

题目：基于Python爬虫的全北京小区信息文本分析报告

学年学期：2023-2024学年秋季学期

课程名称：Python软件设计基础

课程编号：2181030049

课程序号：01

任课教师：王潞

姓 名：庄钰莹

学 号：2021218033012

**评分区域（由阅卷老师填写）：**

结课成绩：

总评成绩：

提交时间：2024年1月12日

**基于Python爬虫的全北京小区信息文本分析报告**

摘 要：本文利用Python爬虫技术从链家网搜集了北京地区小区信息，并采用文本分析方法深入研究。通过LDA主题模型分析、TF-IDF关键词提取、TextRank关键词提取方法、K均值多因素聚类分析等文本分析方法，揭示了小区配套设施、命名特点、房价因素等多个方面。研究结果为房地产市场决策和城市规划提供了重要参考，为北京房产市场的特点和潜在市场趋势提供了深入洞察。

关键词：小区；文本分析；爬虫

正 文：

随着城市化的不断发展和人口增长，房地产市场成为一个备受关注的领域。在这个背景下，房产交易信息平台的崛起为购房者提供了更为便捷和全面的选房途径。作为房产交易信息平台的一员，链家网以其知名度和专业性成为房产信息的主要提供者，特别在北京地区拥有丰富的房源信息。本研究以链家网为数据来源，通过爬虫技术获取了大量北京地区小区数据及文本信息，旨在通过深入的文本分析揭示北京小区及介绍的一系列特征和趋势。

**一、网站选取**

链家网是知名的房产信息平台，汇集了丰富的房源信息，覆盖广泛的地区和多样化的房产类型，对于分析北京地区的房产市场具有显著优势。其定期更新和专业性确保所获取数据具有及时性、准确性和可靠性。本研究针对小区信息，主要关注小区名称、销售情况、建筑年份、房价、在售房数、小区介绍等关键数据及文本。通过获取这些数据，全面了解不同小区的特征，为地产市场研究、城市规划和房地产投资决策提供可靠依据。

**二、爬取过程**

为了获得丰富而全面的数据，本研究选择链家网的小区板块进行数据爬取共获取了北京地区7946个小区的相关数据，数据截至日期为2023年12月20日，爬取北京小区介绍2254条，数据截至日期为2024年1月2日，涵盖了东城、西城、朝阳、海淀、丰台等17个区域。爬取过程主要分为以下三个部分：

（一）数据准备阶段

通过驱动打开 Chrome 浏览器并打开链家网站首页，为避免元素的遮挡，通过driver.maximize\_window()将窗口放大。

对页面源码进行分析后，通过Xpath定位到页面上的{小区}元素并点击，具体如图1所示，进入小区对应页面，并通过driver.switch\_to.window(driver.window\_handles[1])定位到最新打开窗口。

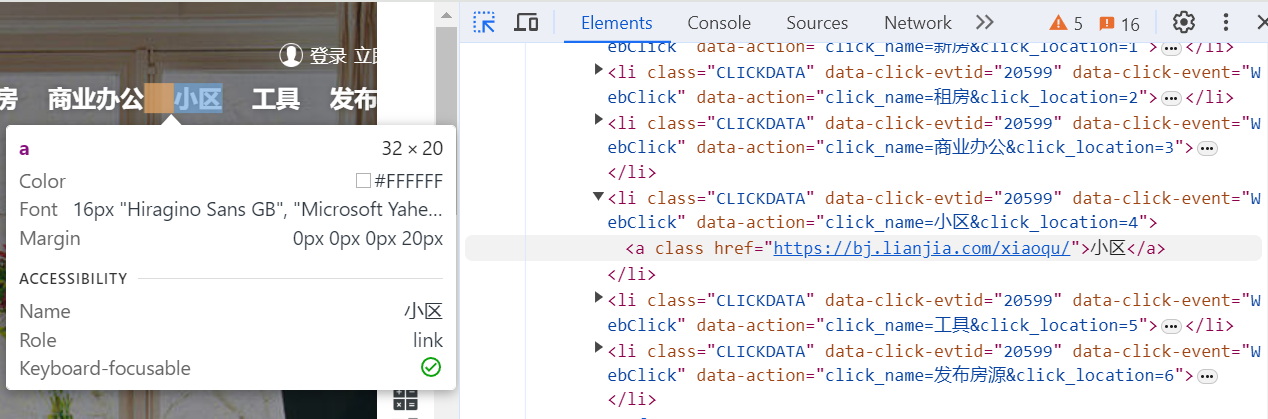


图 1 导航栏定位示意

对小区首页源码进行分析后，通过Xpath路径在用于选择各区划容器中获取跳转“西城”、“东城”等区划链接，具体如图2所示，并将区划及对应链接保存至Excel 文件中备用。

