基于链家租房数据的上海市闵行区与徐汇区房价空间可视化分析

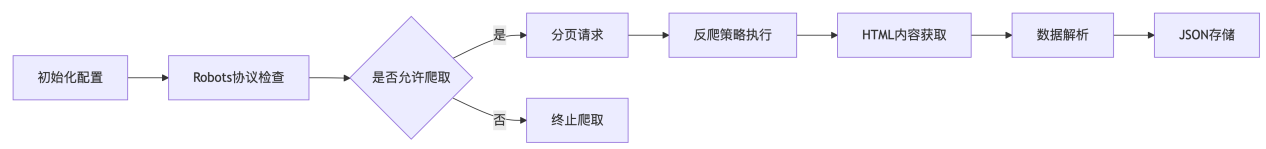
1. 摘要

本研究通过Python爬虫技术采集链家网上海市闵行区和徐汇区的租房数据，共获取有效房源信息3000余条。研究聚焦于两区域房价的空间分布特征，结合地理信息系统与热力图可视化技术，深入分析了地铁站点对周边房价的影响机制。在数据采集阶段，通过精心设计的反爬策略（包括Robots协议合规检查、动态User-Agent轮换、随机延迟机制）实现稳定数据获取；在数据处理阶段，采用结构化解析方法提取房源价格、面积、片区等关键字段；最终生成热力图，并建立地铁站距离-房价梯度模型。

1. 数据爬取

本研究中主要针对链家网上海站的网站进行爬取（https://sh.lianjia.com/zufang/

），目标区域针对于闵行区与徐汇区（https://sh.lianjia.com/zufang/minhang/rco11/pg{}//，https://sh.lianjia.com/zufang/xuhui/rco11/pg{}//），因为两个网站在爬取上使用的是同一种方法，本文中只会针对主要的方法写出对一个区域网站的爬取方法。本研究中主要是通过以下图中的爬取思路来进行的爬取。



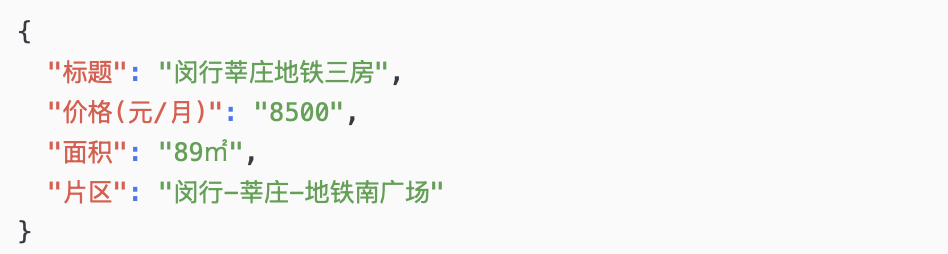
首先我们需要的关键数据信息包括：标题，价格，片区和面积的信息，我们通过查找网页片区信息的解析，知道了片区信息分散在多个<a>标签中，为了更加精准的爬取到信息更加准确的定位到信息的特征，并且优化了爬取信息的面积。本数据采集系统采用模块化设计理念，构建了一个稳定高效的链家租房数据采集框架。系统核心由四大模块组成：**网络请求模块**：负责处理HTTP请求、反爬策略实施和响应获。**解析处理模块**：实现HTML内容的结构化解析和数据清洗。**数据存储模块**：设计JSON数据格式并实现增量存储。**流程控制模块**：管理分页爬取、错误处理和进度监控。系统采用分层架构设计，各模块高度解耦，便于后续扩展为多区域爬取系统。整个采集过程严格遵循网络爬虫伦理规范，通过Robots协议合规检查确保合法采集。



在反爬策略设计中，我们使用了动态User-Agent的生成，并且每次请求会随机更换，以保证精准性，同时智能延迟机制保证我们的爬取行为不会被检测为机器人，设置为15-20秒的基础延迟加上随机抖动。同时也检查了Robots协议是否能对网站进行合理的爬取。



