爬虫步骤:

1. 用requets.get()方法获取网页源代码的响应(把user-agent添加到参数中,模仿浏览器行为)
2. 利用xpath解析网页,获取自己想要获取的数据(可以进行数据清理,比如去掉面积的单位)
3. 可以设置一个循环,来循环爬取网页的数据(按页爬取,可以添加time.sleep防止ip被服务器封掉)
4. 存储到excel中或者数据库中

爬虫

import requests  
from lxml import etree  
import pymongo  
import time  
from openpyxl import Workbook  
  
# 定义一个列表来存储下面要获取的数据  
data\_list = []  
# 连接mongodb需要用到的  
mongo\_py = pymongo.MongoClient()  
  
collection = mongo\_py['lianjia\_sh']['data']  
  
# 获取响应的方法(需要传入地址参数)  
def get\_response(url):  
 # 请求头添加user-agent，模仿浏览器（让服务器不知道这是爬虫程序，以为是浏览器）  
 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"  
 }  
 # requests的get方法获取url的响应  
 response = requests.get(url, headers=head)  
 # 响应码为200就返回网页源代码（response.text）  
 if response.status\_code == 200:  
 return response.text  
 else:  
 # 否则就返回访问错误  
 print("访问网页错误")  
  
# 解析响应的内容,获取自己想要获取的内容(传入响应的参数)  
def parse\_data(response):  
 # 利用lxml库的etree模块解析网页的结构,生成xpath可以解析的对象  
 response = etree.HTML(response)  
 # 使用xpath解析网页  
 datas = response.xpath('//\*[@id="content"]/div[1]/div[1]/div')  
 for data in datas:  
 # 获取自己想要的数据  
 # 简介  
 title = data.xpath('./div/p[1]/a/text()')[0].replace(" ", "").replace("\n", "")  
 # 月租的具体的数值(比如2400)  
 price\_one = data.xpath('./div/span/em/text()')[0]  
 # 月租的单位(比如元/每月)  
 # 把月租的具体的数值和单位分开,防止数据发生不必要的错误  
 price\_two = data.xpath('./div/span/text()')[0]  
 # 合并月租的具体数值和单位  
 price = (price\_one + price\_two).replace(" ", "")  
 # 循环七次,因为下面的xpath包含七个值(我们只需要获取我们想要的值即可)  
 for i in range(1, 8):  
 # 获取area(面积)的值(因为有些面积数据为空,所以下面要对面积的只进行判断)  
 area = data.xpath('./div/p[2]/text()[' + str(i) + ']')  
 # 如果为空(有些网页上也没有数据),继续循环  
 if area == []:  
 continue  
 else:  
 # 如果不为空,就先替换空格,  
 area = area[0].replace(" ", "")  
 # 如果㎡在里面,就说明找到了我们想要的数据,就退出循环  
 if "㎡" in area:  
 break  
 # 否则就继续  
 else:  
 continue  
 # 换行去掉(有些数据)  
 area = area.replace("\n", "")  
 # house:(房屋户型)这个数据和面积数据一样(也是在七个数据之间的一个数据,做法和面积一样)  
 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", "")  
 # 通过xpath获取图片  
 img\_url = data.xpath('./a/img/@data-src')[0]  
 # 最后把各个数据(字段)添加到data\_list列表中  
 data\_list.append([title, price, area, house, img\_url])  
  
  
def main():  
 # 循环100次,链家只能爬取到第100页  
 for i in range(1, 101):  
 print(i)  
 # 因为网站有限制,所以要爬4个页面(有点数据需要点进去(点击图片或者点击文字)才能爬取)  
 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"  
 # 每循环一次,等两秒再进行下一次循环,防止网页封ip  
 time.sleep(2)  
 # 传入地址参数  
 response = get\_response(url4)  
 # 解析获取的响应  
 parse\_data(response)  
 print(data\_list)  
  
 # 把数据保存到mongodb数据库  
 for i in range(len(data\_list)):  
 collection.insert\_one({  
 "简介": data\_list[i][0],  
 "房租": data\_list[i][1],  
 "面积": data\_list[i][2],  
 "户型": data\_list[i][3],  
 "img\_url": data\_list[i][4]  
 })  
  
 # 把数据保存到excel里面  
 workbook = Workbook()  
 sheet = workbook.active  
 # 先把字段名添加到最开头  
 sheet.append(["简介", "房租", "面积", "户型", "img\_url"])  
 # 把data\_list的数据添加到xls中  
 for i in data\_list:  
 sheet.append(i)  
 # 保存到xls文件中,文件名为lianjia  
 workbook.save('lianjia.xls')  
  
  
if \_\_name\_\_ == '\_\_main\_\_':  
 main()

直方图

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']  
  
# 文件路径  
file\_path = 'lianjia.xls'  
# 使用pandas读取lianjia.xls  
data = pd.read\_excel(file\_path)  
# 设置显示全部列(字段),pandas只会显示部分列(字段)  
pd.set\_option('display.max\_columns', None)  
# print(data.head()) 查看前5条数据  
# print(data.info()) 查看数据的信息,看有没有缺失值之类的  
# 获取面积这一列(字段)  
area\_data = data['面积']  
# 使用列表存储面积的数值,下面需要把单位去掉(方便制作直方图)  
final\_data = []  
for i in area\_data:  
 # 如果有-按-分隔为列表(有的面积是范围)  
 if (i.find('-') != -1):  
 # 取-前面的数据,  
 final\_data.append(int(i.split('-')[0]))  
 else:  
 # 如果没有-就直接去掉面积的单位  
 i = int(i.replace('㎡', ''))  
 # 添加到列表中  
 final\_data.append(i)  
  
# 计算极差，看大概需要画多少个竖条行  
area\_range = max(final\_data) - min(final\_data)  
print(area\_range)  
  
# 设置图形大小20为宽，8为高，像素（dpi）  
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，透明度（alpha）  
plt.grid(alpha=0.3)  
# 保存  
plt.savefig('lianjia.jpg')  
plt.show()

散点图

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']  
  
# 文件路径  
file\_path = 'lianjia.xls'  
# 利用pandas读取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()) 显示前五行  
# 获取房租列(字段)  
price\_data = data['房租']  
  
# 创建一个列表存放房租数值的数据  
final\_data = []  
for i in price\_data:  
 # 如果有-就按-分隔,并取前面的数据(例如有的数据是4000-5000元/月)  
 if (i.find('-') != -1):  
 final\_data.append(int(i.split('-')[0]))  
 else:  
 # 其他的就直接去掉元/月方便画图  
 i = int(i.replace('元/月', ''))  
 final\_data.append(i)  
  
# 将列表随机打乱(random:随机 shuffle:打乱),你也可以试一下不打乱  
random.shuffle(final\_data)  
  
# 设置图片大小20为宽,8为高,dpi(像素)  
plt.figure(figsize=(20, 8), dpi=80)  
# 每隔五个取一个值,因为值太多了,点就重合了,就不太好看  
# 绘制散点图  
plt.scatter(list(range(len(final\_data)))[::5], final\_data[::5])  
# 添加标题和标签  
plt.title('链家房租价格分布图')  
plt.xlabel('编号')  
plt.ylabel('价格(元/月)')  
# 添加网格并设置透明度alpha  
plt.grid(alpha=0.3)  
# 保存图片为price\_lianjia.jpg  
plt.savefig('price\_lianjia.jpg')  
plt.show()