

# 爬虫任务报告

## 爬虫任务

十堰问政 <http://m.shiyan.gov.cn/zwhd/web/webindex.action>

## 任务团队

粉红橘子：范思棋、冯淑杰、高乐天

## 数据概况

从2007.10.22--2021.11.26 共十二万条

## 任务完成情况

- 成功爬取十二万条问政记录，存入result.xlsx中
- 更新函数每次更新可以做到将新产生的数据存入原result.xlsx中，并产生一个带有时间的副本文件，类似于，newdataNov 17 17-25-35 2021.xlsx

## 数据说明

- 关于问政信息分别爬取了：编号、提问人、类型、浏览次数、受理单位、标题、提问内容、办结回复、回复时间、办理部门

Q120072															
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	0	编号	提问人	类型	提问时间	浏览次数	受理单位	标题	提问内容	办结回复	回复时间	办理部门			
2	0	332138	青蛙	咨询	2021-11-26 15:18:236		市医保局	网上医保	请问建行手机银行怎	您好！您所反映的	2021-11-26 15:50:43.	C市医疗保障服务中心			
3	0	332117	ly1987	咨询	2021-11-26 11:13:245		市卫生健康委员会	什么时候实行适	请问十堰地区什么时	经了解，十堰市自	2021-11-26 16:28:55.	C市卫生健康委员会			
4	0	332089	笑下去	咨询	2021-11-25 22:22:180		市房地产服务中心	预售证问题	惠泽春风里预售迟迟	您好，企业《施工	2021-11-26 14:16:12.	C市审批科			
5	0	332086	ykm	其他	2021-11-25 21:18:267		市自然资源和规划	表扬不动产登记	表扬不动产登记	感谢您肯定与鼓	2021-11-26 14:43:25.	C市自然资源和规划局			
6	0	332078	云夕	咨询	2021-11-25 17:22:315		郧西县	福银高速郧西路	福银高速郧西收费站	您好，您所提及的	2021-11-26 15:16:16.	C郧西 交通局			
7	0	332068	李刚12345	投诉	2021-11-25 15:50:227		市公安局	东风阳光城4期5	由于东风阳光城4期小	感谢您对交通管理	2021-11-26 15:33:07.	C市交管局办公室			
8	0	332065	lisa61904	求助	2021-11-25 14:38:348		市市场监督管理局	美善小区暖气费	张湾区艳湖社区的	美善依据《湖北省物	2021-11-25 16:19:49.	C市市场监督管理局			
9	0	332063	十堰一路	求助	2021-11-25 14:13:316		市市场监督管理局	上海名都物业乱	上海名都物业乱收	费，	问政人2021-11-25 15:21:03.	C市市场监督管理局			
10	0	332061	神龙	咨询	2021-11-25 13:56:290		竹溪县	两病补贴何时到	医保局所说的两病补	神龙先生您好：关	2021-11-26 11:13:55.	C竹溪县 医保局			
11	0	332050	问政007	求助	2021-11-25 11:20:259		市公安局	外地驾照转入十	请问一下，外地驾	照转回十堰，已	2021-11-25 12:09:44.	C十堰经济技术开发区			
12	0	332037	逸666	求助	2021-11-25 10:25:266		十堰经济技术开发区	用电需求问题	领导好！我是茂达花	园区域改办回复：已	2021-11-25 10:23:36.	C市交管局车管所			
13	0	332033	www123	咨询	2021-11-25 10:06:231		市人力资源和社会保障	一级残疾无劳动	领导们	你好 您好，经与您	电话2021-11-25 17:48:49.	C市人力资源和社会保障局			
14	0	332015	荷风呈祥	求助	2021-11-24 20:02:313		市房地产服务中心	翰林世家小区上	今年寒冬季	您好！我市不属于	2021-11-25 11:19:27.	C市房地产服务中心			
15	0	332007	张生2	咨询	2021-11-24 17:14:164		市社会保险局	新开单位社保账	公司新成立，现在想	工作人员与当事人	2021-11-26 10:04:14.	C市社会保险局			
