蝈蝈过节在家大吃特吃胖了不少，越发懒散，某天接到了新任务要爬去房地产信息，实在是不想接啊，bs4爬网页各种正则，真是太麻烦了。这时候我脑中想到了一个python很著名的爬虫框架-------scrapy，既然是框架肯定很好用咯。

1. 准备工作

语言：Python3

工具：Pycharm，Firefox

技术: Scrapy，Matplotlib

1. 流程概述

框架安装

项目生成

完成items.py

完成爬虫逻辑

完成pipelines.py

完成settings.py

根据 数据作图

1. 详细实现

框架安装，这个特别需要说下，scrapy既然是个框架就会安装一大堆包，有些包需要C++编译，蝈蝈不是写C#的并没有visual studio，可以去网上下载已经编译好的twisted包，然后应该就能安装成功了

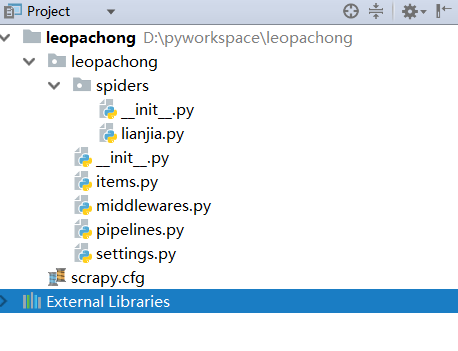
Scrapy 和我们以前的代码有点不同，他需要生成一个项目

scrapy startproject leopachong

Cd leopachong

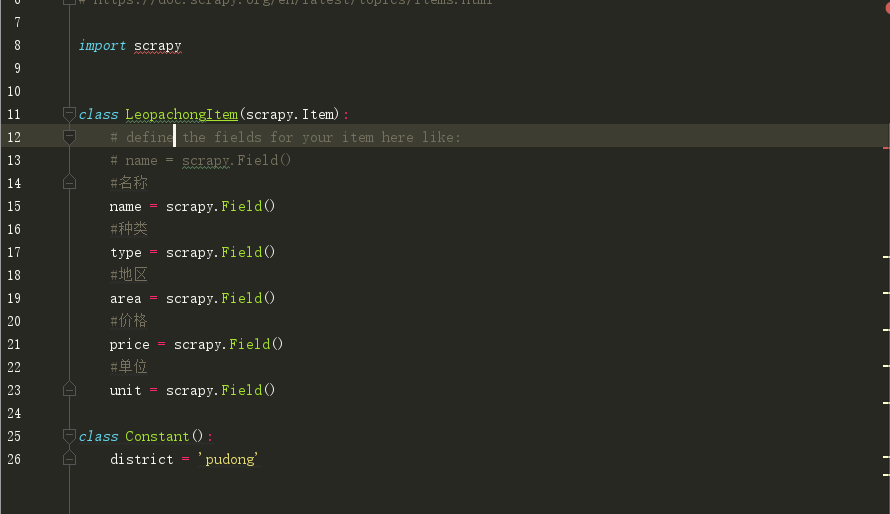
scrapy genspider lianjia lianjia.com

需要输入以上命令



这是完成以后项目内的情况

Items.py 是我们需要抓取的参数名

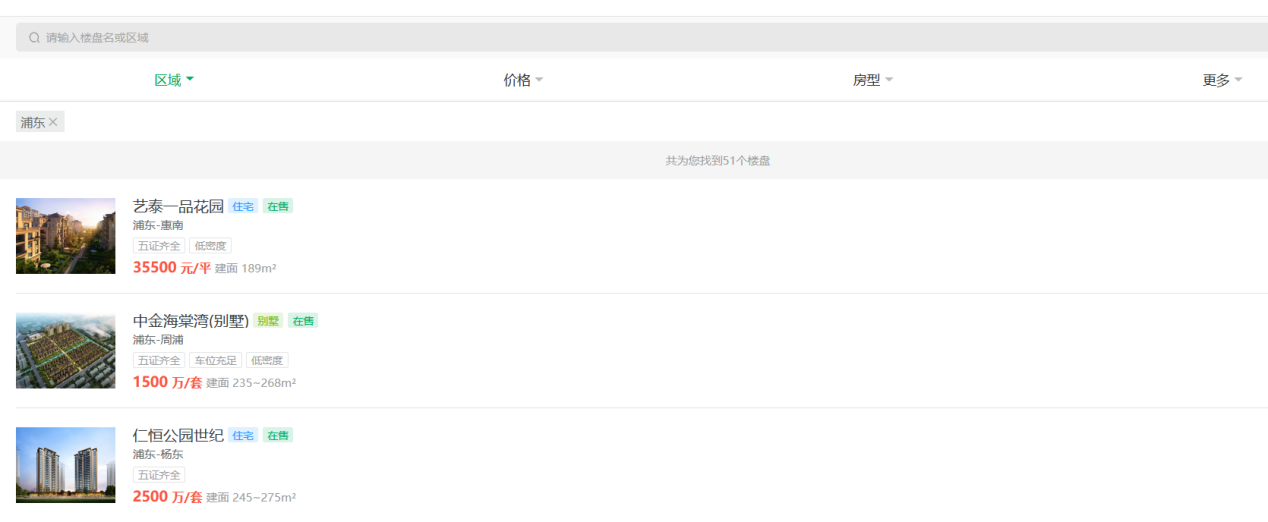


根据提示完成，毫无难度。

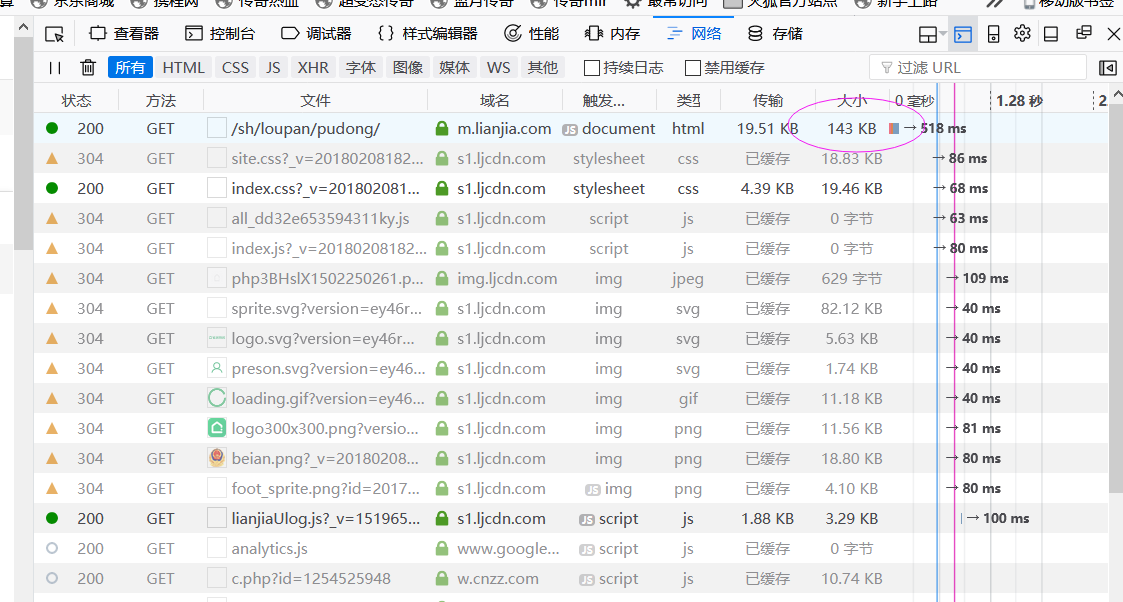
我在最后放了个记录常量的类，python是没有直接生成常量的方法的，暂时用一个类表示