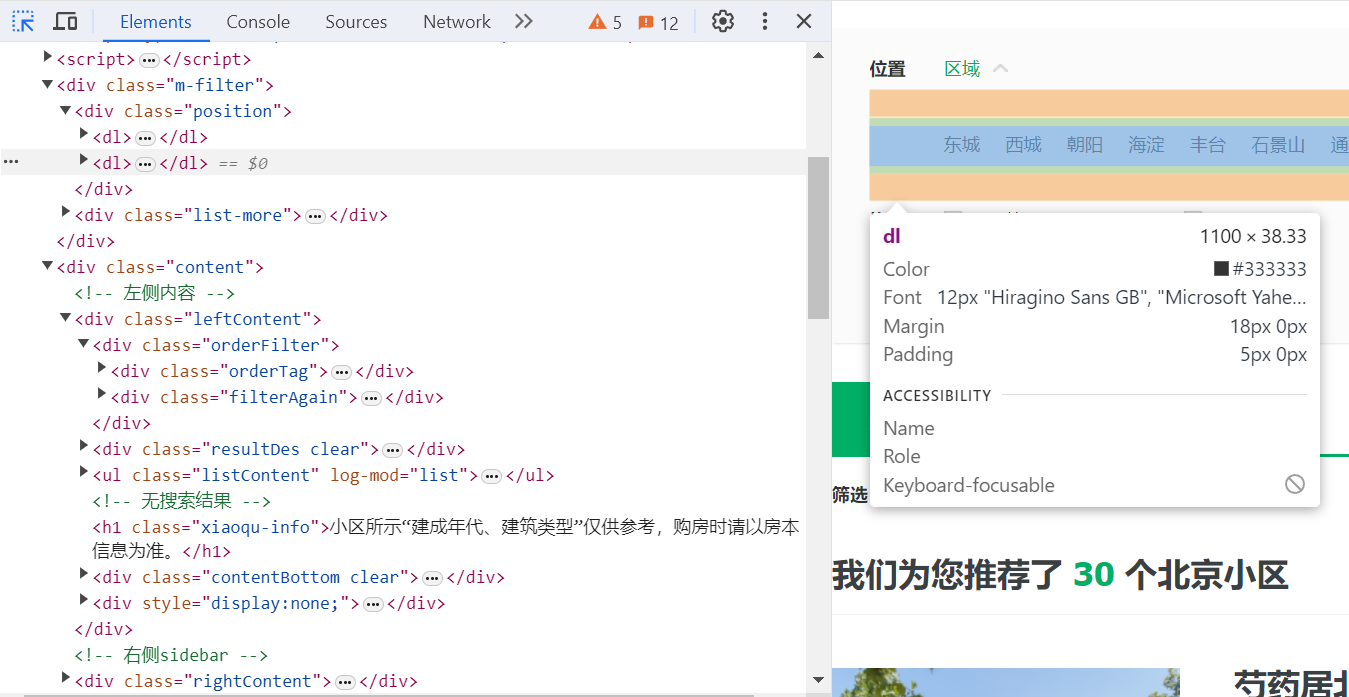


图 2 各区划链接获取示意

（二）数据爬取阶段

从之前保存的 Excel 文件中读取各区划链接列表。并逐个遍历这些链接，发送请求获取小区页面内容并解析，通过css选择器对信息进行定位并提取，例：sale = li.css('.houseInfo a::text').getall()，具体如图3所示。

