

湖南网络工程职业学院毕业设计答辩记录表

学生姓名	罗奥	学号	202165110019	指导老师	刘芳
二级学院	信息工程学院	专业	大数据技术	班级	大数据技术二班
毕业设计题目	基于智房网二手房数据分析与可视化报告				
答辩日期	2024 年 5 月 21 日		答辩地点	线上	
答辩小组成员	曹清清、马萌、宁可、银奕淇				
学生自述要点： 二手房市场的发展是关系到民生福祉的重要问题。随着国家对新建商品房市场的调控力度加大，二手房市场正在逐步发展成熟，并与一手房市场形成竞争与共存的关系。因此，通过大数据技术分析二手房数据，可以更好地了解市场状况，对规范二手房市场秩序具有重要意义。 毕设中，主要进行了以下工作： 使用爬虫技术从智房网抓取二手房数据，包括楼盘名称、户型、建筑面积、地址、价格、总价等字段。对抓取的数据进行清洗，删除重复值、处理异常值、修改数据类型等操作。 利用 Matplotlib 和 Echarts 进行数据可视化，展示各户型的占比、不同户型的平均价格、楼盘总价前十名等数据。 根据数据分析结果，提出了购房建议，如选择热门户型、关注高价户型、选择高端楼盘等。 在数据分析和处理过程中，掌握了 Python 爬虫、Pandas 数据处理、Matplotlib 可视化等技术，提升了自身的数据分析能力。					
答辩组提问及学生回答情况简述： 问题 1：你如何确保从智房网爬取的二手房数据是准确和完整的？ 回答：使用爬虫技术获取智房网二手房数据的接口，并解析网页以获取需要的数据字段。对获取的数据进行人工检查，确保数据的准确性。保存数据到 CSV 文件中，并进行数据清洗，包括删除重复值、处理异常值等，以确保数据的完整性。利用可视化工具检查数据的分布情况，发现异常值后进行删除处理。最终获得的数据经过清洗和处理后，可用于后续的数据分析。 问题 2：你在数据处理中遇到了哪些困难？是如何解决的？ 回答：数据中存在重复值和异常值，需要进行删除和处理。部分数据存在缺失值，需要进行填补。数据的格式不统一，需要进行格式转换。部分数据的单位不一致，需要进行单位转换。为了解决这些问题，我使用了 Python 的 Pandas 库，利用其提供的函数和方法对数据进行清洗和处理。具体包括：使用 drop_duplicates() 函数删除重复值。使用 fillna() 函数填补缺失值。使用 replace() 函数进行数据格式和单位转换。使用 astype() 函数修改数据类型。通过这些操作，我成功地对数据进行了预处理，使其更适合进行后续的数据分析。 问题 3：你认为在毕业设计中还有哪些可以改进的地方？ 回答：可以考虑扩展数据爬取范围，涵盖更多城市和地区的二手房数据，以提高分析结果的全面性。优化数据处理流程，实现数据清洗的自动化，减少人工干预，提高效率和准确性。加入时间序列分析、房价预测等高级分析技术，深入挖掘数据价值。整理分析结果，形成系统化的二手房市场分析报告，为相关决策提供更有力的参考。通过这些改进，毕业设计将更加完善和实用，对二手房市场的分析也会更加深入和全面。					
记录员（签名）	马萌		答辩小组长签名	曹清清	

