文档的执行顺序为：

1. 先在 sql 数据库中创建一个名为 housingprice 的数据库

create database hp;

2. 执行 step1.py 文件创建 datalist 表
3. 创建成功后，执行 getMsg.py 文件爬取郑州市房价信息，结果保存在 hp 数据库中的 datalist 表中

ID	NAME	STATE	PRICE	address_1	address_2	address_3	huxing	square	totalprice	primaryID	type
243334	正商华钻汇景公寓	推荐	10000	管城	管南片	中州大道与南三环交汇处向西300米路北	商住	41	410000	243334410	1
243334	正商华钻汇景公寓	推荐	10000	管城	管南片	中州大道与南三环交汇处向西300米路北	商住	45	450000	243334450	1
246939	和昌博景国际	在售	9300	中原	中原万达	陇海路与西环交叉口东南角	商住	39	362700	246939390	1
246939	和昌博景国际	在售	9300	中原	中原万达	陇海路与西环交叉口东南角	商住	44	409200	246939440	1
246939	和昌博景国际	在售	9300	中原	中原万达	陇海路与西环交叉口东南角	商住	50	465000	246939500	1
246939	和昌博景国际	在售	9300	中原	中原万达	陇海路与西环交叉口东南角	商住	59	548700	246939590	1
249416	万科城	在售	13000	高新区	高新区	科学大道与西四环交汇处	商住	33	429000	249416330	1
250725	鑫苑二七鑫中心	在售	11500	二七	陇三环	大学南路与南四环交汇处西南角	商住	50	575000	250725500	1
251206	郑西鑫苑名家	在售	5800	荥阳市	荥阳市	荥阳市郑上路广武路西南	商住	43	249400	251206430	1
251206	郑西鑫苑名家	在售	5800	荥阳市	荥阳市	荥阳市郑上路广武路西南	商住	52	301600	251206520	1

NAME 表示小区名称，SQUARE 表示面积大小，STATE 表示房子属性（推荐、在售、待售、售罄），PRICE 是房子均价（如果-1 表示当前房子不在销售中），ADDRESS_1 为所在市区，ADDRESS_2 和 ADDRESS_3 再细分了一下地理位置，HUXING 表示户型情况，TOTALPRICE 表示房屋总价，TYPE 表示房源类型 1=新房 2=二手房

4. 将一手房的信息注释掉，打开执行二手房的程序，再执行一次（二手房的网站的灵活性很强，有时候代码执行不正确，需要微调）

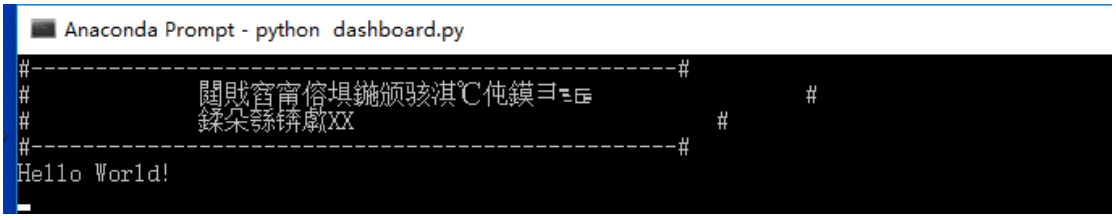
```
216
217 if __name__ == '__main__':
218     #print datetime.now()
219     main()
220     print datetime.now()
221     #secmain()
222
```

```
216
217 if __name__ == '__main__':
218     #print datetime.now()
219     #main()
220     print datetime.now()
221     secmain()
222
```

