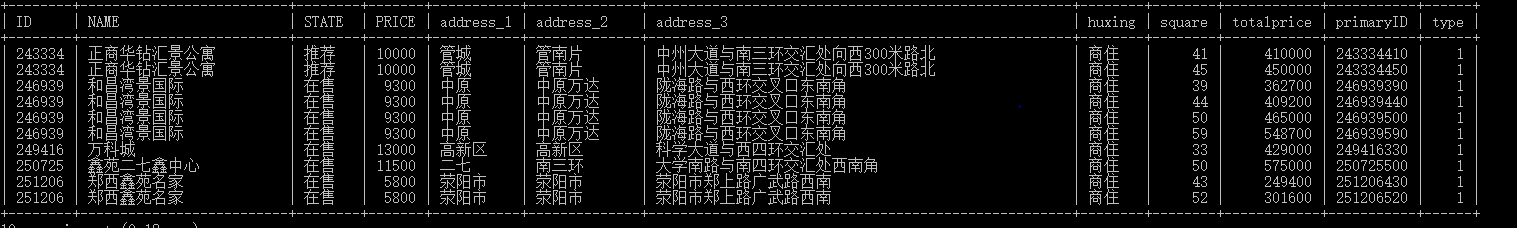
文档的执行顺序为：

1. 先在sql数据库中创建一个名为housingprice的数据库

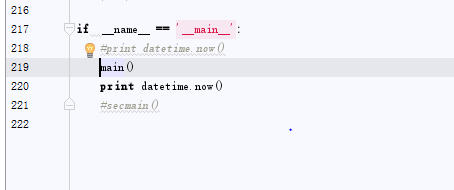
create database hp;

1. 执行step1.py文件创建datalist表
2. 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



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

1. 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）

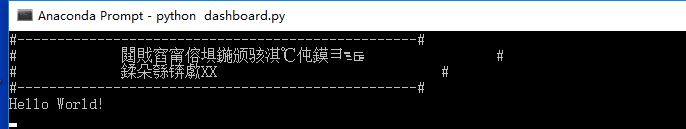




1. 爬取数据结束后，执行dashboard.py即可实现查询

切换到文件夹目录下

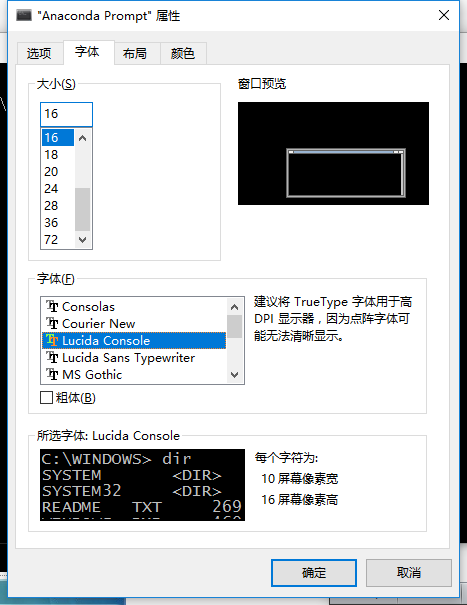
运行python dashboard.py会发现



所以要把CMD终端改为UTF-8格式



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>

