爬虫项目系统设计文档

小组成员：同二鹏，郑黎明，张耀元

在北上广的大城市，有很多就业机会，吸引着许多年轻的程序员，但是大城市的房价之高也成为年轻人立足的一个最大的障碍。在大数据时代，我们的数据分析技术可以应用生活的方方面面，在爬虫大项目的构思阶段，我们希望分析一下大家最关心的房价和工资数据，给程序员们提供一个更好地选择。我们小组的爬虫项目爬取的数据是：各个城市的房价和工资。系统架构图如下所示：

利用jsoup抽取网页链接

利用jsoup抽取数据并进行数据分割

分析目标网站中的源代码

读取文件中的数据，求出各城市的平均房价和个城市的平均工资

将分割之后的数据存储到文件中

分析房价和工资性价比，用条形图表示最终结果

**1. 利用jsoup提取网页链接，解析网页数据**

**1.1 获取房价数据**

**目标数据**：各城市的各行政区平均房价

**数据来源**：<http://www.fangjia.com/>

首先，查看网站全国所有城市房价导航页面的源代码

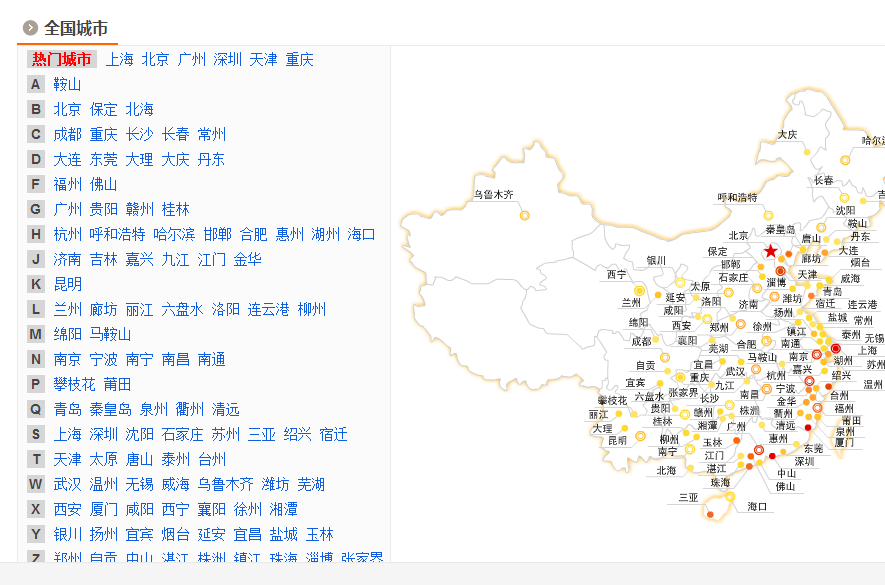


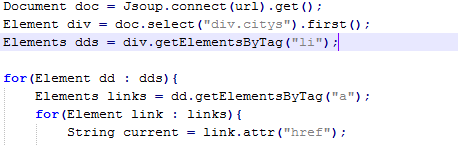
图1 房价网全国所有城市导航页面



图2全国所有城市房价导航页面的源代码

由图2可以看出各城市网页链接位于<div class=”citys”><li><a>标签中的href属性。

利用Jsoup抽取网页链接：



然后，查看单个城市的房价网页，如北京市的房价页面，见图3。

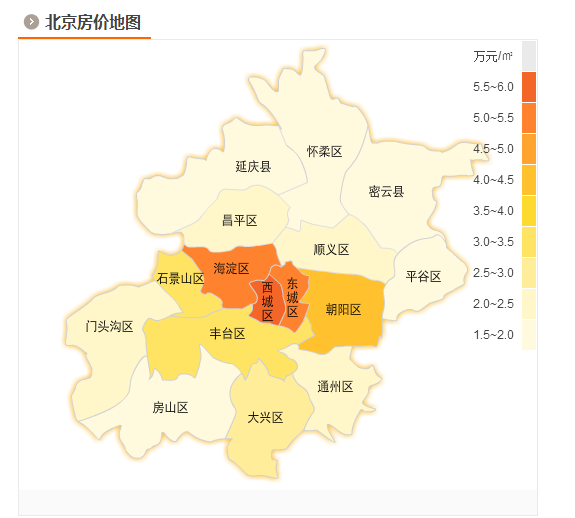


图3北京市区房价网页

在这个网页中，可以看到北京市的行政地图。点击任意一个北京市的区，可以看到区级的房价信息。依旧查看源代码，见图4：

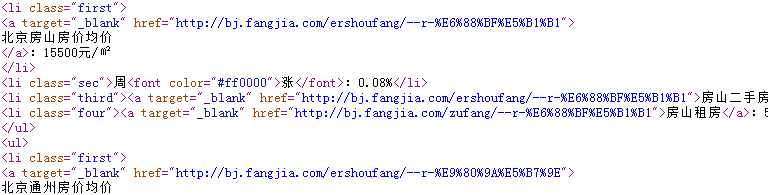
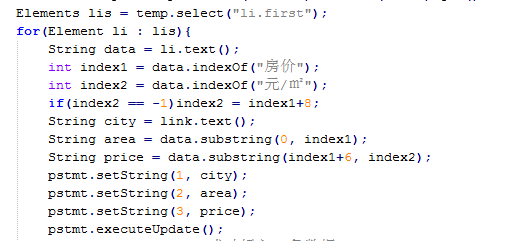