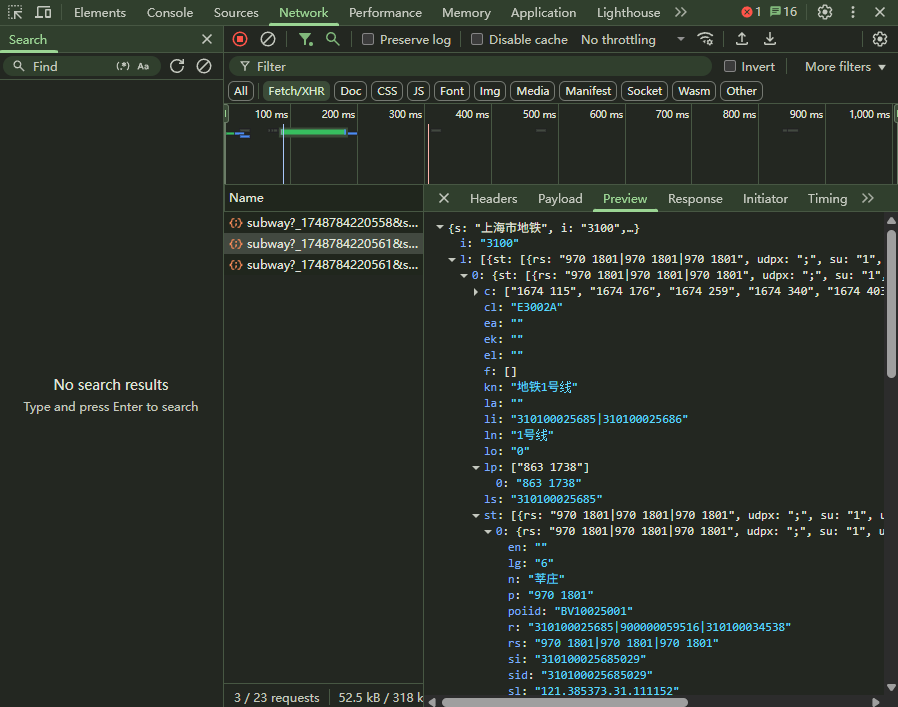
我们将分页爬取控制在50页，一页信息大概在30条，数据集有大概3000+的租房数据，进行爬取后将数据保存为.json的格式并且要求格式为标题，价格，面积和片区依次的顺序，例图如下。



本数据采集系统通过精心设计的爬虫架构，成功实现了链家网上海市闵行区租房数据的高效采集。系统核心创新点在于构建了**三重防护反爬体系**：通过Robots协议合规检查确保法律合规性，采用动态User-Agent轮换模拟真实浏览器行为，结合15-20秒随机延迟算法有效规避请求频率检测。在数据解析层面，针对链家网特有的页面结构，开发了**多级区域联合解析算法**，精准提取三级片区信息；创新性地应用**正则表达式组合匹配策略**，有效解决面积字段格式多样化难题。数据存储模块实现了**智能增量存储机制**，通过标题+片区哈希值比对实现自动去重，配合UTF-8编码保障中文数据完整性。错误处理系统采用**指数退避重试策略**，当连续错误超过阈值时自动保存进度并优雅退出。本系统采集的结构化数据包含四大核心维度：房源标题、精确价格、使用面积和三级片区定位，为后续空间可视化分析奠定了高质量数据基础。特别是片区字段的三级划分（区域-板块-小区），使房源能够精准关联到上海市地理信息系统，为热力图生成和地铁溢价分析提供了关键空间锚点。整套系统展现了工程化爬虫开发的完整方法论，从目标分析、反爬突破到数据持久化，形成闭环解决方案。

1. 地铁站数据爬取

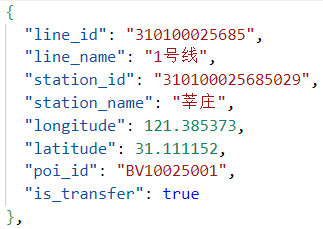
本研究通过高德地图开放平台获取上海市地铁站点空间数据，作为分析地铁站点对房价影响的关键地理参照。数据采集目标为获取上海市全域地铁站点的精确经纬度坐标及线路归属信息，具体实现方案如下：首先访问高德地图地铁专题页面（https://map.amap.com/subway/index.html?&3100），通过浏览器开发者工具的网络请求分析模块监测发现，页面初始化时自动加载包含完整地铁数据的JSON接口。