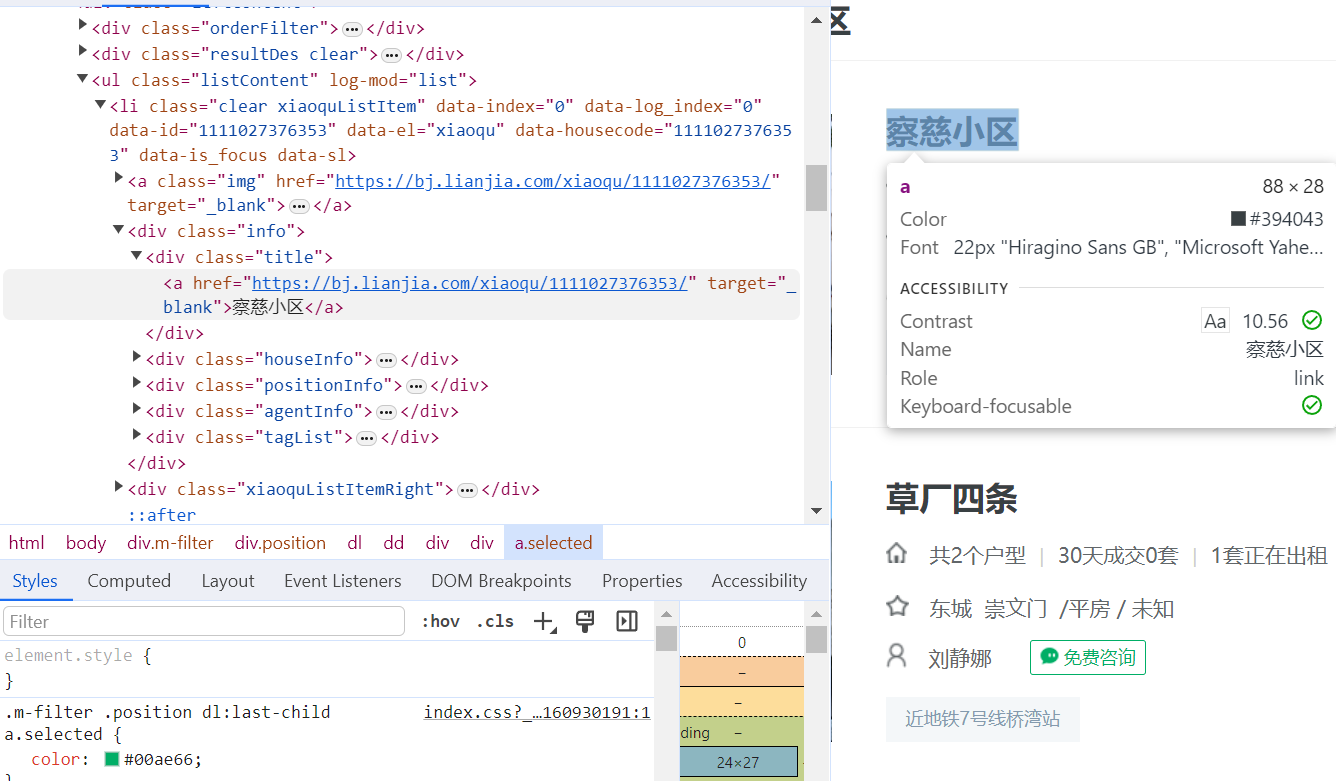


图 3 各小区信息获取示意

由于其中建筑年份与材质位于同一容器下且内含多个空格及换行符，故对其进行数据清洗与拆分。在区划内小区数据提取完毕后，将每个小区的信息存储到列表中，循环各区域网页，并获取每一页的内容，继续提取相关信息，

（三）数据保存阶段

将存储有小区信息的列表转换为 Pandas 的 DataFrame 格式，并保存到另一个Excel 文件中。

完整爬取过程的步骤流程如图4所示，爬取北京小区简介数据的过程与上述步骤基本一致。

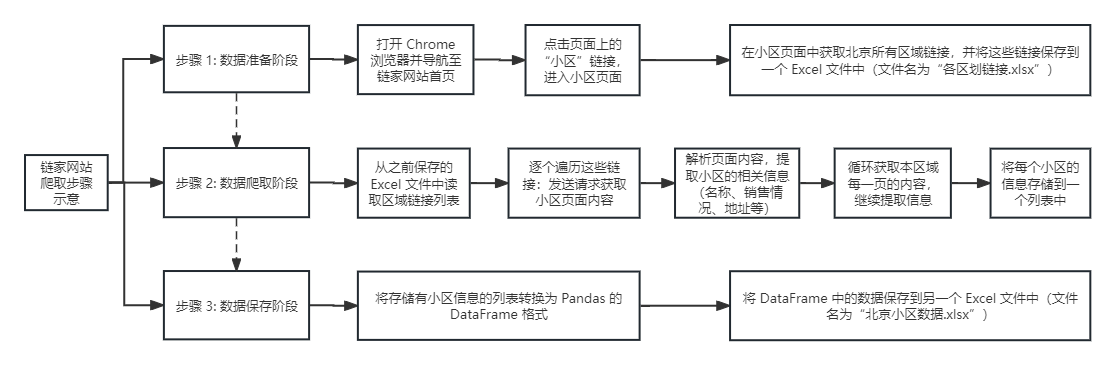


图 4爬取步骤流程图

**三、文本分析方法及过程**

（一）高频词云图

词云图是一种可视化技术，用于突出显示文本中出现频率较高的关键词。根据文本中不同词语的出现频率，词云图将词语按照词频大小进行布局并显示，词频越高的词在图中显示得越大。这种技术可以使浏览者能够快速感知最突出的文字，迅速抓住重点，了解主旨。词云图可用于展示文本数据中的关键词、主题、情感等，帮助人们快速抓住文本数据的重点和特征[[[1]](#endnote-0)]。