16	0	332006	断了线的	批评	2021-11-24 16:52:477		市医保局	凭什么指定建行	最近连续两天就为	了您好！您所反映的	2021-11-25 10:33:38.	C市医疗保障服务中心			
早期数据中受理单位办理部门数据不全，并不是爬取问题，是网站本身数据不全。															
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
89723	0	169650	gongqiang528	咨询	2015-12-11 17:24:11.0	578	市公交公司	关于机场	看了贵公	您好！对您的	2015-12-12 11:03:24.				
89724	0	169649	小炮啊啊啊	求助	2015-12-11 17:20:14.0	745	市自然资源和规划局	北京小镇	你好，北	江您好！来信收	2015-12-13 18:06:47.0				
89725	0	169646	1819522901	求助	2015-12-11 16:19:49.0	518	市公路养护中心	现在316国	现在在316国	您好，减速带	2015-12-18 10:43:16.0				
89726	0	169645	luchen880828	咨询	2015-12-11 16:18:25.0	612	市公交公司	机场公交	方案1：红	您好！对您的	2015-12-12 10:35:34.0				
89727	0	169644	玻璃杯	批评	2015-12-11 16:17:08.0	478	十堰移动公司	信号弱，	林湖二	号尊敬的用户您	2015-12-20 14:23:20.0				
89728	0	169643	yunyangshui	咨询	2015-12-11 15:55:27.0	526	市公安局	红绿灯	浙江路	您好，我单位	2015-12-11 16:56:23.0				
89729	0	169641	1226505795	批评	2015-12-11 15:29:14.0	631	郧阳区	面子工程	五峰乡政	您好！感谢	2015-12-14 16:31:19.0				
89730	0	169640	346778839	求助	2015-12-11 15:21:44.0	1038	市发展和改革委员会	十堰高速	2015年7月	您好，感谢	2015-12-16 10:51:14.0				
89731	0	169639	白秉锋	投诉	2015-12-11 14:21:45.0	366	市房地产服务中心	房产证办	我是天津	经调查核实，	2015-12-14 10:33:24.0				
89732	0	169638	韩克利	批评	2015-12-11 14:19:22.0	465	郧西县	办理分户	我是郧西	根据县有关文	2015-12-28 11:51:13.0				
89733	0	169636	舒克。	求助	2015-12-11 14:03:25.0	1744	市城投公司	火车站北	在以前的	南广场及铁路	2015-12-11 16:42:34.0				
89734	0	169635	舒克。	求助	2015-12-11 13:55:58.0	102203		增加站台	前段时间	您好，十堰站	2016-01-07 12:38:29.0				
89735	0	169634	舒克。	求助	2015-12-11 13:46:45.0	1323	市城投公司	火车站北	请问之前	火车站北广场	2015-12-11 16:41:42.0				
89736	0	169633	舒克。	求助	2015-12-11 13:36:11.0	606	市公交公司	公交编号	请问贵公	您好！目	2015-12-12 11:35:10.0				
89737	0	169632	我！我！我！	求助	2015-12-11 13:06:26.0	1038	市教育局	2016年十	月我想问	“2016年市直	2015-12-11 16:24:53.0				
89738	0	169631	陈志明	其他	2015-12-11 12:46:26.0	709	市公交公司	拟开通的	市公交	您好！我	2015-12-15 11:05:05.0				
89739	0	169630	7758521	咨询	2015-12-11 12:01:10.0	456	市公交公司	关于机场	看了贵公	您好！对您的	2015-12-11 16:47:16.0				
