学号 专业 姓名

实验日期 教师签字 成绩

实验报告

【实验名称】 数据处理

【实验目的】

用matlab从链家上爬取某地区二手房信息（各小区的名称、价格及坐标），将所有信息绘制在坐标图中

【实验原理】

正则表达式，matlab处理文件等部分内容

【源码】

clc;close all;clear;

Data = {};

address1 = {};

address2 = {};

x = [];

y = [];

for pageNo = 1:1 %为方便展示，只读取一页信息

try

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

str = urlread(['https://nj.lianjia.com/xiaoqu/',num2str(pageNo),'/']);

Number = regexp(str, '(?<=data-housecode=").{13,13}','match');

name = regexp(str,'(?<=<div class="title">\n\s{6}<a href=).\*?(?=</div>)','match');

name = cellfun(@(s)s(s>999),name,"UniformOutput",false);

temp = regexp(str,'(?<=<span class="positionIcon"></span>).\*?(?=</div>)','match');

loc = cell(size(temp,1),2);

for i = 1:length(temp)

loc(i,:) = regexp(temp{i},'(?<=>)[^\n].\*?(?=<)','match');

end

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

str2 = urlread('https://nj.lianjia.com/xiaoqu/');

Number = regexp(str2, '(?<=data-housecode=").\*?(?=")','match'); %找到每个小区对应的网址

for fNo = 1:30

clc;

fprintf("%d\n",fNo)

ulr = ['https://nj.lianjia.com/xiaoqu/',strjoin(cellstr(Number(fNo))),'/']; %获取对应网址的信息

str1 = urlread(ulr);

lattitude1 = regexp(str1,'(?<=" xiaoqu="[).\*?(?=,)','match');

lattitude2 = regexp(str1,'(?<=,).{7,17}(?=]" class=")','match'); %找到对应的经纬度

if isempty(lattitude1)

address1 = [address1;'0'];

else

address1 = [address1;lattitude1'];

end

if isempty(lattitude2)

address2 = [address2;'0'];

else

address2 = [address2;lattitude2'];

end

end

Data = [Data;name',loc,price',address1,address2];

catch

break;

end

end

fprintf("Data读取完成");

for j = 1:length(Data)

x(end+1) = str2num(cell2mat(address1(j)));%把元胞数组转化为矩阵，在把矩阵中数字转化为字符串

end

for j = 1:length(Data)

y(end+1) = str2num(cell2mat(address2(j)));

end

plot(x,y,'ok');

text(x,y,price);

xticks(118.7:0.1:119.1);

yticks(31.5:0.1:32.3);

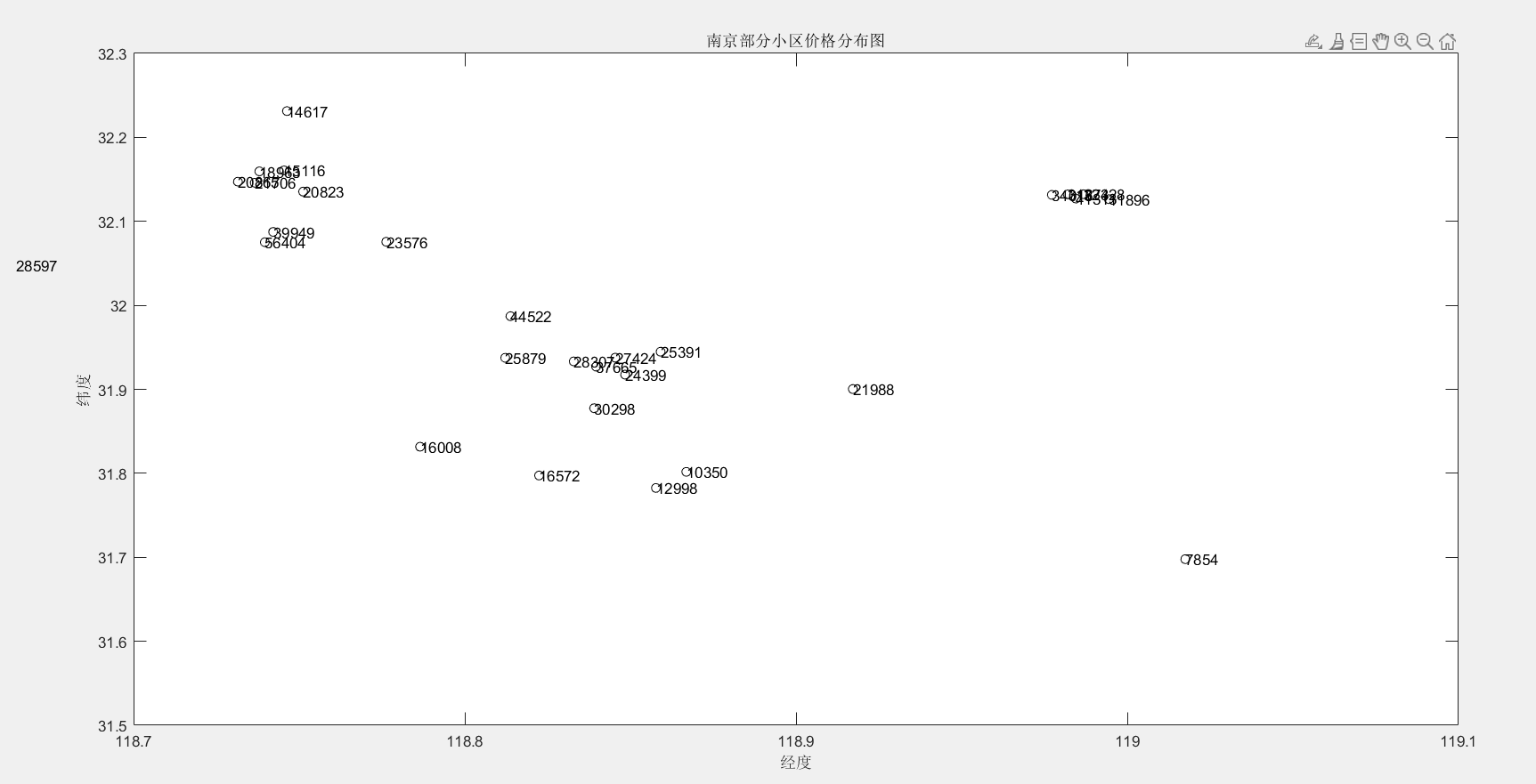
set(gca,'XLim',[118.7,119.1],'YLim',[31.5,32.3]);%定义x，y坐标边界

xlabel('经度');

ylabel('纬度');

title('南京部分小区价格分布图');

【运行结果截图】



【实验总结】

根据PPT上已有的代码，再补充上抓取每个小区独立网页的URL，在每个小区独立网页抓取该小区的经纬度数据即可完成本任务；

比较繁琐的地方是，抓取来的字符串放在元胞数组中，需要用cell2mat将其转化成矩阵，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。