（二）TextRank关键词提取

TextRank是一种基于图的排序算法，主要应用于关键词提取和文本摘要。它借鉴了PageRank的思想[[[2]](#endnote-1)]。该算法将词句间的相似关系看成是一种推荐或投票关系，据此构建TextRank网络图，并通过迭代计算得到词句的权重值[[[3]](#endnote-2)]。

（三）TF-IDF关键词提取

本部分主要使用了TF-IDF（词频-逆文档频率）和PCA（主成分分析）两种技术进行文本特征处理和降维可视化。

TF-IDF特征提取：TF-IDF是信息检索和文本挖掘中常用的技术。它通过统计词语在文本中的频率来衡量词的重要性[[[4]](#endnote-3)]。这种方法的优点是能够过滤掉一些常见但不重要的词语，同时强调了稀有但可能更具代表性的词语。

PCA降维：PCA是一种常用的降维技术，用于将高维数据降至低维，同时保留最重要的特征。PCA通过构建一个新的特征空间，将原始数据映射到这个新的低维空间中。

（四）LDA主题模型分析

潜在狄利克雷分配（Latent Dirichlet Allocation，简称LDA）是一种主题模型，由Blei等人于2003年提出。该模型假设每篇文档的主题以概率分布的形式存在。通过对一系列文档进行分析并提取其主题分布，进行主题聚类或文本分类。LDA模型涉及到的数学知识包括二项分布、Gamma函数、Beta分布、多项分布、Dirichlet分布、马尔科夫链、MCMC、Gibbs Sampling、EM算法等[[[5]](#endnote-4)]。

应用LDA方法进行主题模型构建时，主题数量的选取至关重要，能够直接影响模型的识别结果，为了使主题结果更加科学、客观，本研究采用主题一致性和困惑度二者相结合的方式来确定最优LDA主题数。[[[6]](#endnote-5)]运用LDA主题模型抽取的不同主题数量的一致性及困惑度变化趋势如图5、6所示，由图可知，主题数设置为5时，困惑度最低，包含足够多的信息，主题效果较好，因此本研究将LDA模型拟抽取的最终主题数确定为5。

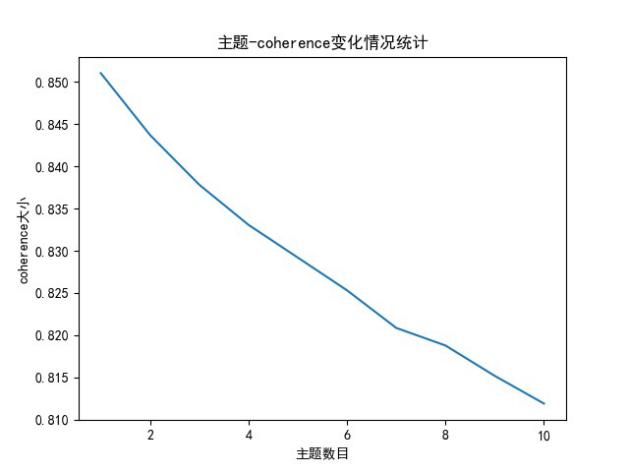
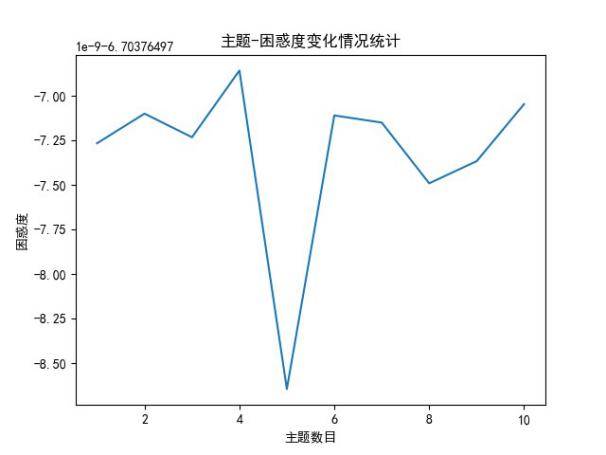