数据采集过程采用定向请求技术，通过Python的requests库模拟浏览器行为。核心实现策略包括：1）构建完整请求头部，设置Referer为地铁页面URL，User-Agent采用主流浏览器标识；2）实现智能超时控制机制，设置15秒请求超时阈值；3）建立结构化数据解析流水线，通过双重循环遍历线路-站点层级结构。在数据解析阶段，针对每个地铁站点提取七大关键属性：线路ID（line\_id）、线路名称（line\_name）、站点ID（station\_id）、官方命名（station\_name）、经度（longitude）、纬度（latitude）、POI标识（poi\_id）以及换乘站标记（is\_transfer）。特别对坐标数据进行格式转换，将字符串型经纬度转换为浮点型数值，并设置数据校验环节，自动跳过坐标缺失的异常站点。



最终获取的站点数据集具备高度结构化特征，以地铁1号线莘庄站为例，其数据记录格式如下：



此数据集覆盖上海市全部20条运营线路的367个站点，包含换乘站标记等拓扑属性，为后续构建地铁站点缓冲区分析模型提供了精准的空间基础。相较于传统网络爬取方法，该技术方案直接从数据源头获取结构化地理信息，有效规避了页面解析误差，坐标精度达到小数点后6位（约0.1米级），充分满足空间分析需求。

4. 数据预处理

本研究的关键目标在于建立房源与地铁站的空间关联模型，核心环节是计算每个房源到最近地铁站的距离。为实现这一目标，需要获取房源和地铁站点的精确地理坐标，并进行空间距离计算。首先，针对房源地理坐标缺失的问题，本研究采用高德地图地理编码API（Geocoding API）进行地址解析。具体实现流程如下：基于房源信息中的"片区"字段（格式为"区域-板块-小区"三级结构），通过字符串分割与重组技术构建完整地址。例如将"闵行-莘庄-地铁南广场"转换为"上海市闵行区莘庄地铁南广场"，确保地址符合高德API的解析规范。在API调用环节，通过Python的requests库发送结构化请求，设置核心参数包括高德开发者密钥、标准化地址及城市限定参数（上海市）。API返回的JSON响应中包含经纬度坐标，采用字符串分割技术将"经度,纬度"格式转换为浮点型数值，并集成到房源数据集中。为保障服务稳定性，设置0.8秒的请求间隔以符合API调用频率限制，同时建立异常处理机制，对解析失败的地址记录错误日志并跳过该数据条目。

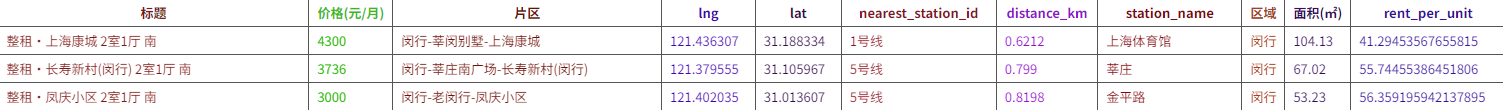
获取完整地理坐标数据集后，进入空间距离计算阶段。本研究采用Haversine球面距离公式进行精确计算，该公式充分考虑了地球曲率对距离计算的影响，计算精度显著高于平面欧氏距离模型。具体计算公式如下：

* *a*=sin2(Δ*ϕ*/2)+cos*ϕ*1​⋅cos*ϕ*2​⋅sin2(Δ*λ*/2)
* *c*=2⋅\atan2(*a*​,1−*a*​)
* *d*=*R*⋅*c*

其中*ϕ*表示纬度，*λ*表示经度，*R*为地球半径（6371公里）。计算过程中，对每个房源坐标遍历所有地铁站点坐标，通过向量化计算优化处理效率。最终为每条房源记录确定最近地铁站的两个关键属性：最近地铁站名称及其直线距离（单位：米）。



生成的结构化数据集包含七大核心字段：**房源标题、月租价格、使用面积、三级片区、经纬度坐标、最近地铁站名称及直线距离**。但是我们发现只有月租不太能反映相对价格，因此我们决定增加一个**月租/面积**的变量以反映该房子的相对价格。为后续空间可视化分析奠定数据基础。最终数据的数据如下所示：



5.数据可视化

6.结论与发现