89740	0	169629	245966511	批评	2015-12-11 11:55:57.0	468	市房地产服务中心	东方国际	我来去年	您好，你反映	2015-12-14 10:33:38.0				
89741	0	169628	吉卜	求助	2015-12-11 11:55:52.0	1173	市医保局	职工医保	职工医	保您好！您这种	2015-12-14 11:09:00.0				
89742	0	169626	8785183	建议	2015-12-11 11:26:29.0	1831	十堰联通公司	铁塔公司	我在五堰	尊敬的用户您	2015-12-15 15:47:22.0				
89743	0	169625	CCXZ	批评	2015-12-11 11:06:01.0	486	市规划局	柳中园	一轻核装	：1169625号网	2015-12-11 16:11:29.0				
89744	0	169624	8785183	建议	2015-12-11 11:02:05.0	737	市卫生健康委员会	退休职工	我们是上	您好！依据《	2015-12-11 16:18:47.0				
89745	0	169622	香草	求助	2015-12-11 10:24:32.0	444	市房地产服务中心	房产登记	您好，我	你好，现不动	2015-12-15 09:50:03.0				
89746	0	169621	一粒尘	咨询	2015-12-11 10:21:13.0	665	武当山旅游经济特区	关于评选	针对武当	你好！你反映	2015-12-14 09:05:46.0				
89747	0	169620	18772200405	求助	2015-12-11 10:13:42.0	483	十堰移动公司	和昌国际	和昌国际	尊敬的客户：	2015-12-20 14:23:46.0				
89748	0	169619	18772200405	咨询	2015-12-11 09:57:49.0	840	市公交公司	S5路公交	喜闻公	交您好！S5路公	2015-12-12 11:02:12.0				
89749	0	169618	晖晖	批评	2015-12-11 09:48:12.0	457	市房地产服务中心	书香阁物	这个所谓	您好，你反映	2015-12-14 10:31:32.0				
89750	0	169616	18772200405	咨询	2015-12-11 09:39:56.0	601	市公交公司	机场2号	线建议机	场您好！对您的	2015-12-11 16:44:43.0				
89751	0	169614	asdg123	批评	2015-12-11 09:16:39.0	599	房县	领导因	人一那姓	您好！由于您	2015-12-14 11:33:04.0				
89752	0	169613	svsybsbsb	批评	2015-12-11 07:56:28.0	595	市人力资源和社会保障	局员工入	职你好，我	你好！请到你	2015-12-21 16:25:28.0				
89753	0	169611	撒帆撒	求助	2015-12-10 22:53:12.0	557	市住房公积金中心	公积金可	以国家发	布尊敬	的网友，	2015-12-11 16:35:32.0			
89754	0	169610	太美丽邵阳	投诉	2015-12-10 22:39:56.0	774	郧阳区	请向下	身为郧阳	郧阳区图书馆	2015-12-14 16:29:41.0				
89755	0	169609	cydbf	批评	2015-12-10 22:05:02.0	438	丹江口市	严重拖欠	我们有一	群Cydbf同	志：2015-12-14 15:49:23.0				
89756	0	169607	豆豆果果	投诉	2015-12-10 21:47:50.0	814	武当山旅游经济特区	情况反映	我期望	您好！你反映	2015-12-14 15:07:46.0				
89757	0	169605	961259055	求助	2015-12-10 21:41:48.0	615	市公安局	可以更改	如果不	网友您好！改	2015-12-11 17:07:38.0				
89758	0	169604	zhang1	求助	2015-12-10 21:17:04.0	567	市行政审批中心	审批	工	您好，因	工审批	2015-12-15 09:06:06.0			

