

**Python基础与数据分析**

题 目：上海链家租房数据的爬取与分析

学 系：信息科学学院

专 业：数字媒体技术

学生姓名：周润达，刘晓敏，罗永添，

谢明恒，尤诗颖

学 号： 18054116、18054059

18054100、18054089、18054066

指导教师： 潘志宏（副教授）

二〇一九 年 12 月

**项目分工一览**

|  |  |  |
| --- | --- | --- |
| **姓名** | **学号** | **分工内容** |
| 周润达 | 18054116 | 网页动态爬取数据、  数据保存为Excel、  数据分析、  word文档报告  PPT |
| 刘晓敏 | 18054059 | 数据分析、PPT |
| 罗永添 | 18054100 | 数据分析、word文档报告 |
| 谢明恒 | 18054089 | 数据分析 |
| 尤诗颖 | 18054066 | 数据分析、word文档报告 |

**目录**

[1 .引言 4](#_Toc28444311)

[1.1 选题背景和意义 4](#_Toc28444312)

[1.2 主要研究内容和主要工作 4](#_Toc28444313)

[2 .相关技术介绍 5](#_Toc28444314)

[2.1 xpath 5](#_Toc28444315)

[2.2 pandas 6](#_Toc28444316)

[2.3 matplotlib库 7](#_Toc28444318)

[2.5 Codecs库 8](#_Toc28444319)

[2.6 jieba库 8](#_Toc28444320)

[3.数据分析 9](#_Toc28444321)

[3.1 获取数据 9](#_Toc28444322)

[3.2 分析总结 14](#_Toc28444323)

[3.2.1利用jieba库来分析房子简介信息高频关键字以及关键词出现次数，并根据关键字频率生成词云 14](#_Toc28444325)

[3.2.2分析房子面积的数量 16](#_Toc28444326)

[3.2.3分析房租的数量 19](#_Toc28444327)

[4.总结与展望 22](#_Toc28444329)

# 1.引言

## 1.1 选题背景和意义

​​

随着互联网的普及，线上租房的优点更加突出。线上租房有四大好处:

1、房源获取：用户可以直接通过线上进行在线搜索周边的房源，房源获取更加全面也更加便捷。

2、精准房源：通过输入自己对于房源的需求，可以快速获取满足需求的精准房源。

3、在线交流：对于房价、房间配置等问题，用户都可以通过线上进行与房东进行交流。

4、线上合同：线上合同能够随时满足用户的查看需求，条款条例也更加方便进行查看。

线上租房在如今已经成为了新一代方式，既满足了人们的需求，又方便快捷，甚至有些价格比中介手中的价格低廉，也因此人们大多数都会优先选择在线上租房。为此，针对线上租房，我们对此进行进一步的解析筛选与了解其价格等，作出数据统计分析，以便网民进行大规模的了解参考，作出大概选择，而不是去盲目选择，浪费时间和精力。

## 1.2 主要研究内容和主要工作

###### 1.2.1设计Python网络爬虫进行数据爬取

如何设计一个Python网络爬虫是在短时间内获取大量商品信息的关键。因为网站有一些限制,我们主要是从多页面进行爬取及数据的格式化等方面进行了研究和设计。

###### 1.2.2利用Xpath进行爬取，解析和清洗

以链家网作为对象，利用Xpath技术爬取数据，并且使用Xpath解析数据。并且同时进行数据清洗。并且爬取主图片的url供参考。

###### 1.2.3数据分析

利用pandas库分析面积和房租做数据分析。

# 2. 相关技术介绍

## 2.1 xpath

XPath即为XML路径语言（XML Path Language），它是一种用来确定XML文档中某部分位置的语言。XPath基于XML的树状结构，提供在数据结构树中找寻节点的能力。起初XPath的提出的初衷是将其作为一个通用的、介于XPointer与XSL间的语法模型。但是XPath很快的被开发者采用来当作小型查询语言。

