学号 专业 姓名

理论课任务2

【任务名称】 用MATLAB分析房价

【任务目的】

在链家网上，用matlab抓取某一城市某一区的小区价格

利用直方图函数 histogram() 分析小区价格的分布，并美化图形

【源码】

clc;close all;clear;

Data={};

for pageNo=1:99 %读取1-99页

clc;fprintf("%d\n",pageNo);

str=urlread(['https://hf.lianjia.com/ershoufang/pg',num2str(pageNo),'/']);

price=regexp(str,'(?<=totalPrice totalPrice2"><i> </i><span class="">).\*?(?=</span><i>万</i></div><div class=")','match');

%爬取Price

Data=[Data;price'];%把每一页的Price加入到Data中

end

len=length(Data);

Data1=zeros(len,1);

for i=1:len

Data1(i)=str2num(cell2mat(Data(i))); %把cell型的数据转换成double型

end

xlswrite('E:\Matlab\111\ClassAssignment\Price\_Data.xlsx',Data1); %把数据写入excel表中

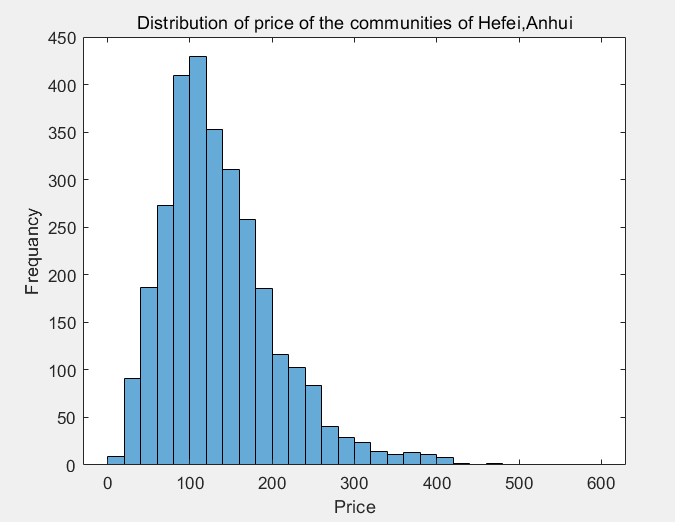
h = histogram(Data1); %绘制直方图

ylabel('Frequancy');

xlabel('Price');

title('Distribution of price of the communities of Hefei,Anhui');%图形美化

【运行结果截图】



【任务总结】

有实验课任务的经验与PPT上已有的代码，可以较简洁的完成本任务，但是题目要求是爬取小区价格，本程序爬取成了房价，不过基本的思想是一致的，只是相关的网址和正则匹配部分需要更改；

与实验课任务一样，本程序依然需要用cell2mat将抓取到的保存再元胞数组中的数据转化成矩阵，再通过str2num转化成字符串，否则就一直出现矩阵对不齐的报错，目前我没有找到其他很好的解决本问题的方法；

在用正则表达式进行匹配时，可以在网页源码中用ctr+f进行寻找，如果正好一些字符串可以找到30个（每页网页由30个数据），那么大概率就没有问题；

本程序在运行较慢，每次报错之后再修改都再次运行都需要花一些时间，所以最好先自己检查一下代码逻辑，减少不必要的报错以节约时间。