对于链家的爬虫

我爬取了链家新房的价格、名称、区、具体位置、状态、类型和面积数据

和链家二手房的区、具体位置、单价、名称、小区名、总价和面积数据

一开始爬二手房的时候，没有发现100页以后也可以直接改url

所以我是用

for i in range(7):

for j in range(100):

url=base\_url+"a"+str(i+1)+"d"+str(j+1)

这样一段代码，（a代表面积标签，d为页数）爬取了700页房屋的信息

后来发现可以直接改url

就只要

for i in range(1000):

url=base\_url+str(i+1)

这样写就可以，爬取了总共1000页的数据。

**爬虫的难点**

我认为主要是把面积数据从一串文本中分离出来，这里用到了split函数，但是因为我用xpath的方法爬虫，它生成的是一个list 所以要转换一下，将这个list里面的字符串遍历，对字符串使用split函数。

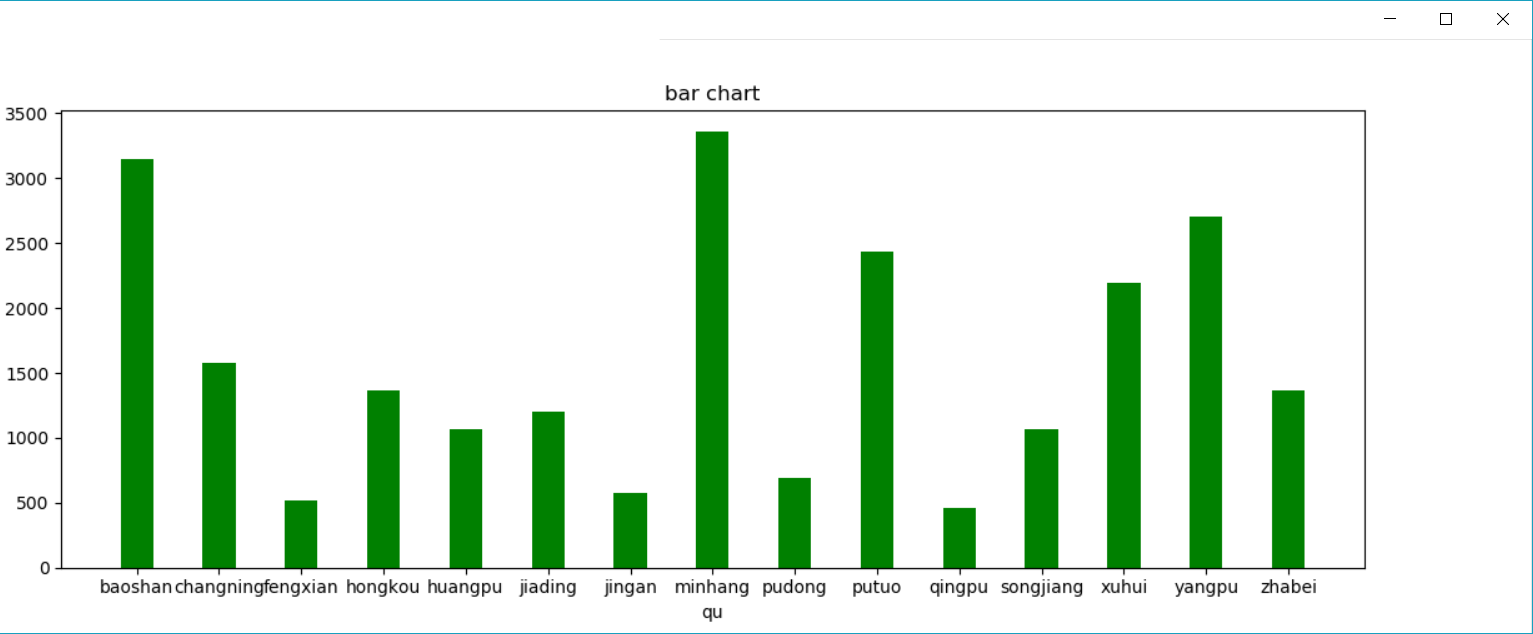
**数据分析**

我用了pandas自带的一些功能和matplotlib实现了数据分析和可视化

首先看到所有的字段都没有空值



统计了每个区房屋的数量



之后用matplotlib绘成了柱状图。这里其实其他的数据没有汇总，是因为面积、单价都是object格式，并且有很多转义字符’t，只要把他们转换成float类型，就可以了。

绘图部分也有很多完善的空间。

新房的数据比较少。后期用excel调整了一下。

**总结**

整个过程是自己看教程慢慢写的，虽然程序很简单，后期分析也因为时间关系，没有能做好。但是我发现python的功能很强大，也很灵活，还想好好学一学，我觉得这个项目很有意义。