图4 北京市房山区房价网页源代码

可以看出我们需要的数据位于标签<li class=”first”>中，利用jsoup抽取数据并进行数据分割：



至此，各城市的行政区平均房价数据的抽取已经分析完毕。图5是爬取的部分数据截图。

**1.2 获取工资数据**

**目标数据**：中国最适合创业的20城市互联网职位工资

**数据来源**：[www.lagou.com](http://www.lagou.com)

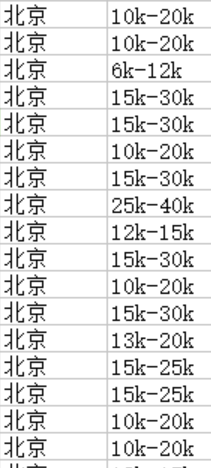
因数据获取的方法与获取房价数据方法类似，在此不做赘述。爬取的部分数据截图见图6

图5爬取的房价 图6 爬取的互联网工资

1. **根据爬取的数据，求每个城市房价和工资的平均值**

在第一阶段我们爬取的房价数据是“区级”房价的平均数值，根据很简单的求均值公式，我们就可以求出每个市的平均房价，图7是根据源数据求出的部分城市房价均值。同理，也可以根据源数据求出每个城市的互联网职位的平均工资，图8是工资的部分数据截图。

图7 平均房价 图8 平均工资

1. **根据已知的房价和工资信息，分析薪房比。**

我们根据爬取的数据求出的平均工资是每个月的月薪，如17K为每月工资为17000元。我们根据爬取的数据求出的平均房价为每平方米的房价。用平均月薪\*12/平均房价，就可以求出，一个人一年的薪水可以买到多少平方米的面积。图9就是我们最终的分析结果。



图9 薪房比较结果

一般而言，当“房薪比”在6倍以内是压力是较小的，也就是说，在6年内的工资可以买到自己住所的人幸福感不会因受到房价高的影响大打折扣。北京、上海、广州的“房薪比”都远远超过6倍，人们的生活压力可想而知。

1. **总结**

本次大作业我们爬取的工资数据较之网站中千万数据量而言是有限，而且网页中的数据都是一个范围值，不是准确的值，所以数据可能和某些数据的实际平均值有出入，这会影响到最终数据的准确性。