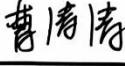
湖南网络工程职业学院毕业设计答辩记录表

学生姓名	饶蕾	学号	202165110021	指导老师	刘芳
二级学院	信息工程学院	专业	大数据技术	班级	大数据技术二班
毕业设计题目	基于智联卓聘网的招聘数据分析和可视化报告				
答辩日期	2024 年 5 月 21 日	答辩地点	线上		
答辩小组成员	曹清清、马萌、宁可、银奕淇				
学生自述要点： 我是湖南网络工程职业学院 21 级大数据技术二班的饶蕾。我的毕业设计报告的题目是《基于智联卓聘网的招聘数据分析和可视化报告》，我的指导老师是刘芳老师。接下来我将以三个方面来介绍我的毕业设计报告。 第一部分，项目背景。随着互联网的飞速发展，大数据的时代已然到来，互联网也为企业和求职者提供一个更加便利、更加广阔的平台，网络招聘逐渐成为企业招聘人才的一种重要途径。对于企业，为了提高企业竞争力，需要招聘更多优秀的人才，因此越来越多企业选择借助网络平台的力量进行工作招聘；对于求职者，想要获得真实有效的招聘信息，就要进行筛选，对信息的真实性进行核实。 第二部分，项目分析。在分析研究就业招聘资源现状的基础上，找出其中存在的问题，并分析其产生的影响因素，提出就业选择的建议，从而提高就业率，促进中国整个社会的进一步发展。运用网络爬虫技术、数据处理技术、数据可视化技术对招聘信息资源进行分析，更好地挖掘各地区招聘资源的有用信息和价值规律，帮助求职者寻找更契合自己的职位。 第三部分，项目介绍。描述中国就业资源、要求、现象和分布状况，分析中国就业资源的特征，找出其存在的问题及影响因素。使用大数据技术进行数据采集、清洗、分析、可视化智联卓聘网的招聘数据，从职位、薪资、学历要求、经验要求、地区等方面直观呈现招聘数据中的有用信息，提高分析的准确性和直观性。					
答辩组提问及学生回答情况简述： 问题 1：你为什么会选择这个题目？ 答：近几年来新冠肺炎疫情蔓延，对我国经济发展的影响不断加重，就业形势严峻。因此，找到一份适合自己的工作，是每一个应届毕业生都需要思考的问题。为了近一步的了解就业形势，获得更加真实的招聘数据，也为了帮助在校大学生能更多地了解就业信息与今后的在校规划。 问题 2：MySQL 数据库是什么，说出它的优缺点？ 答：MySQL 数据库是一种关系型数据库，将数据保存在不同的表中。 优点：1、性能卓越、服务稳定，很少出现异常宕机；2、软件体积小，维护成本低。 缺点：1、如果使用大量存储过程，那么使用这些存储过程的每个连接的内存使用量会大大增加；2、mysql 不允许调试存储过程。					
记录员（签名）	马萌		答辩小组长签名	曹清清	

湖南网络工程职业学院毕业设计答辩记录表

学生姓名	栗杨林	学号	202165110022	指导老师	刘芳
二级学院	信息工程学院	专业	大数据技术	班级	大数据技术二班
毕业设计题目	基于中国地震台网的历史数据分析与可视化报告				
答辩日期	2021 年 5 月 21 日		答辩地点	线上	
答辩小组成员	曹清清、马萌、宁可、银奕淇				
学生自述要点： 1. 项目背景和意义：介绍地震对人类造成的危害，强调分析历史地震数据的重要性，以便加深地震的认识，为采取有效的防灾减灾措施提供支持。 2. 项目技术路线：简要介绍采用的主要技术，包括网络爬虫技术、Python 数据处理和数据可视化技术。 3. 数据采集：阐述如何使用网络爬虫技术从中国地震台网网站爬取所需数据，并将数据存储为 CSV 文件。 4. 数据处理：描述数据处理的主要思路，包括删除重复值、处理缺失值，以及处理震级、经纬度、深度等数据。 5. 数据可视化：说明使用 Echarts 和 Pyecharts 库将地震数据可视化成不同类型的图表，从不同维度分析地震数据。 6. 问题分析和建议：概述从数据可视化结果中得出的结论，并提出相关的建议。 7. 收获与体会：分享在毕业设计过程中的收获，包括知识技能的提高、独立工作能力的培养，以及对地震数据分析的理解加深。					
答辩组提问及学生回答情况简述： 问题 1：你为什么会选择地震数据作为毕业设计的主题？ 回答 1：地震对人类社会造成重大危害，但人类目前仍难以准确预测。因此，通过分析历史地震数据，可以更好地了解地震的危害，为防灾减灾提供支持。 问题 2：在数据采集过程中，你遇到过哪些困难？ 回答 2：在数据采集过程中，需要根据网站结构提取数据，并处理翻页等。这些问题在老师的指导下，通过学习相关知识得到了解决。 问题 3：数据处理中，你采用了哪些方法？ 回答 3：在数据处理中，首先删除了重复值和缺失值，然后对震级、经度、纬度、深度等数据进行了标准化处理。 问题 4：数据可视化采用了哪些技术？ 回答 4：数据可视化采用了 Echarts 和 Pyecharts 技术，将地震数据可视化为条形图、饼图、散点图和折线图，从不同维度分析了地震数据。 问题 5：从可视化结果中，你获得了哪些结论？ 回答 5：从可视化结果中，可以看出不同地区的地震					
记录员（签名）	马萌		答辩小组组长签名	曹清清	