一个网页中通常含有xpath路径表达式，这些路径表达式选取了XML文档中的节点或节点集，与我们在常规电脑系统中看到的表达式十分相似。路径表达式是从一个XML节点（当前的上下文节点）到另一个节点、或一组节点的书面步骤顺序。这些步骤以“/”字符分开，每一步有三个构成成分：轴描述（用最直接的方式接近目标节点）；节点测试（用于筛选节点位置和名称）；节点描述（用于筛选节点的属性和子节点特征）。 本项目在爬取链家网时，基于通过xpath选取关键字爬取。其中，chrome中的插件xpath helper可以帮助我们快速获取xpath的表达式。但是有时候还是得直接编译。

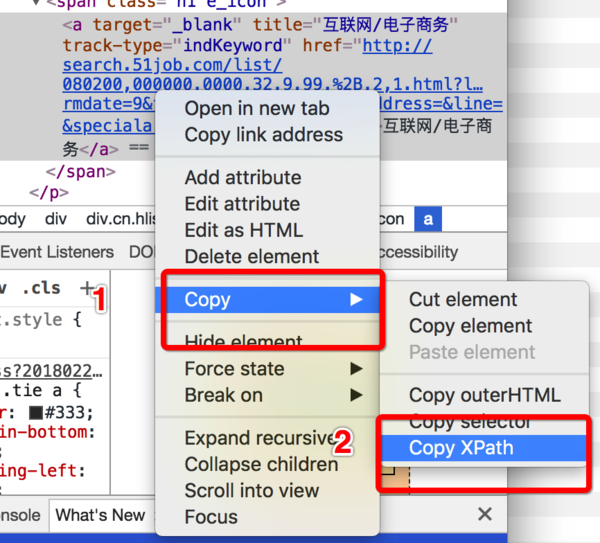


图2.1.1 xpath获取流程

## 2.2 pandas库

在科学运算中最为重要的两个模块，一个是 numpy,一个是 pandas。任何关于数据分析的模块都少不了它们两个。原因有两点：1.运算速度快，numpy 和 pandas 都是采用 C 语言编写, pandas 又是基于 numpy, 是 numpy 的升级版本。2.消耗资源少：采用的是矩阵运算，会比 python 自带的字典或者列表快好多。Numpy与Pandas的不同之处在于，如果用 python 的列表和字典来作比较, 那么可以说 Numpy 是列表形式的，没有数值标签，而 Pandas 就是字典形式。Pandas是基于Numpy构建的，让Numpy为中心的应用变得更加简单。

本项目在进行数据清洗时，主要使用了Pandas模块。Pandas是Python的一个数据分析包，该工具为解决数据分析任务而创建。其融入了大量库和标准数据模型，提供了高效的操作数据集所需的工具；同时还拥有大量能使我们快速便捷地处理数据的函数和方法。Pandas最初被作为金融数据分析工具而开发出来，因此，pandas为时间序列分析提供了很好的支持。 Pandas的名称来自于面板数据（panel data）和python数据分析（data analysis）。panel data是经济学中关于多维数据集的一个术语，在Pandas中也提供了panel的数据类型。Pandas的核心就像操作一个电子表格的无头版本，比如 Excel。你使用的大多数数据集将是所谓的数据帧（DataFrame）。你可能已经熟悉这个术语，它也用于其他语言，但是如果没有，数据帧通常就像电子表格一样，拥有列和行，从这里开始，我们可以利用 Pandas 以闪电般的速度操作我们的数据集。Pandas 还与许多其他数据分析库兼容，如用于图形的 Matplotlib，NumPy。

## 2.3 Matplotlib库

Matplotlib 是一个用于在 [Python](https://www.python.org/" \t "https://wizardforcel.gitbooks.io/matplotlib-user-guide/content/_blank) 中绘制数组的 2D 图形库。虽然它起源于模仿 MATLAB®[1] 图形命令，但它独立于 MATLAB，可以以 Pythonic 和面向对象的方式使用。虽然 Matplotlib 主要是在纯 Python 中编写的，但它大量使用 [NumPy](http://www.numpy.org/" \t "https://wizardforcel.gitbooks.io/matplotlib-user-guide/content/_blank) 和其他扩展代码，即使对于大型数组也能提供良好的性能。Matplotlib 的设计理念是，你应该能够使用几个，或者只有一个命令创建简单的图形。如果你想看到你的数据的直方图，你不需要实例化对象，调用方法，设置属性等等；它应该能够工作。