图 5 小区介绍主题困惑度及一致性

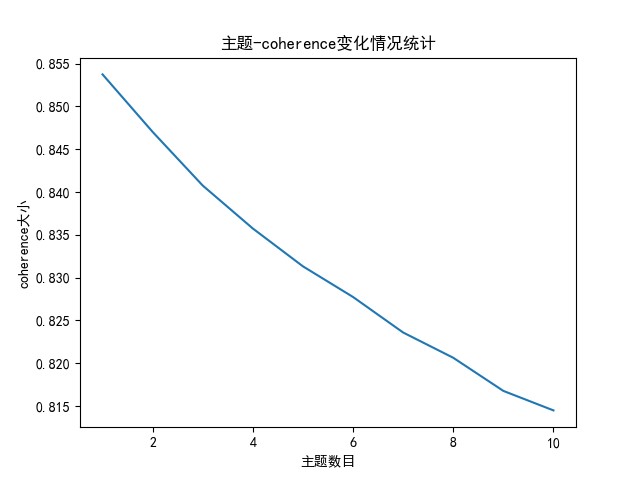
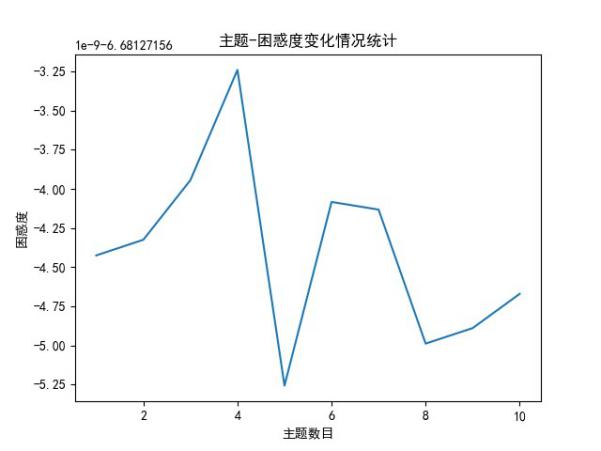


图 6 周边配套主题困惑度及一致性

（五）K均值多因素聚类分析

K均值聚类是一种常用的无监督学习算法，用于将数据点划分到K个不同的簇中，使得同一簇内的数据点彼此相似，不同簇之间的数据点相异。它的功能是将数据分组成K个簇，并且每个数据点都属于其中一个簇，以便更好地理解数据的内在结构和模式。

**四、文本分析结果**

（一）北京小区命名高频词云表



图 7 小区命名高频词云图——类型及用途



图 8 小区命名高频词云图——命名

经过对词频结果的综合分析，将爬取到的小区命名按照频率分为两部分，第一部分主要为建筑类型及用途，如图7所示，第二部分主要为小区名称，如图8所示。本部分文本分析有助于理解地名的多重含义与设计理念。

在第一部分中，出现频率较高的词语主要包括“号院”、“胡同”、“家园”、“小区”、“大厦”、“国际”、“花园”等。这些词语通常用于标识地理位置或指代特定建筑类型。例如，“胡同”通常代表小区具有相对浓厚的历史气息，而“大厦”则一定程度表明小区具备较多的高层建筑。

而在第二部分中，一些突出的词语如“艺境”、“革新”、“光明”、“青春”、“温泉”等，这些词汇往往意味着区域或建筑的独特性或特殊氛围。例如，“艺境”和“创意”可能指示了一个以创意产业或文化艺术为特色的区域，而“革新”和“光明”则可能代表着小区注重展现科技创新和现代性的文化内涵；

（二）TextRank小区介绍关键词提取



图 9 TextRank小区介绍关键词提取词云图

在本研究中，我们运用了TextRank模型对文本进行了深入的分析。结果显示，“小区”（0.028）是文本中的主要关键词，与文本主题契合，表明模型具有一定的准确性。其次是“物业”（0.0089）、“社区”（0.0072）和“公园”（0.0063），这也一定程度上表明了除了小区本身，物业服务、社区活动以及公园或绿地等设施也是消费者对于小区信息关注的重点。这些元素共同构成了一个完整的居住环境，对居民的生活质量有着重要影响。

（三）TF-IDF北京小区命名关键词提取

使用TF-IDF进行关键词提取，排名越靠前表示在小区命名中出现的频率越高，以下词汇在小区命名中频繁出现：

表 1 TF-IDF小区命名高频词

|  |  |  |  |
| --- | --- | --- | --- |
| 顺序 | 词汇 | 顺序 | 词汇 |
| Top1 | 号院 | Top6 | 花园 |
| Top2 | 胡同 | Top7 | 公寓 |
| Top3 | 家园 | Top8 | 大街 |
| Top4 | 小区 | Top9 | 嘉园 |
| Top5 | 大厦 | Top10 | 西里 |

在得出的关键词结果中，一部分词语是描述小区或住宅区域类型的，如“号院”、“胡同”、“家园”、“小区”、“大厦”、“花园”、“公寓”等。另外一些词可能指特定的小区或区域名称，如“嘉园”、“西里”、“国际”等。除此之外，还有一些词则是描述地理位置或区域方位的，如“东区”、“南路”、“西街”、“北区”等，文本关键词分布情况经PCA降维后可视化呈现如图10所示。