看名字就知道我们要爬哪家网站了吧，反正在蝈蝈的城市里哪里都有那家门店

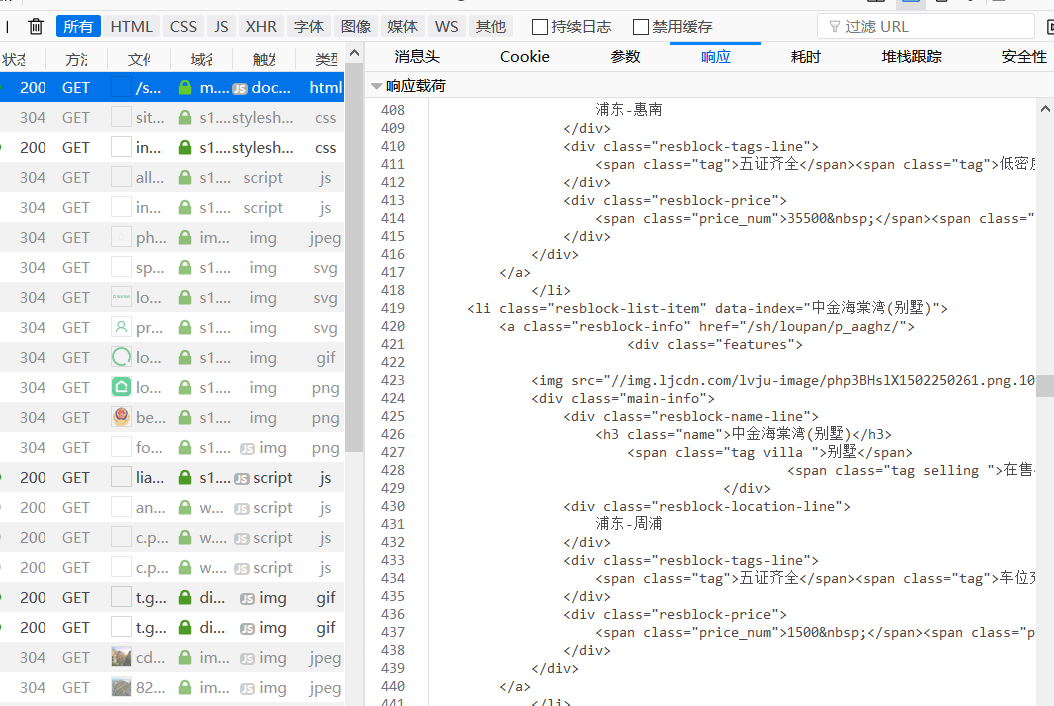
照例直接上那家公司的移动端网站，反爬的风险小些



一顿操作发现了房产信息，这不就是我们需要的嘛？

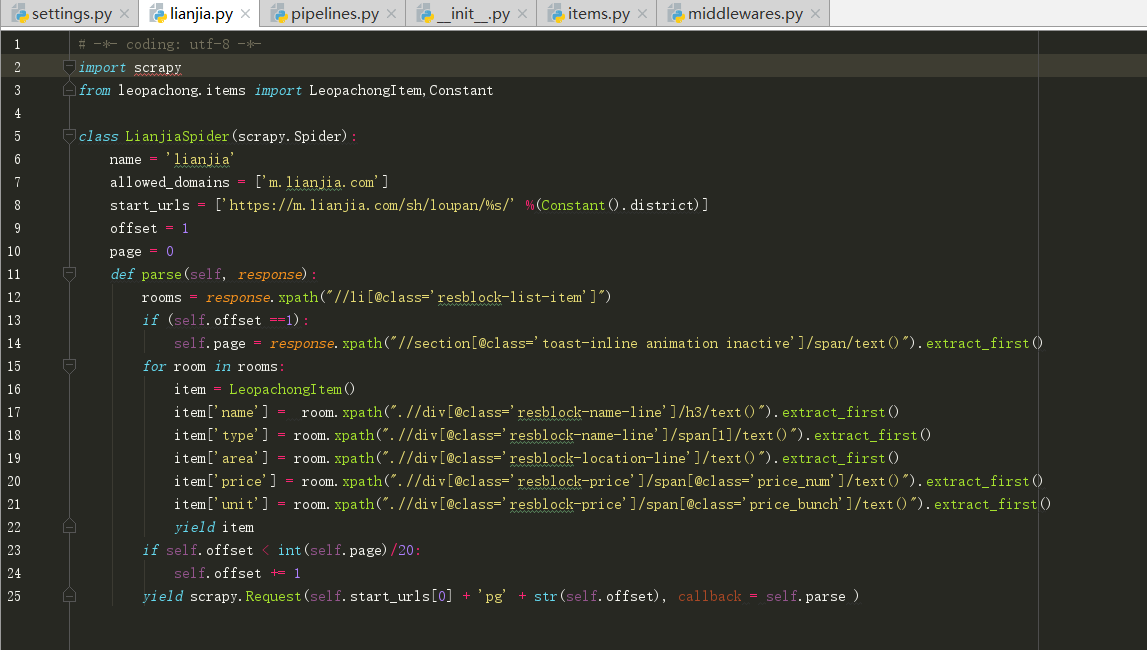


红圈部分的数据量如此巨大，我们点进去看下



不就是我们需要爬取的网站吗？

然后我们完善spiders内爬虫类



完整代码如上，我来分部讲解下

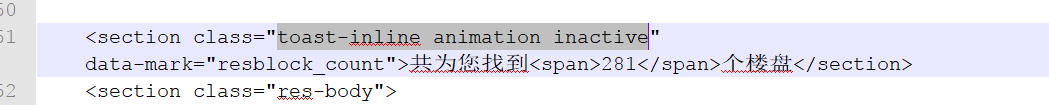
6，7行是默认的，我们可以自己写，也可以不管

8行是我们需要连接的url，浏览器可以看到，这里就是使用了常量类，具体常量需要填哪些等会会提到

9行Offset是页码，从第一页开始

10行是我存放总条数用的

12行我们会得到一个items，后面那堆乱码一样的东东是xpath语法，我敢保证比正则简单多了，网上随便看个说明10分钟就懂了



13，14行的意思就是如果我第一次爬，我要获取上图281那个数字，用来计算页数



15行到22行，就是我利用xpath爬去网页数据

23，24行是限制条件，如果我小于总页数我的页码+1，每页人工计算下是20条

24行是读取下一页，并且回调



Pipelines.py是持久化方法，这是用来存放数据的，我就用csv存放起来

Settings.py是设置参数

ROBOTSTXT\_OBEY = *False*

如果是默认的true会获取不到数据

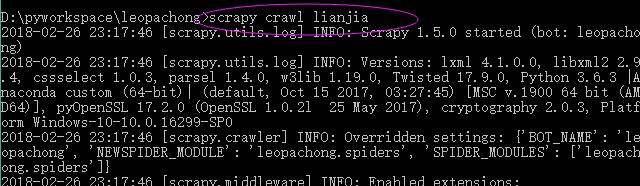
DEFAULT\_REQUEST\_HEADERS = {  
 'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,\*/\*;q=0.8',  
 'Accept-Language': 'zh-CN,zh;q=0.8,zh-TW;q=0.7,zh-HK;q=0.5,en-US;q=0.3,en;q=0.2',  
 'User-Agent':'Mpzilla/5.0(Windows NT 10.0; Win64; x64; rv:58.0) Gecko/20100101 Firefox/58.0'  
}