多年来，我常常使用 MATLAB 进行数据分析和可视化。 MATLAB 擅长绘制漂亮的图形。当我开始处理 EEG 数据时，我发现我需要编写应用程序来与我的数据交互，并在 MATLAB 中开发了一个 EEG 分析应用程序。随着应用程序越来越复杂，需要与数据库，http 服务器交互，并操作复杂的数据结构，我开始与 MATLAB 作为一种编程语言的限制而抗争，并决定迁移到 Python。 Python 作为一种编程语言，弥补了 MATLAB 的所有缺陷，但我很难找到一个 2D 绘图包

## 2.4 Codecs库

由于python中默认的编码是ascii，如果直接使用open方法得到文件对象然后进行文件的读写，都将无法使用包含中文字符（以及其他非ascii码字符），因此建议使用utf-8编码。

python的内部是使用unicode来处理的。而codecs专门用作编码转，通过它的接口是可以扩展到其他关于代码方面的转换。在python3.x中可以直接将bytes类型的数据转换成其他编码格式，而不用手动先转成unicode。codecs模块为我们解决的字符编码的处理提供了lookup方法，它接受一个字符编码名称的参数，并返回指定字符编码对应的 encoder、decoder、StreamReader和StreamWriter的函数对象和类对象的引用。为了简化对lookup方法的调用， codecs还提供了getencoder(encoding)、getdecoder(encoding)、getreader(encoding)和 getwriter(encoding)方法；进一步，简化对特定字符编码的StreamReader、StreamWriter和 StreamReaderWriter的访问，codecs更直接地提供了open方法，通过encoding参数传递字符编码名称，即可获得对 encoder和decoder的双向服务。

## 2.5 jieba库

jieba库的作用有：①分词。可使用 **jieba.cut** 和 **jieba.cut\_for\_search** 方法进行分词，两者所返回的结构都是一个**可迭代**的 generator，可使用 for 循环来获得分词后得到的每一个词语（unicode），或者直接使用 **jieba.lcut** 以及 **jieba.lcut\_for\_search** 直接返回 list。②添加自定义词典，可以载入以及调整词典。③关键词提取。 ④词性提取。⑤词性标注。⑥并行分词。⑦返回词语在原文的起止位置。

# 3.数据分析

## 3.1 获取数据

数据的作用主要表现在数据的分析使用。通过对信息数据进行分析，不仅能把隐藏的数据挖掘出来，还能通过这些隐藏的讯息，挖掘出有利于决策的信息。

我们的项目通过爬虫，从链家网爬取相关的系列商品，爬取的信息包括房子简介、房租、面积、户型、参考图的网站。这些爬取的信息将作为后续数据分析的数据。

1. 获取结果：



图 3.1数据爬取数据列表

（2）相关代码如下：数据爬取加数据清洗

import requests  
from lxml import etree  
import pymongo  
import time  
from openpyxl import Workbook  
  
data\_list = []  
  
mongo\_py = pymongo.MongoClient()  
  
collection = mongo\_py['lianjia\_sh']['data']  
  
  
def get\_response(url):  
 head = {  
 "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/86.0.4240.75 Safari/537.36"  
 }  
 response = requests.get(url, headers=head)  
 if response.status\_code == 200:  
 return response.text  
 else:  
 print("访问网页错误")  
  
