',
    ]
```

五、运行爬虫，即可得到结果。



	A	B	C	D	E	F
1		title	date	text		
2	0	2020年中	1月19日			针状焦是一种新型优质的炭素材料，外观为银灰色，具有热膨胀系数低、导电率高、电阻率小、机械强度高、孔隙率低、纯度高及易石墨化等特点，目前通常作为生产高功率电极和超高功率电极等的原料，工业应
3	1	2020年中	2021年9月			一、供需情况 钼精矿MoS2为铅灰色，与石墨近似，有金属光泽，属六方晶系，晶体常 呈六方片状，底面常有花纹，质软有滑感，片薄有挠性。 钼在我国储量居世界前列，辽宁锦西、陕西金堆城、吉林、山西
4	2	2020年中	2021年9月			铁矿是金属矿产资源之一。铁是世界上发现最早，利用最广，用量也是最多的一种金属，其消耗量约占金属总消耗量的95%左右。铁矿石主要用于钢铁工业，冶炼含碳量不同的生铁（含碳量一般在2%以上）和钢
5	3	2021年中	2021年9月			一、黄金行业发展现状 1、黄金产业链 黄金（Gold）是化学元素金（Au）的单质形式，是一种软的，金黄色的，抗腐蚀的贵金属。金是较稀有、较珍贵和极被人看重的金属之一。国际上一般黄金都是以盎司为
6	4	2020年中	2021年9月			一、基本情况 矿产勘查是研究矿产形成与分布的地质条件、矿床赋存规律、矿体变化特征及工业矿床最有效查明和评价方法的实用地质学，具有较强的综合性、实践性、经济性和政策性，属经济地质学的范畴，
7	5	2020年中	2021年9月			一、锂资源现状 到2020年底，全球锂资源量约为8600万吨，其中南美“锂三角”地区（智利、阿根廷和玻利维亚交界处的高海拔湖泊和盐沼）的锂资源总量占全球锂资源总量的近60%。其他资源丰富的国家包括
8	6	2020年中	2021年9月			一、铜矿基本概述 铜矿指可以利用的含铜的自然矿物集合体的总称，铜矿石一般是铜的硫化物或氧化物与其他矿物组成的集合体，与硫酸反应生成蓝绿色的硫酸铜。铜的工业矿物有：自然铜 、黄铜矿 、辉铜矿
9	7	2021年中	2021年9月			一、发展现状 钢铁行业作为我国国民经济发展的支柱产业，涉及面广，产业关联度高,向上可以延伸至铁矿石、焦炭、有色金属等行业,向下可以延伸至房地产、汽车、船舶、家电、机械、铁路等行业。钢铁工业
10	8	2020年中	2021年9月			一、应用领域 铅锌矿，是指富含金属元素铅和锌的矿产。铅锌用途广泛，用电气工业、机械工业、军事工业、冶金工业、化学工业、轻工业和医药业等领域。此外，铅金属在核工业、石油工业等部门也有较多的
11	9	2020年中	2021年10			一、铝土矿的地理分布 铝土矿（Bauxite）实际上是指工业上能利用的，以三水铝石、一水铝石为主要矿物所组成的矿石的统称。铝土矿是生产金属铝的最佳原料，也是最主要的应用领域，其用量占世界铝土矿
12	10	2020年中	2021年10			一、石墨产量 石墨是碳的一种同素异形体，为灰黑色、不透明固体，化学性质稳定，耐腐蚀，同酸、碱等药剂不易发生反应。在氧气中燃烧生成二氧化碳，可被强氧化剂如浓硝酸、高锰酸钾等氧化。 天然石墨
13	11	2020年我	2021年10			镁（Magnesium）是人体的必需元素之一，镁资源在自然界分布广泛，在含镁矿物和含镁蒸发型矿物（海水，卤水，盐湖）等都有分布。镁的主要生产工艺包括电解法和硅热法（皮江法）两种，其中电解法的主
14	12	2021年中	2021年10			一、砂石骨料行业发展现状 1、相关政策法规 砂石骨料是建筑、道路、桥梁、水利、水电等基础设施建设用量最大、不可替代、不可或缺的材料,与人类的生存和发展息息相关，中国在相关政府部门的支持下,砂
