煤炭价格爬虫：以金投网为例

url = “https://cngold.org/meitan/list\_112\_all.html”

爬虫流程：准备工作、获取数据、解析内容、保存数据

**Spider for coal infomation：**

1. 先在url中提取2023年的煤炭咨询列表，定位得到2023煤炭价格列表集：**div class=”history\_news\_content”/ul/li**

2. 对于每一个网址，定位得到每天煤炭价格咨询的精确地址：

**根据div class=”border\_top”/a@herf**

3. 打开煤炭价格页面，定位产地、品种、价格和时间等信息：

**tbody/tr/th**

4. 将从上述页面得到的煤炭信息存到csv文件中

寻找额外的煤炭及相关产品的信息网站：

1. 中国煤炭市场网：登陆后无法查看，无会员权限。

2. 锐思数据：数据比较古老，且只能查看两年内的数据，没有最近3年内的数据。

3. 国际能源网：查询煤炭网需要升级1500铜牌会员。

4. 中国煤炭经济网：煤炭价格指数按周记，但已停发两个月。

5. K线交流平台：根据煤炭行情走势进行分析预测，结果可信度未知。

焦炭数据爬取

6. 焦煤在线：[焦煤在线 (sxccol.com)](https://www.sxccol.com/sjtj/jtzs/index.htm?typ=1)

onclick=”refresh(‘413’);” 根据黄色部分定位网址，

<tbody/tr/th> 标题行：指数名称/本期/上期/周涨跌/月涨跌 共5列，指数单位：点

<tbody/tr#山西焦炭综合.table2/th> 山西焦炭价格总指数 D1 D2 D3 D4 记4列。数据周更，量太少，已弃。

7. 焦炭数据金投网爬取

定位网址：[焦炭期货-金投期货-金投网 (cngold.org)](https://futures.cngold.org/jiaotan/list_2533_all.html) 从四月开始记录数据；

打开一个日期，进入该日焦炭期货咨讯；**<li/a@href=(.\*?) a@title=”今日焦炭(.\*?)”>** ，然后抛弃爬取的第一个网址

打开此链接，定位<**tbody/tr/span.time>、<table/tbody/tr/td/span.quotes>**

数据库名称：焦炭主力期货价格查询

日期、开盘价、昨收价

问题1：AttributeError: 'NoneType' object has no attribute 'replace'

原因：爬取网页后，解析数据得到的是空列表

分析：requests获取到的html.text为空字符串；所以下面抛出NoneType异常。原因是请求get需要增加headers来解决反扒；模拟浏览器请求来获取数据；开发者页面：networks——headers

1. 查找喷吹煤价格指数：我的钢铁网站有详细资料，但无权限，需要开会员。

2. 查找无烟煤价格数据：野天鹅网站，以月为单位更新，持续到今年二月份；我的钢铁myteel.com有资料，但也有限制；隆众资讯：要开会员。

3. 煤炭数据省市对应：

3.1 使用geopy自动对城市生成对应的省，但常年不维护，导致爬取失败，baidu.py的api\_path版本从v2改成v3也爬取失败，不可用。

3.2. jieba库安装，下载全国行政区划分词库文件，根据此文件进行分词匹配得到省名，繁琐。

3.3 查询煤炭数据中产地对应的省份名称共24条数据/天，根据数据爬取的天数决定上述省份数据的复制次数，并将煤炭数据与该数据进行拼接操作。

3.4 金投网煤炭信息抓取滞后，例如2023年5月9日煤炭信息在13点13分38秒时于该网站发表，但时隔一小时还未在“2023年煤炭价格资讯列表”中找到，也未在“煤炭价格”中看到。