Header头设置，模拟浏览器

ITEM\_PIPELINES = {  
 'leopachong.pipelines.LeopachongPipeline': 300,  
}

这是pipelines处理的条数，默认300，应该没有那么多

由于是框架，pycharm并没有main方法可以直接调用，所以需要在命令行调用



项目目录下输入scrapy crawl 爬虫名 爬虫就启动了

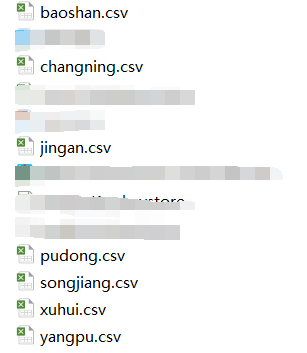
最后看下成果。。。



各个区的常量是怎么来的呢？随便打开一张需要爬的网页，仔细看下，就能看到代表各区的标签了

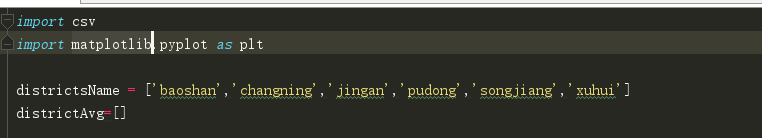


这样我们搜集了一些数据，如下图



说到科学绘图Matplotlib真是太有名了，今天我们来研究下怎么绘图

重新创建一个python类

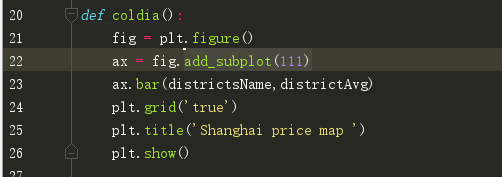


先引入两个类库，一个负责打开我们的CSV文件，一个负责作图

然后创建两个list，给一个list赋上我们的区名



这是一个打开csv文件，然后计算平均数，并且把平均数放到list里的方法



这个就是绘图的方法了，今天我们画的直方图，非常简单

21行调用方法，新建一个制图窗口

22行把整个窗口分成一行一列去其中一块

21，22行比较难懂但又是每次都需要的，大家可以网上搜索下，深化理解

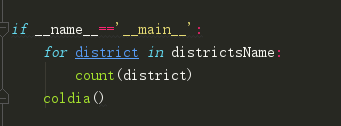
23行就是画图

横坐标就是我们之前输入的区名，纵坐标是我们计算的平均房价

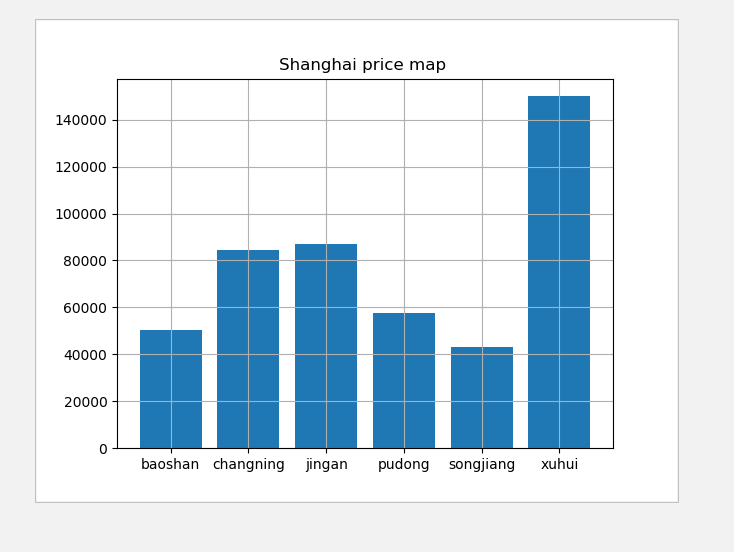
24行显示网格

25行给图输入标题，不过输入中文需要字体文件，暂时输入英文

26行就是显示图片啦



Main方法，最后显示



徐汇由于样本最少，并不足以说明问题

从图中看出市中心的房价依然很是凶残，静安的价格是80000+/平

较为偏远的地区房价就略便宜，松江离市区最远，所以房价很是便宜，不过依然不是普通大众能买得起的 T\_T