15	13	2020年中	2021年9月			一、探矿权审批程序 探矿权是指勘查矿产资源的权利。中国矿产资源法规定，勘查矿产资源，必须依法登记，国家对矿产资源勘查实行统一的登记制度，由国务院地质矿产部门及国务院授权的有关主管部门负责
16	14	2021年全	2021年10			一、概述 钻石是指经过琢磨的金刚石，金刚石是一种天然矿物，是钻石的原石。简单地讲，钻石是在地球深部高压、高温条件下形成的一种由碳元素组成的单质晶体。钻石可以分为真钻石和伪钻石，真钻石又可
17	15	2020年中	2021年10			一、锂资源行业基本概述 锂（Li）是一种银白色的金属元素，原子序数3，原子量6.941，是最轻的碱金属元素，用于原子反应堆、制轻合金及电池等。锂在地壳中的含量约为0.0065%，已知的含锂矿物有150多种
18	16	2020年中	2021年10			一、金属硅（工业硅）产业链 金属硅，又称结晶硅或工业硅，其主要用途是作为非铁基合金的添加剂。金属硅是由石英和焦炭在电热炉内冶炼成的产品，主成分硅元素的含量在98%左右（近年来，含Si量99.99
19	17	2020年中	2021年11			富含银元素的矿产称之为银矿。纯银为银白色，故又称白银。在所有金属中，银的导电性、导热性最高，延展性和可塑性也好，易于抛光和造型，还能与许多金属组成合金或假合金。 全球银矿资源丰富，储量巨
20	18	2020年中	2021年10			1.产业链结构分析 高温合金是以铁、镍、钴为基体元素，能在600℃以上的高温环境下抗氧化或耐腐蚀，并能在一定应力作用下长期工作的一类金属材料。高温合金不仅有优良的高温强度、良好的抗氧化和耐腐
21	19	2020年中	2021年10			矿产，泛指一切埋藏在地下（或分布于地表的、或岩石风化的、或岩石沉积的）可供人类利用的天然矿物或岩石资源。矿产可分为金属、非金属、可燃有机等类别，是不可再生资源。 非油气矿产资源指在地壳内
22	20	2020年中	2021年10			金矿指金矿石或金矿床（山）。金矿石是具有足够含量黄金并可工业利用的矿物集合体。金矿山是通过采矿作业获得黄金的场所，是通过成矿作用形成的具有一定规模的可工业利用的金矿石堆积。 金矿床类型复
23	21	2020年中	2021年10			一、产量 稀土被称之为“现代工业的维生素”，作为国家战略性资源，它不仅在石油、化工、冶金、纺织、陶瓷、玻璃、永磁材料等领域得以重用，而且在国防军工和高新技术产业也起着关键性作用。近年来随
24	22	2020年中	2021年10			1.产业链整体结构 金属粉体材料行业是高端制造领域发展的重要基础，属于国家重点支持的战略性新兴产业领域中的新材料产业。金属粉末主要包括铁基金属粉末、铜基金属粉末两大类，铜基金属粉末是其中重
25	23	2021年中	2021年10			一、钨矿行业现状 1、产业链 钨行业上游以黑白钨矿勘探采选为主，中游主要是"钨精矿-仲钨酸铵（APT）-钨粉"冶炼，下游包括钨材、钨丝、硬质合金等材料加工。从价值链来看，我国钨行业呈现"两头高、中
26	24	2020年中	2021年11			一、玻璃制品的回收利用类型 西欧国家玻璃容器的平均回收率已达30.5%，欧盟各成员国计划2～3年内将各种废旧玻璃的回收量增加1倍。对这“措施据初步估算可节约石油2500万升，节约原料200万吨，节省
27	25	2021年中	2021年11			一、整体进出口情况 天然石墨，顾名思义就是自然界天然形成的石墨，一般以石墨片岩、石墨片麻岩、含石墨的片岩及变质页岩等矿石出现。中国天然石墨出口数量明显大于进口数量，中国是全球天然石墨重要