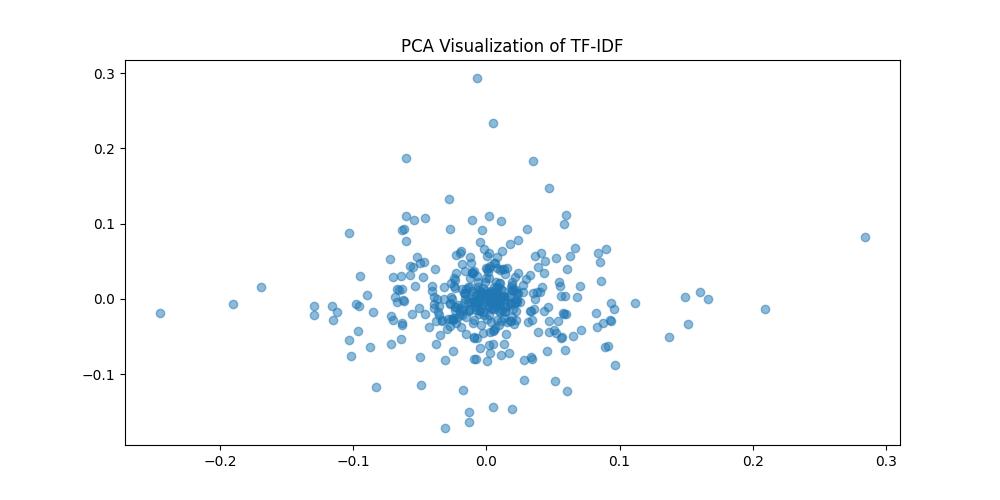


图 10 小区命名TF-IDF关键词分布图

除此之外，使用TF-IDF对链家网中北京部分小区介绍进行关键词提取，得出结果如下表所示：

表 2 TF-IDF小区介绍命名高频词

|  |  |  |  |
| --- | --- | --- | --- |
| 顺序 | 词汇 | 顺序 | 词汇 |
| Top1 | 小区 | Top6 | 建成 |
| Top2 | 社区 | Top7 | 居住 |
| Top3 | 物业 | Top8 | 供暖 |
| Top4 | 物业费 | Top9 | 车位 |
| Top5 | 栋楼 | Top10 | 人车 |

在得出的关键词结果中，常见关键词包括“小区”、“社区”、“物业”及相关信息，物业费、栋楼、建成等词汇也频繁出现，突显了用户在选择小区时对基础设施、物业服务等方面的关注度较高。供暖、车位等词汇也反映了居住舒适性和停车便利性的重要地位。综合而言，通过TF-IDF关键词提取，有助于把握消费者在选择房屋时的主要关切点，关键词分布可视化呈现如图11所示。

