**天津市租房信息数据分析系统**

**­**———————————————

**实习实训项目需求说明书**

小组名称：豆瓣吃瓜小组

小组成员：金成花，于洋淼，杨雄峰，张景瑞

时间：2023.7.21

**目 录**

[一、引言 3](#_Toc140827735)

[二、项目背景 3](#_Toc140827736)

[三、目标用户分析 3](#_Toc140827737)

[租房求租者 3](#_Toc140827738)

[租房提供者 3](#_Toc140827739)

[数据分析师和研究人员 3](#_Toc140827740)

[四、功能需求 4](#_Toc140827741)

[搭建天津市租房信息数据库 4](#_Toc140827742)

[数据采集和处理 4](#_Toc140827743)

[数据展示和搜索 4](#_Toc140827744)

[智能问答机器人 4](#_Toc140827745)

[地理信息分析 4](#_Toc140827746)

[用户管理 4](#_Toc140827747)

[数据可视化和报告生成 5](#_Toc140827748)

[五、非功能需求 5](#_Toc140827749)

[技术支持 5](#_Toc140827750)

[性能要求 5](#_Toc140827751)

[用户界面和易用性 5](#_Toc140827752)

[可维护性和可测试性 5](#_Toc140827753)

# 一、引言

该项目旨在通过对租房信息进行数据分析，为用户提供有关租房市场的价格对比、区域热度、区域评价等信息，以帮助目标用户做出更明智的租房决策。为此，我们将搭建一个综合的租房信息数据库，并利用爬虫技术收集租房数据，通过数据分析和可视化展示工具，以及智能问答机器人，提供全面的数据洞察和决策支持。

# 二、项目背景

在天津市，租房市场竞争激烈，对租房信息的需求量大。租房求租者希望能够找到适合自己的租房信息，了解租金水平、地理分布和舒适程度等因素。租房提供者希望通过了解市场需求和租金变化等信息，优化自己的租房房源。数据分析师和研究人员对租房市场的趋势、价格走向和人口流动等因素感兴趣。然而，市场上缺乏一个专业的数据分析工具，为用户提供全面的租房信息分析和决策支持。因此，本项目旨在搭建一个租房信息数据库，并提供租房信息的可视化图表展示功能，结合智能问答机器人，帮助用户更好地了解租房市场情况。

# 三、目标用户分析

租房求租者

目标用户之一是寻找租房的求租者。他们希望能够快速便捷地找到适合自己的租房信息，了解租金水平、地理分布和舒适程度等因素。他们希望通过可视化图表展示了解租房市场的趋势和热点区域，以便做出明智的租房决策。

租房提供者

目标用户之二是租房提供者，如房产中介或业主。他们希望通过了解市场需求和租金变化等信息，更好地定价和推广自己的租房房源。他们可以通过该平台查询租房需求、了解市场的供需情况，并根据需求进行调整和优化。

数据分析师和研究人员

目标用户之三是数据分析师和研究人员，他们对租房信息市场感兴趣，希望通过该平台收集和分析租房数据，研究市场趋势、价格走向和人口流动等因素。他们希望能够获取原始数据和灵活的数据查询功能，以支持他们的研究工作。

# 四、功能需求

搭建天津市租房信息数据库

数据采集和处理

* 从“安居客”平台收集租房信息数据，并进行数据清洗、去重和格式转换。

将得到的DataFrame增加列，以方便更进一步地进行数据分析。比如在本次分析中我们利用爬虫得到的房源表中的‘街道’一列，对其内容进行split字符串切片操作，再进行一定的字符串拼接最终便得到了区域内容；

将DataFrame按照列标签提取指定列组成新的DataFrame，然后去重某一列后便实现了初步的数据清洗，后续再进一步series化，可以求出相关数目与均值；

对于单个表而言，绘制图表时直接将经过数据清洗后的列进行list序列化；

对于多表联立而言，首先需要将表中想要联立的属性列加入到其中的一个DataFrame中，然后再对加入列的DataFrame进行更进一步的数据清洗和去重工作；

绘制图表时，需要将数据进行更进一步的格式转换，主要采用对指定提取列进行list序列化的方式。如果后续要绘制热力图，则需将两个list进行按照字典形式合并成为一个新的list

* 将数据进行可视化处理后可视化为图表。

利用经过list序列化处理的数据，使用pyplot绘制条形图、饼状图、直方图、多条形直方图，plotly绘制网页表格，pyecharts绘制网页ECharts热点图

数据展示和搜索

* 使用Echarts可视化工具将数据库中的租房信息进行可视化展示，包括但不限于租金分布图、热门区域图、房屋类型图等。用户可以通过交互界面对图表进行筛选和操作，以获得更具体的数据洞察。
* 实现SQL搜索功能，允许用户根据特定的条件查询数据库中的租房信息。用户可以使用SQL语句来对数据进行灵活的筛选和排序，以满足个性化的数据需求。
* 允许用户通过交互界面筛选和操作图表，获取更具体的数据洞察。

