爬取 租房/买房/二手房 信息 可行性

爬取 图书馆 信息 可行性

房子

在浏览器 ----> 所见(到租房信息)即所得(对应 HTML 源码) 而后者即是我们需要在 代码中（爬虫）得到的 “文本”信息

我所做的基本上与我们正常点击上网的方式是一样的，只是我们要 **显式** 地提供一些信息去从对方服务器获取（”下载”）相应代码文本；而后者只是 **隐式**地（将这些信息随着我们的点击事件的发生一并发送给了）对方服务器 从而“下载”到我们的本地，同时浏览器读取“理解”这些文本信息，并显示出来。

所以爬虫所做的事情就是自动化 模拟 浏览器去获取信息，但我需要做的不仅仅是这点事情，还要包括：

怎么去模拟浏览器
怎么能让对方服务器愿意给我数据（对方会判断我是否一个机器人爬虫-反爬虫）
怎么去模拟我们在浏览器的事件（如点击，滑动等等）
怎么更好地解析所获文本信息（解析规则[一般不变] 但会由该网站更新而改变）
怎么更好得存储这些信息
怎么能够将以上操作更加高效快速（高并发？ 多进程 信息量大）
以上是我这个 Python 爬虫 的**重点难点**！

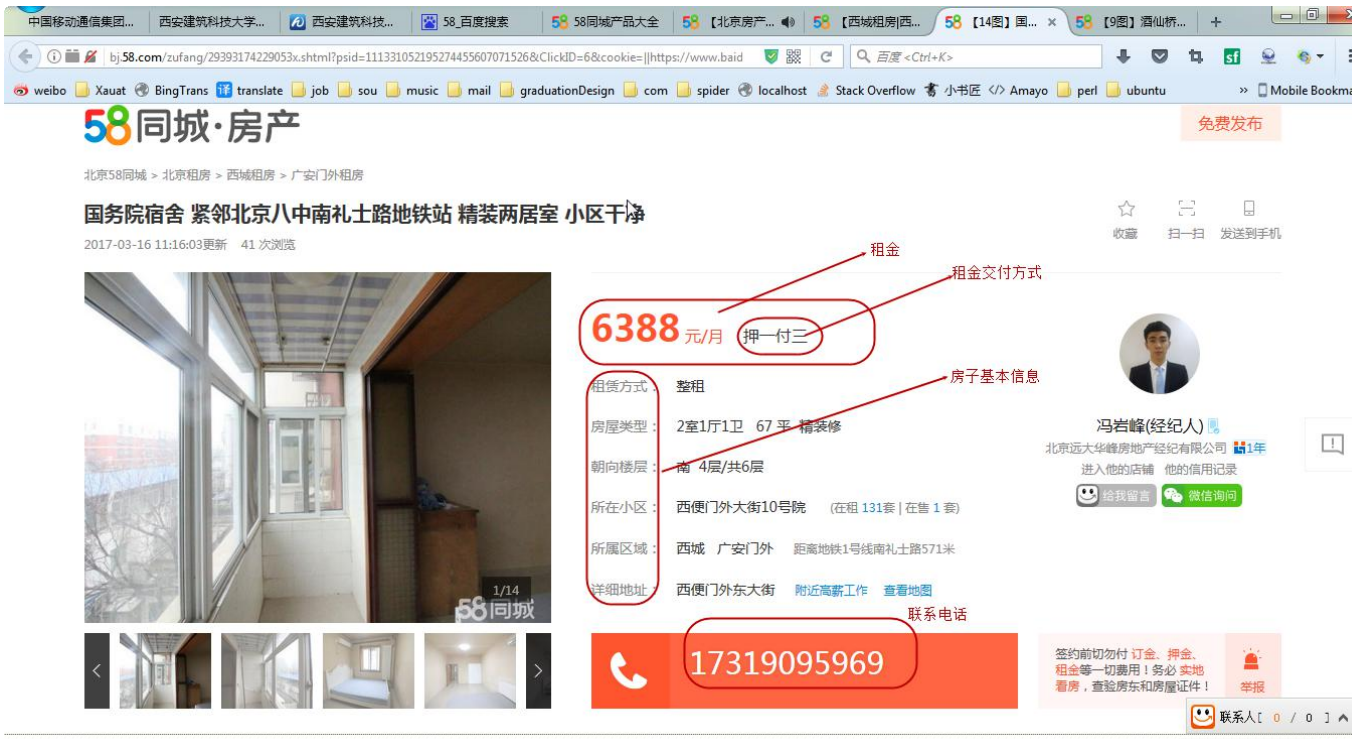
接下来说明 租房 与 图书馆 这两个爬虫方向的比较