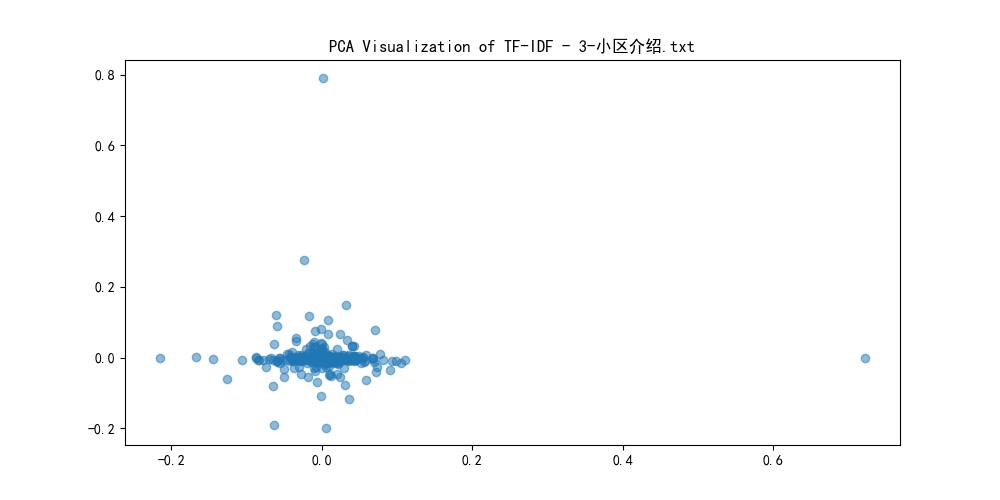


图 11 小区介绍TF-IDF关键词分布图

（四）LDA北京小区介绍主题建模分析

通过计算困惑度及内部一致性将主题数设定为5，并对小区介绍进行LDA主题建模，得出主题分布如下图所示。

表 3 LDA小区介绍各主题分布情况

|  |  |  |
| --- | --- | --- |
| 主题 | 中心词 | 数量占比 |
| 主题一：配套设施 | 广场，单位，安保，建筑，孩子，楼道，物业公司，车库，摄像头，热水，中国银行，建设银行，公摊面积，买菜，氧吧，人口，政府，道路，小桥流水，理发店，永定河，民生银行，亭台，耀莱，京东，商用，蔬果，开发区，北京医院，协和医院，文娱活动，复兴路，首都医科大学 | 20.4% |
| 主题二：建设质量 | 物业，物业费，户型，车位，房，卧室，客厅，阳台，市政，楼间距，便利，单元，业主，面积，分房，两居室，住宅，楼龄，置业，居室，产权，地热，路面，新建，三梯，花草，大院，别墅，公摊，换房，卖房，人口密度，景物，墙体，建房，混合结构，贴砖 | 20.1% |
| 主题三：内部户型 | 购物，成熟，房屋，北京市，门禁卡，人员，规划，宜居，成本，供暖费，安全性，餐厅，外墙，北京城，购房，便利店，楼盘，建筑面积，交易，住宅小区，低密，合围，年龄段，房本，地段，地区，大户型，房率，密度板 | 20.0% |
| 主题四：社区服务 | 绿化，方便，保安，花园，停车位，格局，南门，停车费，外立面，住房，福利，小孩，北区，民水民电，年轻人，入户，暖气片，游泳馆，万达，假山，图书馆，底商，问题，治安，车辆，宽敞，养房，上班族，规模，城建，形式，老年人，公寓，幸福，活动场所，居民楼，草坪，天然气，社区卫生，车位费，治安管理，保障性，生态 | 19.9% |
| 主题五：经济因素 | 化工，风景线，家庭和睦，季度，覆盖率，新贵，材质，定义，主义，新铺，地板，民水，科技，全面，领海，彩虹，干洗店，广电，节省时间，人造草坪，昌盛，有名，活动区，设计院，有色金属，悠久 | 19.7% |

在综合分析各主题内词条，可将其分别命名为配套设施、建设质量、内部户型、社区服务、经济因素，这一定程度上表明了消费者在进行小区选择时，更为倾向于考虑以上五个影响因素。消费者在小区选择中相对更关注配套设施，占小区介绍总文本的20.4%。

具体来看链家网中对于小区配套设施的介绍，可将其总结为就医需求、出行需求、购物需求、教育需求及其他，具体如表4所示：

|  |  |  |
| --- | --- | --- |
| 主题 | 中心词 | 数量占比 |
| 主题一：就医需求 | 医院，老人，产权，医疗，朝南，落地窗，地铁，公交站，散步，公摊面积，协和医院，东四环，老人，百联，宜居，朝阳区，商店，地理位置，邮政局，小孩，交通，电梯，直观，心，宣武医院，太阳，园，金融，农历，康宝路，万寿路，妇幼，实地，河，公交站，健康，电梯，朝南，车位，交通，道路，生态 | 20.2% |
| 主题二：出行需求 | 地图，购房，家乐福，住房，成熟，餐饮，税费，面积，交通，户口，售房，龙湖，年轻，医院，孩子，步行，西南，电梯，楼龄，功能，商圈，功能，布局，健身房，三代人，婚房，环境，自带，孕婴，居民，顺义，玄关，体验，建筑风格，总体，易世达，商店，全明，很大，整洁，充足，高，光线 | 20.0% |
| 主题三：购物需求 | 户型，阳台，南北，配套，契税，商场，楼层，齐全，商品房，客户，交通，通风，正规，落地，居室，森林公园，天街，布局合理，两梯，望京，卖场，交通银行，利用，绿化，公交车站，干净，阳光，商圈，万科，合适，大悦，楼间距，影院，电影院，苏宁，安全，地方，小两口，中信银行，房源，华贸 | 20.0% |
| 主题四：教育需求 | 房子，业主，视野，便利，窗户，北京，家庭，永辉，空间，餐厅，三居室，贷款，清晰，花园，百度，房本，设施，板楼，人群，宽敞，市场，成本，群体，数据，小孩，房龄，房屋产权，经典，车位，早市，门，长阳，户，主题公园，入户，便捷 | 19.9% |
| 主题五：其他需求 | 小区，客厅，格局，号线，社区，菜市场，朝北，舒适，建设银行，环境，低，新，路，华联，百货，诚心，东，楼，换房，南门，北门，购物广场，公房，总医院，一梯，临街，结构，年轻人，全明，朝西，名下，农业银行，肯德基，整体，中国银行，工商，交易，动静，拎包，特莱斯，商城 | 19.9% |