智能问答机器人

* 通过基于网页问答内容编写的智能问答机器人，回答用户对租房相关问题的查询。机器人通过自然语言处理和匹配算法，提供准确的答案，并给出相关的租房建议和推荐。

**模型训练**

**分词**

使用jieba库，将输入的 command 数组中的文本句子进行分词，并以字符串形式返回分词后的结果。comm 是一个NumPy数组，其中存储了原始的文本句子。jieba.lcut(sentence) 将每个句子进行分词，返回一个由分词后词语组成的列表。这个列表会进一步被转换成一个字符串，使用空格将词语连接起来，以便后续的文本向量化。

**模型训练**

获取分词后的文本数据后，使用 TfidfVectorizer 类来构建一个基于TF-IDF的文本向量化模型。TF-IDF是一种常用的文本特征提取方法，它能够量化文本中每个词语的重要性。vectorizer.fit(words) 训练模型，并返回训练后的模型对象。

**保存模型及向量化后的数据集**

调用 getModel() 函数获取训练好的TF-IDF模型，并将其保存为名为 model.pkl 的文件。接下来，通过对文本进行向量化，将处理后的数据保存为CSV文件。首先，model.transform(getWords()) 将之前训练好的模型应用于分词后的文本数据，得到TF-IDF向量。然后，通过 .toarray() 将稀疏矩阵转换成稠密数组，并使用 .reshape(len(command), -1) 重新调整数组的形状，使其符合原始输入 command 的大小。这样得到的 data0 是一个二维数组，其中每行表示一个文本样本的向量表示，并将其保存为csv文件。