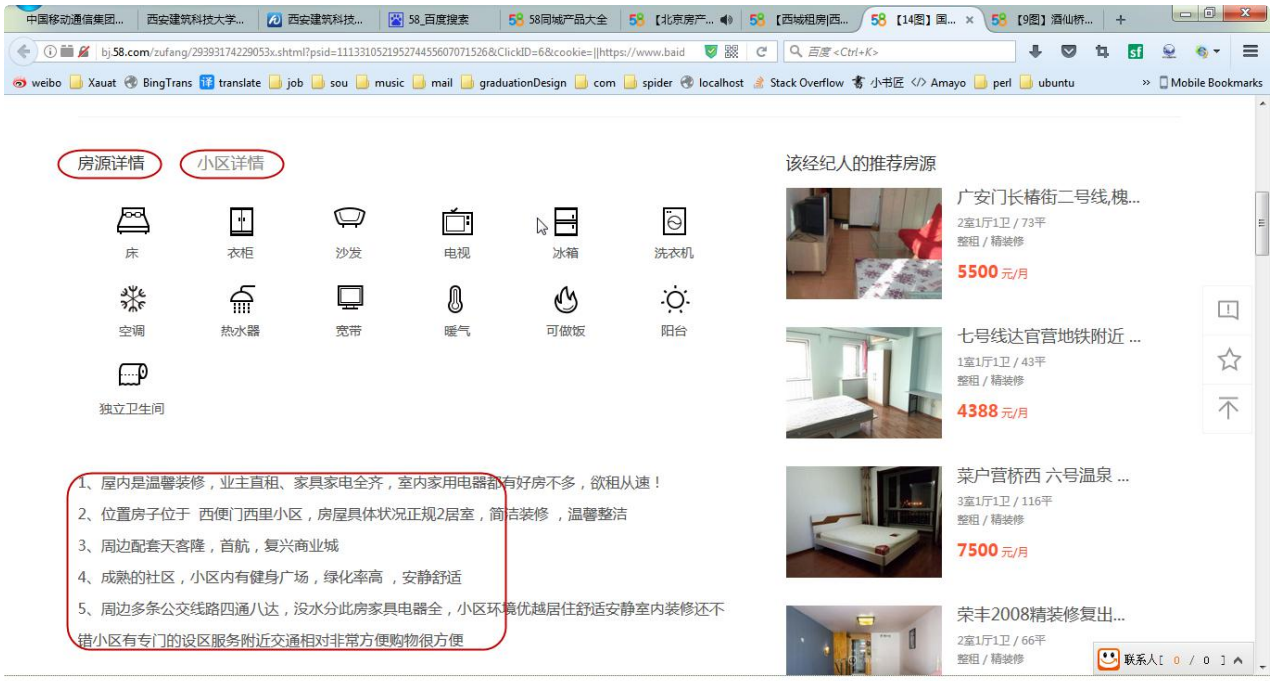
需要老师具体分析比较怎样能让这个两个方向可以更加 有功能 丰富，孰好？

以下是我的简单比较分析结果。

这里主要是针对的网站 58 同城 安居客 我爱我家 链家网

接下来以其中一个租房信息进行具体说明：


58同城·房产
免费发布
北京58同城 > 北京租房 > 西城租房 > 广安门外租房
国务院宿舍 紧邻北京八中南礼士路地铁站 精装两居室 小区干净
2017-03-16 11:16:03更新 41 次浏览
收藏
扫一扫
发送到手机
租金
租金交付方式
6388 元/月
押一付三
房子基本信息
租赁方式： 整租
房屋类型： 2室1厅1卫 67 平 精装修
朝向楼层： 南 4层/共6层
所在小区： 西便门外大街10号院 (在租 131套 | 在售 1 套)
所属区域： 西城 广安门外 距离地铁1号线南礼士路571米
详细地址： 西便门外东大街 附近高薪工作 查看地图
联系电话
17319095969
冯岩峰(经纪人)
北京远大华峰房地产经纪有限公司 1年
进入他的店铺 他的信用记录
给我留言
微信询问
签约前切勿付 订金、押金、租金等一切费用！务必 实地看房，查验房东和房屋证件！
举报
1/14
58同城
联系人[0 / 0]