表 4 LDA配套设施各主题分布情况

基于上表可以得出，消费者对小区附近医疗资源的需求相对明显，其中协和医院、医疗、老人、产权等词汇突显了他们对于社区健康服务的重视，特别是针对老年人及家庭的医疗需求，而消费者对于交通便利性的关注表明他们在选择小区时会考虑到工作、购物和其他活动的出行便捷性，对于教育资源的需求体现在对学校、教育机构的关注上。词语商场、超市等，则突显了消费者对于日常生活便利性和购物体验的重视。综上可得，开发商和房地产从业者在规划小区时应全面考虑以上因素，以满足多层次的消费者需求和提高小区的竞争力。

（五）K-means北京小区多因素聚类分析

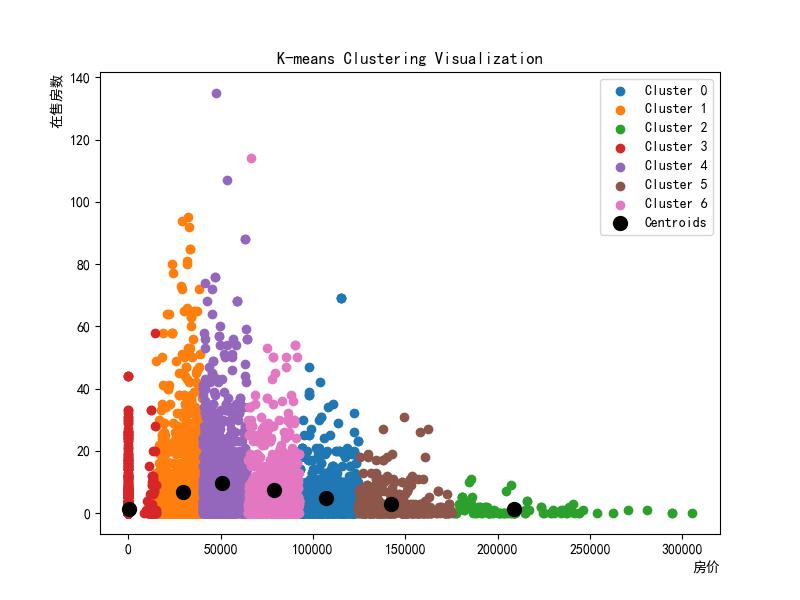


图 12 K-means北京小区聚类分布

基于综合房价、建筑材质、成交数目等因素得出的聚类结果可知，北京小区共被分为6类，房价对于聚类结果的影响相对显著，且基于聚类结果生成散点图可知，房价越高，在售房数越少，房价位于8000万元的小区在售房源较多，房源流动性较强，而高房价区域更加注重稀缺性，房源流动性较弱。

**五、总结**

本研究通过Python爬虫技术从链家网搜集了北京地区的相关小区信息，并采用了多种文本分析方法进行深入研究，揭示了小区配套设施、命名特点、房价因素等多个方面的特征。本研究发现，小区的命名往往反映了其地理位置、建筑类型以及独特的文化氛围或特色。而消费者在选择小区时，更倾向于受到配套设施、建设质量、内部户型、社区服务、经济因素等五个因素的影响，这些因素共同构成了一个完整的居住环境，对小区居民的生活质量产生重要影响。总的来说，本研究为房地产市场决策和城市规划提供了一定参考，为北京房产市场的特点和潜在市场趋势提供了深入洞察。

1. 参考文献：

   [] 余本国,刘宁,李春报.Python大数据分析与应用实战[M].电子工业出版社:202112.357. [↑](#endnote-ref-0)
2. [] MIHALCEA R,TARAU P.TextRank:Bringing order into texts[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing.Barcelona Spain:ACL,2004:404-411. [↑](#endnote-ref-1)
3. [] 井钰,王名扬,周文远.基于BBCM-TextRank的文本摘要提取算法研究[J].东北师大学报(自然科学版),2022,54(03):67-75.DOI:10.16163/j.cnki.dslkxb202107310001. [↑](#endnote-ref-2)
4. [] 李娅琪,裘鸿菲.新冠肺炎疫情前后游人感知比对研究——以武汉东湖风景区为例[J].西北林学院学报, 2022, 37(3):7. [↑](#endnote-ref-3)
5. [] MOHR J, BOGDANOV, P. Introduction — Topic models: What they are and why they matter [J]. Poetics, 2013, 41(6): 545-569. [↑](#endnote-ref-4)
6. [] 杜恒波,王绍运,罗润东.基于LDA主题模型的中国图书出口效果研究——以亚马逊海外读者评论为例[J].经济问题,2024(02):17-23.DOI:10.16011/j.cnki.jjwt.2024.02.017. [↑](#endnote-ref-5)