湖南网络工程职业学院毕业设计答辩记录表

学生姓名	龙晓宇	学号	202165110047	指导老师	马萌
二级学院	信息工程学院	专业	大数据技术	班级	21级大数据一班
毕业设计题目	基于言情小说吧的数据分析与可视化报告				
答辩日期	2024年5月21日		答辩地点	线上	
答辩小组成员	曹清清、马萌、宁可、银奕淇				
学生自述要点：（ 时 分—— 时 分 ） 本毕业设计旨在通过对言情小说吧中的数据进行深入分析和处理，揭示用户行为特征和小说受欢迎程度等重要信息，以此提升小说推荐的准确性、增加用户粘性、提高用户体验等方面的表现和效果。 本毕业设计重点关注言情小说吧中的数据特点和挖掘重点，为了最大程度提升数据分析结果的可靠性和有效性，会选取一些具有代表性的小说和用户数据进行分析和处理。同时，会注重数据隐私安全和合法性的保护，并考虑数据采集和分析的技术限制等因素，以期达成数据分析的最佳效果和范围。 通过本毕业设计的实施和数据分析结果的输出，希望能够为小说行业的各个方面提供更深入、更客观和更可靠的数据支持和参考，以期推动小说推广和读者数量的不断提升，为小说作家和出版社带来更大的利益和发展机会。					
答辩组提问及学生回答情况简述：（ 时 分—— 时 分 ） 1. 你为什么要选这个题目： 基于言情小说吧的数据分析与可视化报告可能是一个针对言情小说吧网站或其相关数据进行的综合性数据分析项目。这个项目旨在通过对网站的数据进行深入分析，以揭示言情小说吧的运营情况、用户行为、内容趋势等方面的信息。通过数据分析和可视化，可以更好地理解用户的兴趣和行为模式，以便更好地满足用户需求和优化网站运营。此外，数据分析还可以帮助了解竞争对手的情况，以便制定更好的市场策略。总之，基于言情小说吧的数据分析与可视化报告是一个非常有意义的项目，可以帮助学生掌握数据分析技能，并提高对实际问题的解决能力。 2. 你觉得你的毕业设计还有哪些方面可以经行完善： 我的毕业设计是基于言情小说吧的数据分析与可视化报告，虽然已经完成了一部分，但还有很多方面可以进一步完善。首先，可以进一步增加数据来源和扩大数据样本。目前，我的毕业设计主要基于言情小说吧网站的数据，但还可以考虑加入其他相关网站或数据来源，以获得更全面的数据样本，提高分析的准确性和可靠性。其次，可以增加更丰富的数据分析方法和技术。可以进一步探索更高级的分析方法和技术，如机器学习、自然语言处理等，以更深入地挖掘数据中的信息。最后，可以增加对用户行为和兴趣的分析和挖掘。目前我的毕业设计主要集中在网站运营和内容分析方面，但还可以进一步探索用户行为和兴趣方面的信息，如用户留存率、活跃度、阅读习惯等，以更好地理解用户需求和优化网站运营。 总之，我的毕业设计还有很多方面可以进一步完善，需要不断地探索和创新。					
记录员（签名）			答辩小组长签名		

湖南网络工程职业学院毕业设计答辩记录表

学生姓名	黄靖	学号	202165110043	指导老师	马萌
二级学院	信息工程学院	专业	大数据技术	班级	21级大数据一班
毕业设计题目	基于 3DM 网站的网络游戏资源数据分析与可视化报告				
答辩日期	2024 年 5 月 21 日		答辩地点	线上	
答辩小组成员	曹清清、马萌、宁可、银奕淇				
学生自述要点：（ 时 分—— 时 分 ） 随着硬件及互联网技术的迅速进步，单机游戏的图像、内容及响应速度都在不断升级，单机游戏的发展也在更符合玩家喜好。以基于 3DM 网站的网络游戏资源数据分析与可视化报告为背景，通过调研、分析已有的数据，利用所学的专业知识采集相关数据，并对游戏姓名，评分，类型，标签，语言，发布时间以及操作平台等数据进行分析及可视化展示，最后为用户提供简单参考和建议，方便人们能更好的查看和理解网络游戏的发展与在游戏市场的格局变化。 1. 数据采集。本项目中主要采集单机游戏数据。涉及网页数据的采集及数据的存储。 2. 数据清洗。对无效数据，如重复数据、错误数据、残缺数据进行清洗处理，并进行存储。 3. 数据分析。对单机游戏数据进行统计分析，给出当前单机游戏在市场的占比的状况，为单机游戏的发展提出简单合理的参考和建议。 4. 数据可视化展示。各评分的游戏数量展示、各类游戏标签的词云图展示、游戏中使用语言展示、各年月份游戏的发布数量展示、热度前十的游戏标签视图展示。 5. 建议：通过数据分析提出对单机游戏市场发展的相对意见。					
答辩组提问及学生回答情况简述：（ 时 分—— 时 分 ） 1. 你为什么要选择这个题目？ 基于近几年游戏行业的迅速发展，游戏行业的发展前景空前繁荣。但是游戏的高质量发展以及全面发展还有所欠缺，本次选题的原因就是为了分析出游戏行业发展的主要方向以及各方面全面发展提出有利的建议。 2. 你的数据采集使用的什么技术实现的？最后有多少条数据。 设计中所涉及的专业技术有以下：Python 网络爬虫技术、Python 数据处理技术、数据可视化技术。在经过以上技术的筛选下，最后剩下了 1000 余条数据。 3. 你觉得本次设计对你有什么提升吗？ 通过这这次的设计，我发觉到我有将自己曾经学习过的知识联系并运用起来。让我充分认识到大数据这门专业的实用性，也让我对自己学习过的知识得到了巩固和熟悉。在本次过程中，经过自己从错误的代码到正确运行和在错误中的摸爬滚打，我知晓了自己学习的还只是太片面了，很多知识还是得靠我自己去慢慢摸索。					
记录员（签名）	马萌		答辩小组长签名	曹清清	