5. 爬取数据结束后，执行 dashboard.py 即可实现查询

切换到文件夹目录下

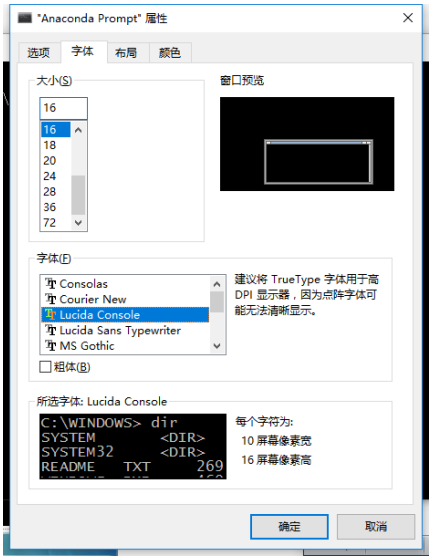
运行 python dashboard.py 会发现



所以要把 CMD 终端改为 UTF-8 格式

```
C:\Users\TinkleG\PycharmProjects\0416113>chcp 65001
```

修改 CMD 属性



再开始运行 python dashboard.py

Python 爬虫用到的技术

urllib2 https://docs.python.org/2/library/urllib2.html

```
urllib2.urlopen(url)
```

和

```
req = urllib2.Request(url)
print url
req.add_header('Referer', url)
req.add_header('User-Agent', 'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/57.0.2987.133 Safari/537.36')
res = urllib2.urlopen(req)
```

http://www.jb51.net/article/51941.htm

http://www.pythontab.com/html/2014/pythonhexinbiancheng_1128/928.html

beautifulsoup https://www.crummy.com/software/BeautifulSoup/bs4/doc/index.zh.html

```
soup = BeautifulSoup(res, 'lxml')  # lxml 解析成 BS4 数
for listall in soup.find_all('div', 'house-details'): 获取名为 house-details 类的内容

    msg = listall.find(class_='details-item').text.split('|') 获取名为 details-item 类的内容
href = listall.find(class_='houseListTitle').get("href") 获取链接地址
```

re https://docs.python.org/3/library/re.html

```
re.findall('[0-9]+', msg[2])[0]找出一个字符串中所有 0-9 的数字
```