28	26	2020年稀	2021年11			稀有稀土稀散金属（简称三稀）是与新兴产业发展密切相关的战略性矿产资源。战略性新兴产业是引导未来经济、社会发展的重要力量，发展战略性新兴产业也已成为世界主要国家抢占新一轮经济和科技发展制高
29	27	2020年中	2021年11			一、产业链结构 药用玻璃是指专门用于药品包装的玻璃，一般是指玻璃瓶罐及器皿。按玻璃材质，可分为钠钙玻璃和硼硅玻璃。按制造工艺可分为模制瓶、管制瓶、安瓿。化学性能是衡量药用玻璃质量高低的重
30	28	2020年中	2021年11			一、铜加工 铜是一种金属元素，也是一种过渡元素，化学符号Cu，原子序数29。纯铜是柔软的金属，表面刚切开时为红橙色带金属光泽，单质呈紫红色。延展性好，导热性和导电性高，因此在电缆和电气、电
31	29	2020年中	2021年11			石墨是碳的一种同素异形体，为灰黑色、不透明固体，化学性质稳定，耐腐蚀，同酸、碱等药剂不易发生反应。在氧气中燃烧生成二氧化碳，可被强氧化剂如浓硝酸、高锰酸钾等氧化。可用作抗磨剂、润滑剂，
32	30	2020年中	2021年11			一、产业链结构 硬质合金是以高硬度难熔金属的碳化钨粉末（硬化相）为主要成分，以钴、镍等金属（粘结相）为粘结剂，经球磨、压制、烧结而成的具备极高硬度和良好耐磨性的合金材料。通常，硬化相主要
33	31	2020年全	2021年8月			一、概述 天然气是指自然界中天然存在的一切气体，包括大气圈、水圈、和岩石圈中各种自然过程形成的气体（包括油田气、气田气、泥火山气、煤层气和生物生成气等）。而人们长期以来通用的“天然气”的
34	32	2021年上	2021年7月			能源，是指能够提供能量的资源。通常指热能、电能、光能、机械能、化学能等。能源是关系到人类社会发展和国民经济增长的最基本的要素之一。 一、原煤 中国是全球产煤大国，近年来，中国原煤产量呈逐
35	33	2020年智	2021年8月			一、大禹节水 大禹节水集团股份有限公司是一家集节水灌溉材料的研发、制造、销售与节水灌溉工程的设计、施工、服务为一体的专业化节水灌溉工程系统的企业,公司主要产品为滴灌系统等节水工程所需的内销
36	34	2021年上	2021年7月			一、 用电量情况 2011年至2020年我国全社会用电量将近长了快一倍，2021年上半年中国全社会用电量为39339亿千瓦时，同比增长16.2%。2014-2021上半年中国全社会用电量统计 资料来源：中电联、智研咨询
37	35	2021年上	2021年7月			全社会用电量是一个电力行业的专业词汇，用于经济统计，指第一、二、三产业等所有用电领域的电能消耗总量，包括工业用电、农业用电、商业用电、居民用电、公共设施用电以及其它用电等。 近年来中国全
38	36	2020年全	2021年8月			一、概述 风力发电是指把风的动能转为电能。风能是一种清洁无公害的可再生能源能源，很早就被人们利用，主要是通过风车来抽水、磨面等，人们感兴趣的是如何利用风来发电。 风是没有公害的能源之一。
39	37	2020年中	2021年11			磷矿是指在经济上能被利用的磷酸盐类矿物的总称，是一种重要的化工矿物原料。用它可以制取磷肥，也可以用来制造黄磷、磷酸、磷化物及其他磷酸盐类，以用于医药、食品、火柴、染料、制糖、陶瓷、国防等
40	38	2020年全	2021年7月			一、概述 煤炭资源是能源矿产资源之一。煤炭是世界上储量最多、分布最广的常规能源，也是重要的战略资源，广泛应用于钢铁、电力、化工等工业生产及民民生活领域。 煤炭的应用领域 资料来源：智研咨询