## 针对提速做的尝试

由于十堰问政网站的数据比较多，约为十二万条，本小组针对提速做了不同尝试。

- 打开8个jupyter notebook页面同时爬取
  - 我们首先去获取url，将八千条左右的url分成了8份，然后同时进行爬取。结束后，再将8份内容用concat合并起来。
  - 总耗时在3小时以上。
- 使用多进程
  - 使用进程池默认进程数量，即cpu核心数量
  - 总耗时在9993.40s，即166.5分钟
- 寻找合适进程池数量
  - 由于进程池默认进程数量比较小，我们决定增加进程数量。

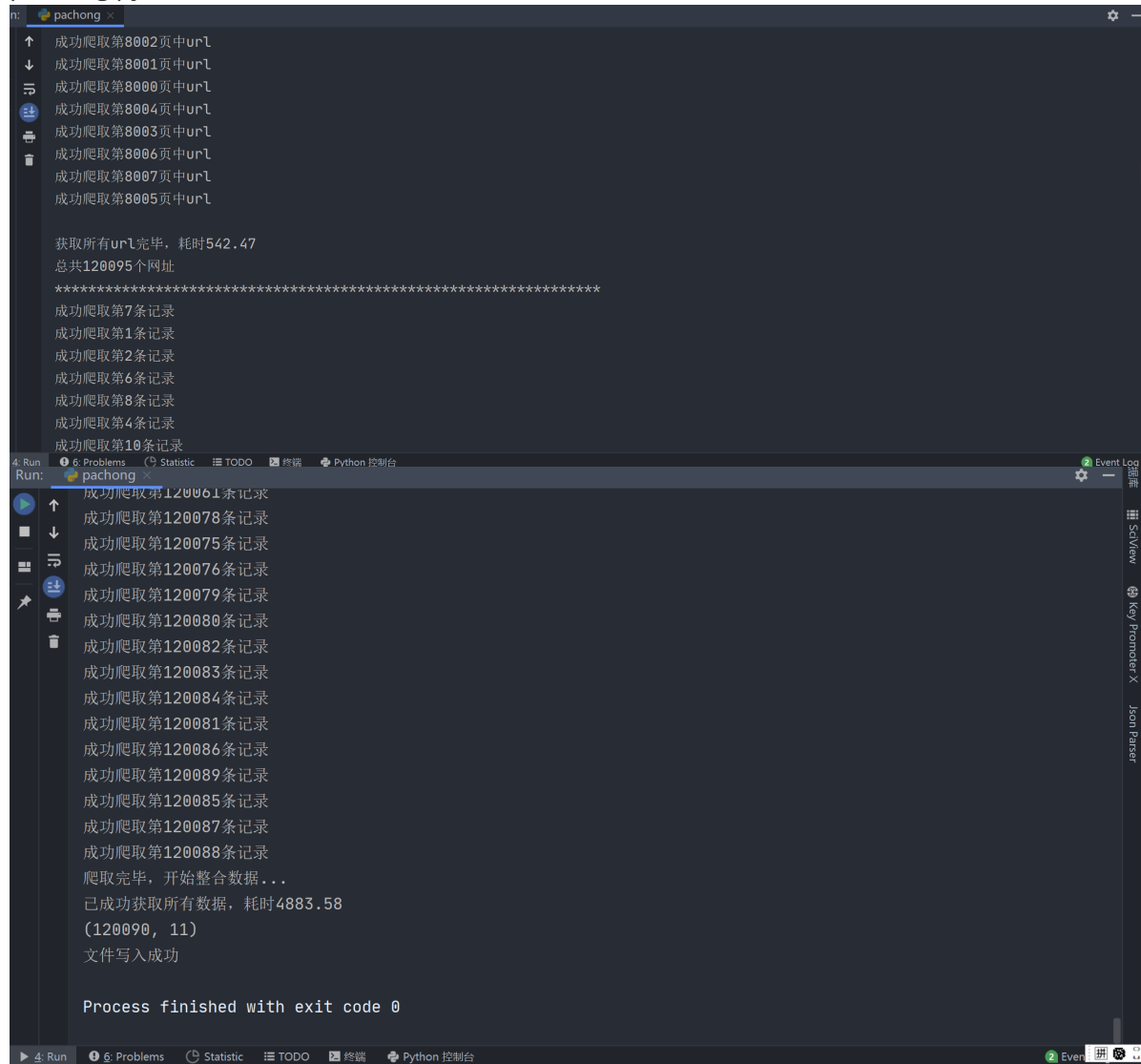
进程数量 \ 网页数量	爬取500个网页	爬取1000个网页	爬取2000个网页
默认	65.04s		
开8个	67.05s		
开10个	55.52s	105.42s	
开12个	47.51s	90.03s	
开14个	44.24s	80.47s	
开16个	39.77s	76.32s	
开18个	39.64s	73.95s	
开20个	40.61s	73.62s	138.89s
开22个	41.93s	72.81s	137.03s
开24个			138.17s

- 通过简单的测试，非常粗略的估计在22个进程数量时，时间比较短，可以又可以提速。
- 22进程总耗时：524.47+4883.58=5408.05s，即90分钟。
- 相比默认cpu核心数量提速将近一倍