湖南网络工程职业学院毕业设计答辩记录表

学生姓名	曾素琴	学号	202165110051	指导老师	马萌
二级学院	信息工程学院	专业	大数据技术	班级	21 大数据一班
毕业设计题目	基于前程无忧网的招聘数据分析与可视化报告				
答辩日期	2024 年 5 月 21 日		答辩地点	线上	
答辩小组成员	曹清清、马萌、宁可、银奕淇				
学生自述要点：（ 时 分—— 时 分 ） 本毕业设计以前程无忧网为背景，利用大数据分析技术，对前程无忧网中的招聘数据中岗位名称、工作地点、月薪范围、工作年限、学历要求、公司性质等数据进行采集，并对无效数据，如重复数据、错误数据、残缺数据进行清洗处理，对前程无忧网的不同岗位的数量统计、不同岗位的工作地点分布、不同岗位的月薪范围排行、不同岗位的工作年限、学历要求、公司性质、公司规模进行数据分析及可视化展示，最后提供求职者和企业的简单参考。					
答辩组提问及学生回答情况简述：（ 时 分—— 时 分 ） 1. 你为什么要选这个题目？ 招聘市场是一个与我们日常生活息息相关的领域，每年都有大量的求职者和企业参与其中。通过这个题目，我可以更深入地了解招聘市场的现状和趋势，包括不同行业、不同职位的招聘需求和竞争情况等，从而更好地为自己的职业规划和发展做出决策。 2. 你的数据采集使用的什么技术实现的？最后有多少条数据？ 使用 python 网络爬虫技术，使用爬虫中的 selenium 模块，通过加载本地浏览器驱动，访问目标站点，获取目标站点的数据。最后有 1000 条数据。 3. 你觉得你的毕业设计还有哪些方面可以进行完善？ 增加更多行业和职位的数据分析：前程无忧网是一个综合性的招聘网站，涉及到的行业和职位非常广泛。在未来的分析中，可以针对不同行业和职位的招聘数据进行深入的分析和挖掘，以更好地了解各行业的招聘需求和趋势。 引入更多的数据可视化方法：可视化报告是数据分析的重要部分，可以通过图表、图形等形式呈现数据。未来可以考虑引入更多的数据可视化方法，如时间序列图、热力图等，以更全面地展示招聘市场的数据和趋势。 加入对求职者行为的分析：除了对招聘方和职位的数据进行分析之外，还可以考虑加入对求职者行为的分析，如求职者地域分布、年龄结构等，以更全面地了解招聘市场的参与者和动态。					
记录员（签名）	马萌		答辩小组长签名	曹清清	