  
def parse\_data(response):  
 response = etree.HTML(response)  
 datas = response.xpath('//\*[@id="content"]/div[1]/div[1]/div')  
 for data in datas:  
 title = data.xpath('./div/p[1]/a/text()')[0].replace(" ", "").replace("\n", "")  
 price\_one = data.xpath('./div/span/em/text()')[0]  
 price\_two = data.xpath('./div/span/text()')[0]  
 price = (price\_one + price\_two).replace(" ", "")  
 for i in range(1, 8):  
 area = data.xpath('./div/p[2]/text()[' + str(i) + ']')  
 if area == []:  
 continue  
 else:  
 area = area[0].replace(" ", "")  
 if "㎡" in area:  
 break  
 else:  
 continue  
 area = area.replace("\n", "")  
 for i in range(1, 8):  
 house = data.xpath('./div/p[2]/text()[' + str(i) + ']')  
 if house == []:  
 continue  
 else:  
 house = house[0].replace(" ", "")  
 if "室" in house:  
 break  
 else:  
 continue  
 house = house.replace("\n", "")  
 img\_url = data.xpath('./a/img/@data-src')[0]  
 data\_list.append([title, price, area, house, img\_url])  
  
  
def main():  
 for i in range(1, 101):  
 print(i)  
 # 因为网站有限制,所以要爬4个网站  
 url1 = "https://sh.lianjia.com/zufang/pg"+str(i)+"rp1rp2rp3rp4"  
 url2 = "https://sh.lianjia.com/zufang/pg"+str(i)+"rp5"  
 url3 = "https://sh.lianjia.com/zufang/pg"+str(i)+"rp6"  
 url4 = "https://sh.lianjia.com/zufang/pg"+str(i)+"rp7"  
 time.sleep(2)  
 response = get\_response(url4)  
 parse\_data(response)  
 print(data\_list)  
  
 # 把数据保存到excel里面  
 workbook = Workbook()  
 sheet = workbook.active  
 sheet.append(["简介", "房租", "面积", "户型", "img\_url"])  
 for i in data\_list:  
 sheet.append(i)  
 workbook.save('lianjia.xls')  
  
  
if \_\_name\_\_ == '\_\_main\_\_':  
 main()

把爬取的数据保存在lianjia.xls文件中

## 3.2 分析总结

### 3.2.1利用jieba库来分析房子简介信息高频关键字以及关键词出现次数，并根据关键字频率生成词云

import xlrd  
from matplotlib import pyplot as plt  
from wordcloud import WordCloud  
from PIL import Image  
import numpy as np  
import jieba  
  
  
def wordcloud(text):  
 # 分词  
 cut = jieba.cut(text)  
 string = " ".join(cut)  
 # 打开遮罩图片  
 img = Image.open(r"prcie.jpg")  
 # 将图片转换为数组  
 img\_array = np.array(img)  
 # 设置参数  
 wc = WordCloud(  
 background\_color='white',  
 mask=img\_array,  
 font\_path="msyh.ttc"  
 )  
 wc.generate\_from\_text(string)  
 fig = plt.figure(1)  
 plt.imshow(wc)  
 # 是否显示坐标轴  
 plt.axis('off')  
 # 显示生成的词云图片,可以直接保存就注释了  
 # plt.show()  
 # 保存  
 plt.savefig(r'wordcloud.jpg', dpi=500)  
  
data\_list = []  
wb = xlrd.open\_workbook("lianjia.xls")  
sh = wb.sheet\_by\_name('Sheet')  
for i in range(1, 10201):  
 data\_list.append(sh.cell(i, 0).value)  
text = "".join(data\_list)  
wordcloud(text)



图3.2.1关键词词云

### 3.2.2分析房子面积的数量

1. 代码如下：

import pandas as pd  
import numpy as np  
from matplotlib import pyplot as plt  
import xlrd  
from pylab import mpl  
  
# 图形显示中文设置  
mpl.rcParams['font.sans-serif'] = ['SimHei']  
  
file\_path = 'lianjia.xls'  
  
data = pd.read\_excel(file\_path)  
# 设置显示全部列(字段)  
pd.set\_option('display.max\_columns', None)  
# print(data.head())  
# print(data.info())  
area\_data = data['面积']  
final\_data = []  
for i in area\_data:  
 if (i.find('-') != -1):  
 final\_data.append(int(i.split('-')[0]))  
 else:  
 i = int(i.replace('㎡', ''))  
 final\_data.append(i)  
  