**计算余弦相似度并返回最优答案**

将用户问题分词后，调用已经训练好的模型将用户问题向量化，计算该问题与向量化后的与数据集中的问题的余弦相似度，返回最优的答案作为answer。

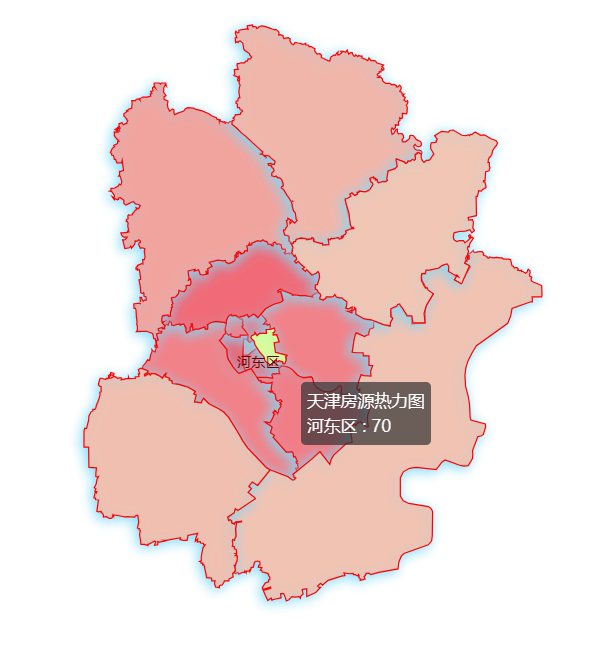
地理信息分析

* 利用echarts提供的地图文档，将租房信息与地理位置相结合。

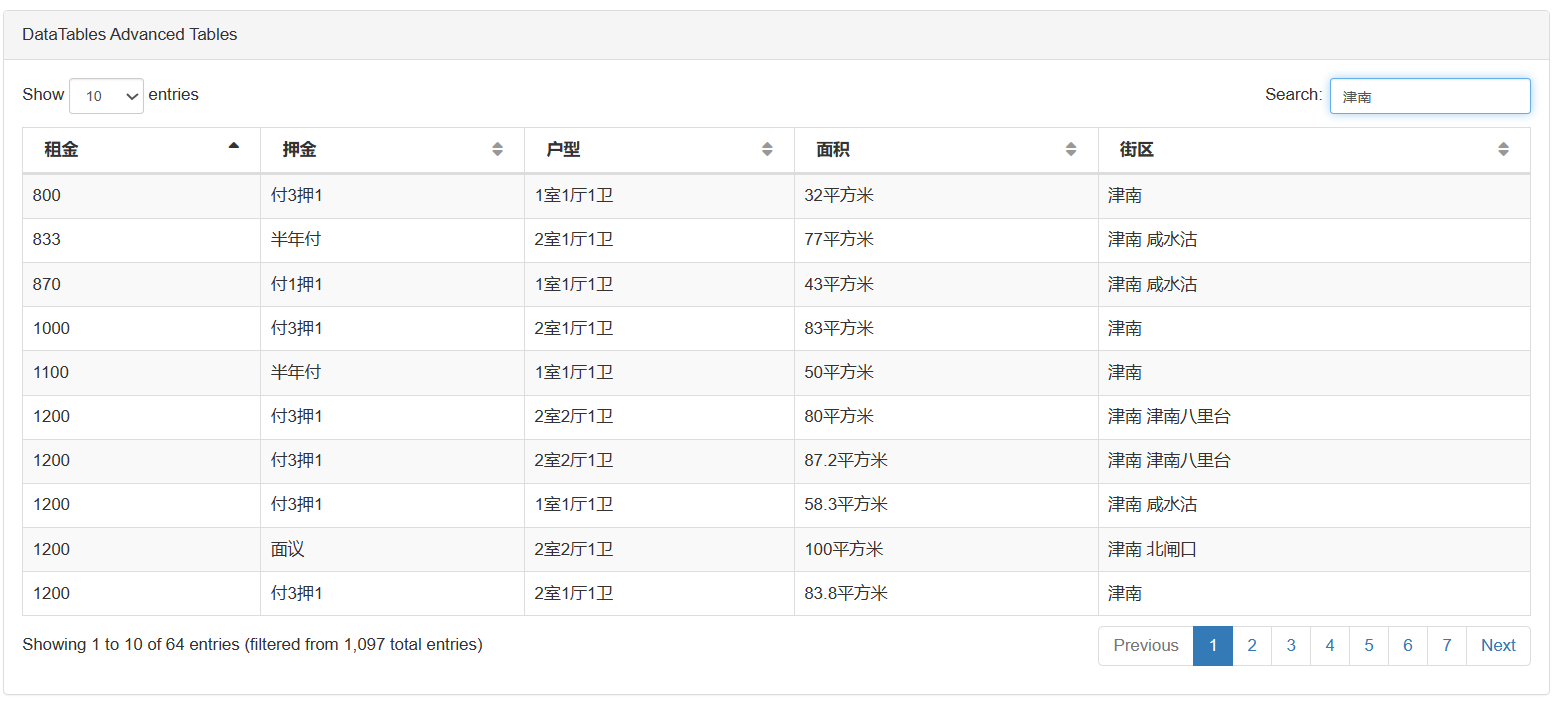




* 可视化展示租房信息在地图上的分布，帮助用户了解各地区的租房热度。

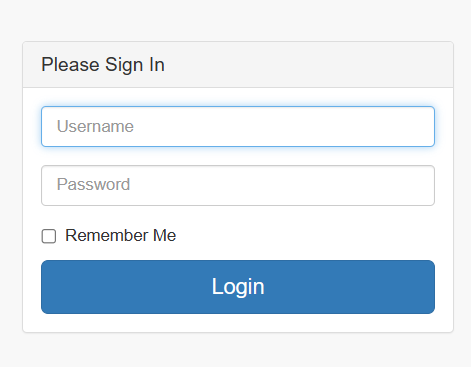


* 提供针对特定地区的租金分析，并支持用户根据地理位置进行数据查询和筛选。



用户管理

* 实现用户的登录注册功能，可在本地数据库新建天津市租房信息数据库。



数据可视化和报告生成

* 利用图表、报表和可视化图形展示统计结果和趋势分析。
* 提供图表导出功能，允许用户生成报告或分享分析结果。

# 五、非功能需求

技术支持

为了实现上述目标，本项目将采用多种技术来支持平台的开发和运行。

* 首先，通过爬虫技术从多个租房网站和房产中介平台收集租房信息，并将其存储在统一的数据库中。然后，通过数据分析技术对数据进行清洗、去重和格式转换，以保证数据的准确性和一致性。
* 接下来，采用Echarts可视化图表工具，将数据转化为直观的图形展示，包括但不限于租金分布图、热门区域图和房屋类型图等。
* 最后，基于Django网页框架搭建用户界面，实现交互功能和智能问答机器人，为用户提供便捷的查询和解答服务。

综上所述，本项目旨在建立一个综合的天津市租房信息数据分析平台，通过收集、处理和展示租房数据，为求租者、租房提供者以及数据分析师和研究人员提供全面的数据洞察和决策支持。同时，通过爬虫技术、数据分析技术、Echarts可视化工具和Django网页框架的运用，实现了数据的及时性、准确性和可视化展示，提升了用户体验和数据分析的效果。

性能要求

* 系统应具备高性能和响应速度，能够在大量数据下快速进行数据处理和查询。系统需要具备并发处理能力，能够同时处理多个用户的请求，并保持稳定的性能。

用户界面和易用性

* 系统应提供直观、易用和用户友好的界面设计，降低用户学习成本
* 界面能够适应不同的设备和屏幕尺寸，提供优良的用户体验。

可维护性和可测试性

* 系统应易于维护和测试，具备良好的代码结构和文档，方便团队成员进行维护、测试和升级。
* 提供相关文档和注释，方便团队成员进行维护、测试和升级