湖南网络工程职业学院毕业设计答辩记录表

学生姓名	曹景琦	学号	20216511001 1	指导老师	宁可
二级学院	信息学院	专业	大数据技术与应用	班级	21 大数据二班
毕业设计题目	基于上海黄金交易所的数据分析和可视化报告				
答辩日期	2024 年 5 月 21 日		答辩地点	网上答辩	
答辩小组成员	曹清清、银奕淇、宁可、马萌				
学生自述要点： 1、项目背景：主要介绍了基于上海黄金交易所的数据分析与可视化报告的创作背景 2、项目数据采集与存储：对该项目的数据爬取、数据处理、数据可视化的具体分析 3、项目数据可视化：主要介绍说明了该项目的问题所在 4、项目解决问题和方法建议：根据提出来的问题，结合可视化展示出来的数据，总结出上海黄金交易所的投资决策					
问题 1. 为什么要选择上海黄金交易所进行爬取分析？ 答：上海黄金交易所在中国黄金市场具有重要地位，其交易数据和价格波动对市场有着较大影响，因此分析这些数据可以帮助了解中国黄金市场的整体走势和特点。2. 在爬取数据的过程中出现了哪些困难？ 问题 2：在数据采集中为什么选择 BeautifulSoup 方法？ 答：BeautifulSoup 可以解析各种类型的 HTML 和 XML 文档,包括不规范的文档。BeautifulSoup 提供了一种灵活的选择器语言,可以方便地定位文档中的元素。 问题 3：在毕业设计中遇到的问题有哪些如何解决？ 答：1. 时间管理困难。可以通过制定详细的时间表和任务划分，设定每天可实现的目标。 2. 资料收集困难。可以积极寻找各种资料来源，如参加学术会议和研讨会、利用互联网搜索引擎和学术数据库、与导师和其他专业人士进行专业讨论和咨询。 3. 技术困难。可以寻求导师和同学的帮助，进行技术交流和讨论，利用在线资源如编程论坛、教程和文档学习，使用在线编程平台或开源代码库。					
记录员（签名）	马萌		答辩小组长签名	曹清清	

湖南网络工程职业学院毕业设计答辩记录表

学生姓名	向政武	学号	202165110075	指导老师	宁可
二级学院	信息工程学院	专业	大数据技术	班级	21级大数据技术一班
毕业设计题目	基于大江大河水情数据分析与可视化报告				
答辩日期	2021年5月21日	答辩地点	终教楼 1107		
答辩小组成员	曹清清、银奕淇、宁可、马萌				
学生自述要点： 本毕业设计是基于中国土地市场网采取的土地数据，进行数据分析和可视化，主要有如下几点： 1. 数据采集。本项目中主要采集水情数据。涉及网页数据的采集及数据的存储。 2. 数据清洗。对无效数据，如重复数据、错误数据、残缺数据进行清洗处理，并进行存储。 3. 数据分析。对当前水资源数据进行统计分析，给出当前水清具体的状况，为水情利用提出简单合理的参考和建议。 4. 数据可视化展示。分析流量和警戒水位之间的关系，分析各个地区的平均流量进行可视化展示。 5. 建议：通过数据分析提出水情合理利用的建议。					
答辩组提问及学生回答情况简述： 1. 你在实现设计的过程中遇到了哪些困难？以及如何解决？ 答：在爬取网站数据的过程中，爬取出来的数据达不到自己的要求，解决办法就是检查自己的爬虫代码并进行优化，实在解决不了就请教老师指导，最终使其爬取的数据达到自己的期望。 2. 你的设计是为了解决什么问题而提出来的建议？ 答：我的设计是分析江河水情的数据，分析其中农业用水、工业用水、住房用水、绿化用水等的占比，然后分析怎样才能使其平衡，经济才能平稳发展，而不出现过度使用水情的现象。 3. 你的数据采集使用的什么技术实现的？最后有多少条数据？ 答：我的数据采集是使用 python 语言进行编程，代码中使用了 selenium 方法，爬取出来的数据有 3200 行。					
记录员（签名）	马萌		答辩小组长签名	曹清清	