threading

http://blog.csdn.net/zhanh1218/article/details/32131385

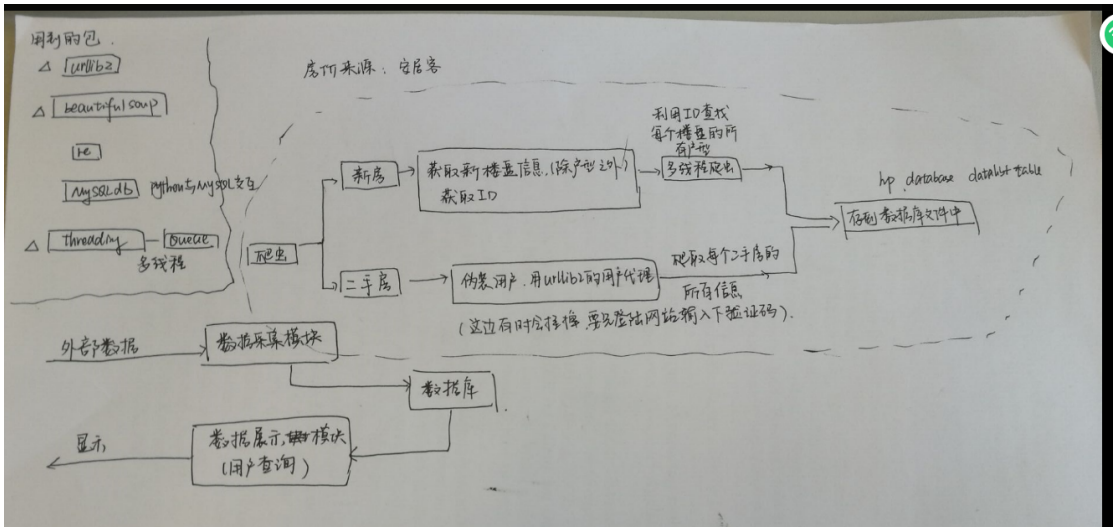
网页搜索 python threading 多线程

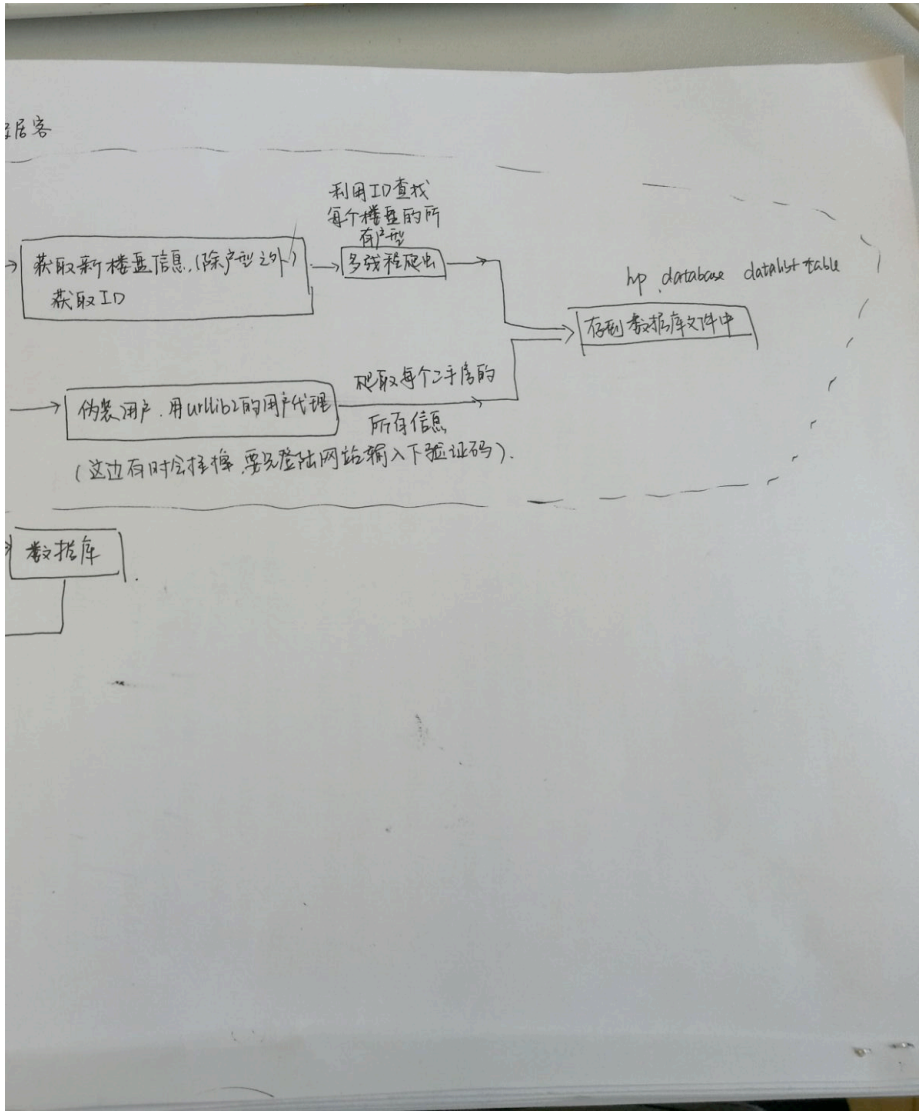
超时装饰器（用于解决网页解析超时线程被暂用无法释放的问题）

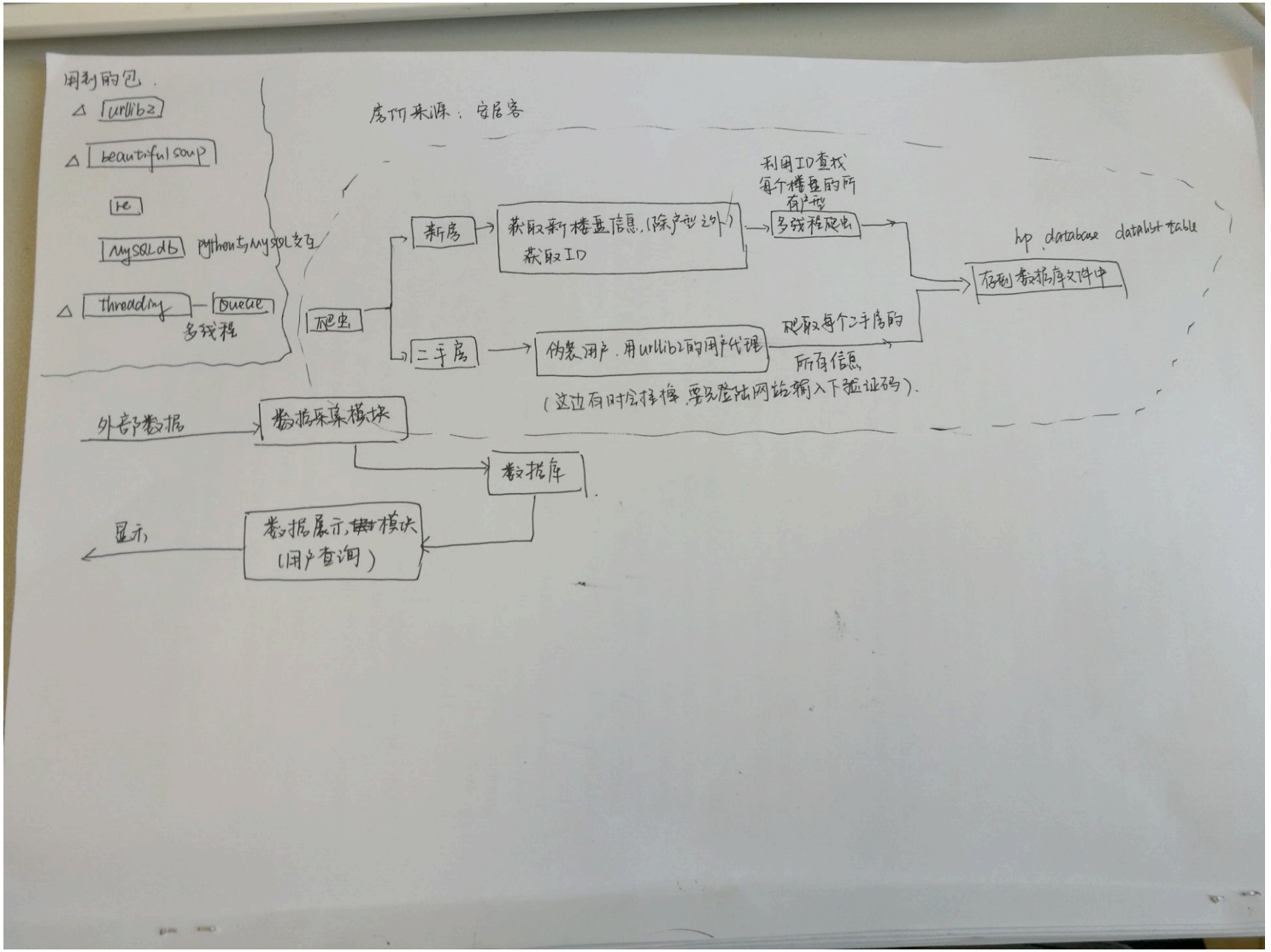
http://www.cnblogs.com/fengmk2/archive/2008/08/30/python_tips_timeout_decorator.html

mysql 与 Python 的交互（都在 create_table.py 文件中）

http://blog.csdn.net/zgl_dm/article/details/8710371




获取新楼盘信息.(除产权之外)
获取ID
利用ID查找
每个楼盘的所
有户型
多线程爬虫
hp database datalist table
存到数据库文件中
伪装用户.用urllib2的用户代理
爬取每个二手房的
所有信息
(这边有时会报错,要先登陆网站输入下验证码).
数据库


用到的包：
urllib2
beautiful soup
re
MySQLdb
python与MySQL交互
threading
Queue
多线程
房价来源：安居客
爬虫
新房
获取新楼盘信息（除户型之外）
获取ID
利用ID查找每个楼盘的所有户型
多线程爬虫
二手房
伪装用户，用urllib2的用户代理
爬取每个二手房的所有信息
（这边有时会弹框，要先登陆网站输入下验证码）
存到本地数据库文件中
hp database datalist table
外部数据
数据采集模块
数据库
显示
数据展示模块（用户查询）