# 计算极差  
area\_range = max(final\_data) - min(final\_data)  
print(area\_range)  
  
# 设置图形大小  
plt.figure(figsize=(20, 8), dpi=80)  
# 设置组距  
d = 30  
# 绘图  
plt.hist(final\_data, range(min(final\_data), max(final\_data) + d, d))  
# 设置x的刻度  
plt.xticks(range(min(final\_data), max(final\_data) + d, d))  
# 设置y的刻度  
y\_list = [500, 1000, 1500, 2000, 2500, 3000]  
plt.yticks(y\_list)  
  
# 添加标题  
plt.title('链家住房面积分布图')  
plt.xlabel('面积(㎡)')  
plt.ylabel('数量(个)')  
  
# 绘制网格  
plt.grid(alpha=0.3)  
# 保存  
plt.savefig('lianjia.jpg')  
plt.show()

1. 柱状图展示：

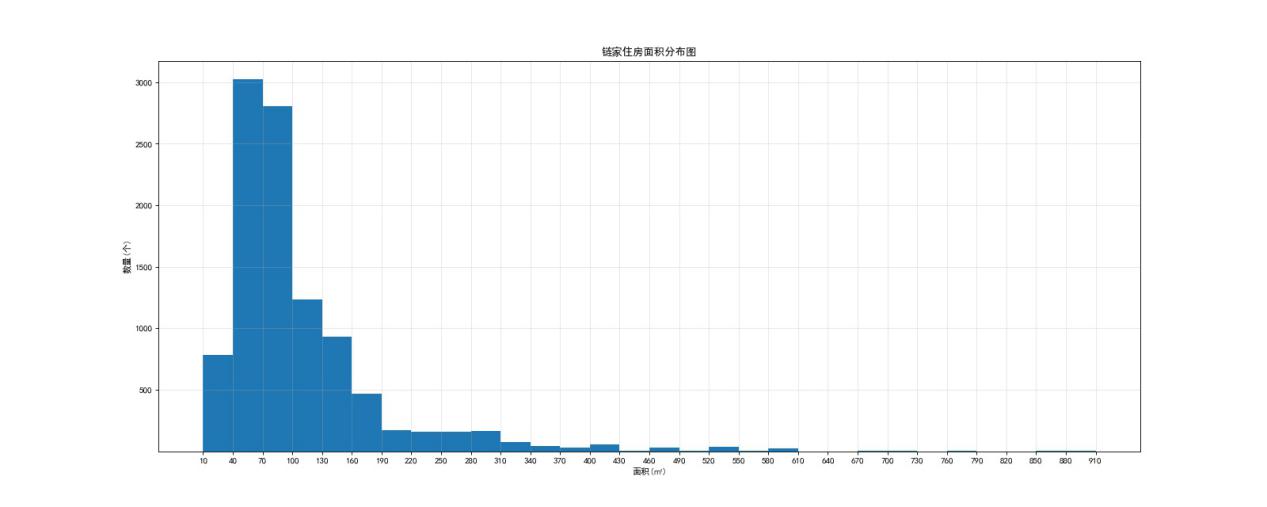


图3.2.2-柱状图

### 3.2.3分析房租的数量

根据数据生成房租价格范围，可以给网民一个大概的价格估摸范围

（1）代码如下：

import random  
import pandas as pd  
import numpy as np  
from matplotlib import pyplot as plt  
from pylab import mpl  
  
# 图形显示中文设置  
mpl.rcParams['font.sans-serif'] = ['SimHei']  
  
file\_path = 'lianjia.xls'  
  
data = pd.read\_excel(file\_path)  
# 设置显示全部列(字段)  
pd.set\_option('display.max\_columns', None)  
# print(data.head())  
price\_data = data['房租']  
  
final\_data = []  
for i in price\_data:  
 if (i.find('-') != -1):  
 final\_data.append(int(i.split('-')[0]))  
 else:  
 i = int(i.replace('元/月', ''))  
 final\_data.append(i)  
  