房源详情
小区详情
床
衣柜
沙发
电视
冰箱
洗衣机
空调
热水器
宽带
暖气
可做饭
阳台
独立卫生间
1、屋内是温馨装修，业主直租、家具家电全齐，室内家用电器都有好房不多，欲租从速！
2、位置房子位于 西便门西里小区，房屋具体状况正规2居室，简洁装修 ，温馨整洁
3、周边配套天客隆，首航，复兴商业城
4、成熟的社区，小区内有健身广场，绿化率高 ，安静舒适
5、周边多条公交线路四通八达，没水分此房家具电器全，小区环境优越居住舒适安静室内装修还不错小区有专门的设区服务附近交通相对非常方便购物很方便
该经纪人的推荐房源
广安门长椿街二号线,槐...
2室1厅1卫 / 73平
整租 / 精装修
5500 元/月
七号线达官营地铁附近 ...
1室1厅1卫 / 43平
整租 / 精装修
4388 元/月
菜户营桥西 六号温泉 ...
3室1厅1卫 / 116平
整租 / 精装修
7500 元/月
荣丰2008精装修复出...
2室1厅1卫 / 66平
整租 / 精装修
联系人[0 / 0]

以上我们在浏览器看到的租房信息都可以在其对应源代码可以一一查看到的，若是您想在自己浏览器上看到这个 HTML 可以打开 对应网页，可以参看以下方法（查看源代码）链接：

http://jingyan.baidu.com/article/09ea3ede21b18cc0aede39d7.html

以上的 HTML “文本” 信息即需要我们继续进行 **“解析”**
解析之后获得以下“可见”信息

毕设功能分析

从以上简单网页截图说明目前可以获知以下简单信息：

1 租房信息
2 租金
3 租金规则
4 租赁方式
5 房屋类型
6 朝向楼层
7 所在小区
8 所在区域（可定位到 区级 街道级别）--> 可进行区域分析
9 详细地址
10 联系电话
11 房源提供（房子可以提供些什么）

<这里是重点>

但无法获取到买卖双方个人隐私信息（如年龄，性别）

其他租房网站基本相同

就目前功能，我们能够做到的就是 通过分析 可以获得以下关系：

地区、类型、方位 <--> 租金 （3 个关系）

当然我可以在通过对应地址（经纬度？）来去获取附近公司信息

进而分析 其中关系

代码说明

目前房源量到底是多少未知，但是以上四个租房网站所对应的爬虫代码都有人实现过了，已经有了较多的代码基础，我可以拿来并进一步加工使用。但并不能保证都可以正常运行，因为其他人与我的配置，应用时间，网站是否更新都有很大关系！