## 针对异常处理

- 由于爬取网页耗时很长，所以异常处理和过程输出显得尤为重要。
- 我们使用traceback库，可以即时获取各种异常信息，发现问题。
- 同时如果个别网页出现爬取异常，整个进程并不会被中断。事后回溯还可以找到问题网页的具体报错。

pachong.py文件爬取全部数据时 控制台的部分输出展示：



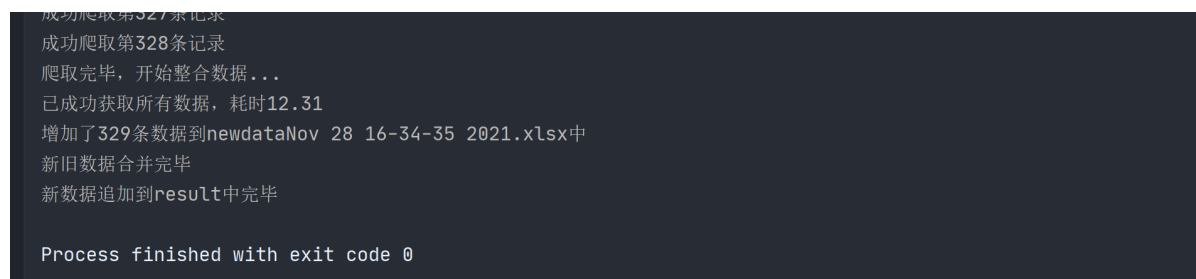
```
↑ 成功爬取第8002页中url
↓ 成功爬取第8001页中url
≡ 成功爬取第8000页中url
📄 成功爬取第8004页中url
📄 成功爬取第8003页中url
📄 成功爬取第8006页中url
📄 成功爬取第8007页中url
📄 成功爬取第8005页中url

获取所有url完毕，耗时542.47
总共120095个网址
*****
成功爬取第7条记录
成功爬取第1条记录
成功爬取第2条记录
成功爬取第6条记录
成功爬取第8条记录
成功爬取第4条记录
成功爬取第10条记录

成功爬取第120061条记录
成功爬取第120078条记录
成功爬取第120075条记录
成功爬取第120076条记录
成功爬取第120079条记录
成功爬取第120080条记录
成功爬取第120082条记录
成功爬取第120083条记录
成功爬取第120084条记录
成功爬取第120081条记录
成功爬取第120086条记录
成功爬取第120089条记录
成功爬取第120085条记录
成功爬取第120087条记录
成功爬取第120088条记录
爬取完毕，开始整合数据...
已成功获取所有数据，耗时4883.58
(120090, 11)
文件写入成功

Process finished with exit code 0
```

更新函数的控制台输出：



```
成功爬取第327条记录
成功爬取第328条记录
爬取完毕，开始整合数据...
已成功获取所有数据，耗时12.31
增加了329条数据到newdataNov 28 16-34-35 2021.xlsx中
新旧数据合并完毕
新数据追加到result中完毕

Process finished with exit code 0
```

## 复杂性比较高的bug

- 12万条记录爬取下来后，保存为dataframe类型写入文件时报错，报错信息显示数据长短参差不齐。
  - 回溯控制台输出的内容，发现极个别网页爬取时报错，显示失败
  - 分析原因：beautifulsoup解析网页时，极个别网页和其他网页源代码有差异，相同的模式爬取不到信息，导致数据长短参差不齐，无法以dataframe类型写入文件之中。
  - 最后我们调整beautifulsoup抓取的内容，白框处改为-1得以解决

```
# 受理单位
unit = data.find_all('span', 'mtp3')
unit = [x.text for x in unit][-1]
unit_.append(unit)
```

2.
  - 12万条记录爬取下来，保存为dataframe后，写入excel时又失败。
  - 多方查找发现，原来pandas导出到Excel会报URLS数量超出65530警告。
  - 我们尝试过这种方法：

```
writer = pd.ExcelWriter('xxx.xlsx',engine='xlsxwriter',options=**
{'strings_to_urls': False})
```

将url变成字符串保存进去。但是失败了。

- 最后一致认为，excel中没必要保存每条记录的原网址，可以直接将url删除