# 将列表随机打乱(random:随机 shuffle:打乱)  
random.shuffle(final\_data)  
  
# 设置图片大小  
plt.figure(figsize=(20, 8), dpi=80)  
plt.scatter(list(range(len(final\_data)))[::5], final\_data[::5])  
plt.title('链家房租价格分布图')  
plt.xlabel('编号')  
plt.ylabel('价格(元/月)')  
plt.grid(alpha=0.3)  
plt.savefig('price\_lianjia.jpg')  
plt.show()

（2）生成散点图：

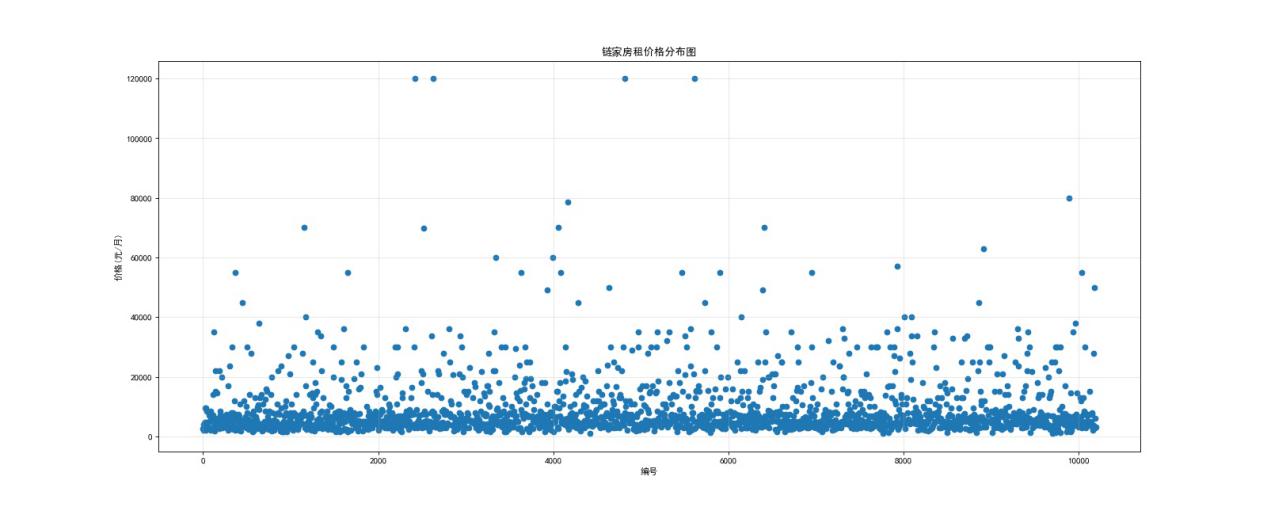


图3.2.3房租-散点图

分析小结：

1.由词云图可以看出,上海租房都是整租，没有合租

2.由柱状图可以看出，40平方米到70平方米的数量最多，70平方米到100平方米的排第二，大部分都在160平方米以下。

3.从散点图可以很直观地看出，大部分价格都在20000以下

# 

# 4总结与展望

通过对本次链家网数据的爬取以及数据的分析，可以看出，爬虫爬取的数据是有一定的价值的，可以人们对租房有一个基本的认识与选择。使用 Xpath等手段,进行爬取数据同时，更进一步巩固了我们Python的知识学习，及其应用。 在获取过程中，由于网站有所爬取限制,所以选用爬取4个网页来得到全部的具体的数据才得以解决问题。在本项目开始指出，我们观看了一些爬虫案例,发现了把数据先保存到数据库,再导出到excel中比较保险。但是,由于数据量不是很多,所以就直接保存到excel中了。

另外，在使用pandas的库的时候读取不了xlsx文件的数据,通过查阅网上的资料发现是xlrd库的版本不支持,所有需要卸载xlrd库,重新安装一个低版本的xlrd才能正常运行。

最后，我们发现，爬取主图的地址下来可以更好更为直观的观察房子的样式，也为网名增加了更为方便的查看和选择途径。