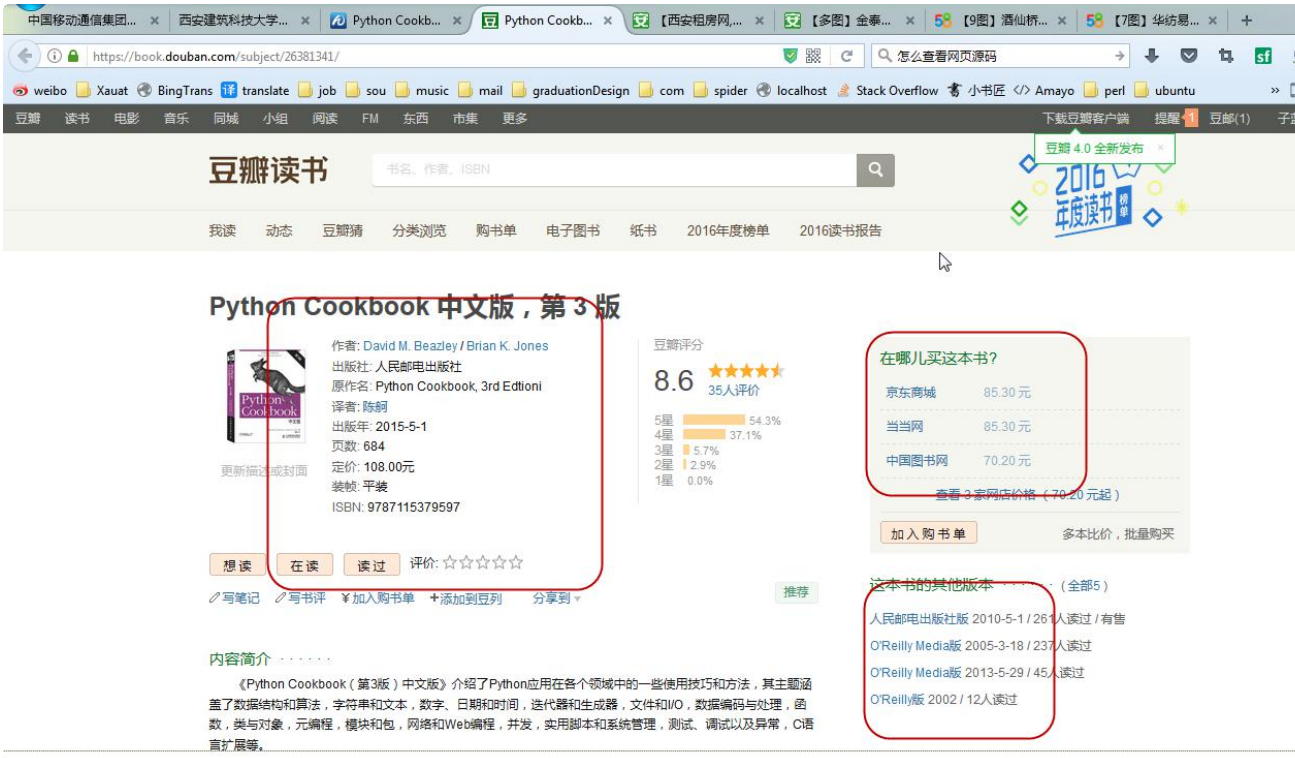
若是老师有什么再次基础上可以得到更多“关系”（功能），希望可以继续添加。

图书馆

上文已经详细简单说明了爬虫原理，这里就简单分析一下建大图书馆检索系统

接下来我要以 Python CookBook 这本书进行介绍

我在此网页上连接到 豆瓣 网上 可见更多信息



正是因为 每个 图书条目的右边 有 豆瓣 当当网 的更多详细说明，才让之前我的图书馆 爬虫计划更加 丰富！

<这里是重点>

就目前来说，我可以获得以下信息：

在馆图书的基本信息
- 图书书名
- 编号
- 简单介绍等其他信息
- 藏书地址
- 校内借阅量

当我连接到 豆瓣， 当当网后，可获得就更多了

- 豆瓣评分
- 购买量
- 购买地址
- 其他版本信息
- 已知阅读量
- 更重要的是 这个两个网站 支持 评论！
- 我可以获得
- 评论者简单信息，进而有一个简单的“隐私”分析

代码说明

就目前资料来说， 图书馆有关爬虫信息不多且不一样，这里是个小难点，但是豆瓣，当当网的爬虫已经有很多人写过了，我们需要的信息与他人虽然不同但可参考，基本上这个方向上的代码需要完全自己写，没有太多参考，与租房的那四个网站一样，这两个外网的架构未知，尤其是因为爬的人太多了，他们公司内部肯定有一个反爬虫机制！

功能总结说明

两个方向不同，实现的功能虽然不同，但是实现的基本方法是一样的，通过寻找关系，进而分析其中关系，可以利用某些统计方法，进一步分析出其中关系，为了显示出所有功能，可以通过在论文上陈述图表，网页表格，网页图形，EXECL 图形等等进行可视化，只要有大量数据就行。

关系即图表。

以上就是我今天下午具体针对这两个方向做的简单分析，如果老师有什么更加合适的点子可以及时沟通哦！