湖南网络工程职业学院毕业设计答辩记录表

学生姓名	朱振阳	学号	202165110010	指导老师	刘艳
二级学院	信息工程学院	专业	大数据技术	班级	21 大数据二班
毕业设计题目	《基于苏州看房网的二手房数据分析与可视化报告》				
答辩日期	2024 年 5 月 20 日		答辩地点	线上	
答辩小组成员	刘芳、刘艳、曹守富				
学生自述要点：（5 分钟） 我的毕业设计题目是《基于苏州看房网的二手房数据分析与可视化报告》，随着二手房的大众化的发展，二手房资源总量在不断增加，二手房的产业不断壮大。同时，二手房在我国各地的区域分布区域化和消息的滞后化认然制约着二手房的全面发展。因此，有必要分析研究二手房分布的基础上，找出其中存在的问题，并分析其产生的影响因素，提出解决二手房消息的滞后化和区域化，帮助更多的人解决住房需要。 在本次毕业设计中，使用了网络爬虫技术，python 数据处理技术和数据可视化，对苏州看房网的二手房数据进行了深入研究。并完成了以下工作： 1. 数据采集。本项目中主要采集苏州二手房数据。涉及网页数据的采集及数据的存储。 2. 数据清洗。对无效数据，如重复数据、错误数据、残缺数据进行清洗处理，并进行存储。 3. 数据分析。对当前苏州二手房进行统计分析，给出当前人们对二手房的选择的状况，为人们选择二手房提出简单合理的参考和建议。 4. 数据可视化展示。各个小区在售数量、各个小区平均价格前十的小区、二手房房屋装修情况、二手房各户型的数量、二手房层数与装修的价格图可视化展示。 通过以上工作，我得出了一些有意义的结论。列如，我发现不同区域的二手房价格存在较大差异，并且房屋面积与价格之间存在一定的相关性。此外，我还通过可视化展示，清晰地展示了二手房市场的发布情况和趋势。 当然，在项目实施过程中，我也遇到了一些挑战。例如，数据的复杂性和多样性给分析带来了一定的困难。但是，通过不断学习和尝试，我最终克服了这些问题。					

答辩组提问及学生回答情况简述：（5 分钟）

问题 1：你为什么选这个题目？

1. 实践意义：分析二手房数据能帮助理解市场动态，为投资者、开发商和政策制定者提供参考。
2. 数据驱动：通过对二手房数据进行挖掘和分析，可以揭示隐藏的模式和趋势。
3. 技术挑战：该课题涵盖数据采集、清洗、分析、建模和可视化等环节，对综合技能是一次锻炼和提升。
4. 兴趣驱动：对房地产市场和数据分析有兴趣，希望将兴趣和专业技能结合，做出有意义的研究成果。

问题 2：你在实现的过程中遇到了哪些困难？你是怎么解决的？

1. 数据收集难度：苏州看房网的数据并没有直接可下载的数据库，因此我需要自行编写爬虫程序进行数据抓取。在这个过程中，我发现网站有反爬虫机制，导致数据爬取困难。为了解决这个问题，我学习并实现了使用代理服务器和设置爬虫爬取延时的方法，以模拟正常用户的行为，最终成功绕过了网站的部分反爬策略，并高效获取了所需的数据。
2. 数据清洗与处理：收集到的数据中包含大量的缺失值和异常值。处理这些数据时，我首先使用了 Python 中的 Pandas 库来评估数据的质量和完整性，然后根据每种情况采用了不同的策略，如填充缺失值、删除异常记录或转换数据类型，以确保数据分析的准确性。
3. 数据可视化的表达难题：为了使复杂的分析结果易于理解，我使用了 Matplotlib 进行了数据的可视化表示。最初，我不确定如何选择最有效的图表类型来表达特定的数据关系。通过不断尝试和优化，以及参考设计和可视化的最佳实践，我最终能够创造出既信息丰富又视觉吸引的图表。

问题 3：你的数据采集使用的什么技术实现的？最后有多少条数据？

数据的采集是通过 Python 编程实现的。我主要使用了 Python 中的几个强大的库来编写数据爬虫，具体包括：

1. Requests：用于向网页发送 HTTP 请求，获取网页内容。
 2. BeautifulSoup：从网页抓取的 HTML 内容中解析所需的数据。
 3. Scrapy：这是一个更高级的框架，用于开发复杂的爬虫程序。我用它来处理反爬虫技术和大量数据的抓取。
 4. Pandas：在数据抓取后用于数据清洗和预处理。
- 最后处理完还有 800 条左右。

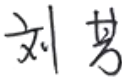
记录员（签名）

曹宇富

答辩小组长签名

刘芳

湖南网络工程职业学院毕业设计答辩记录表

学生姓名	谭汉雄	学号	202165110001	指导老师	刘艳
二级学院	信息工程学院	专业	大数据技术	班级	21 大数据二班
毕业设计题目	基于飞卢小说网的小说数据分析与可视化报告				
答辩日期	2024 年 5 月 20 日		答辩地点	线上	
答辩小组成员	刘芳、刘艳、曹守富				
学生自述要点：（5 分钟） 1. 该毕业设计题目为《基于飞卢小说网的小说数据分析与可视化报告》，现在网络小说盛行，用户小说网里面寻找小说时很容易苦恼，对满目的小说种类有选择困难症，基于本毕业设计以飞卢小说网为数据来源基础进行可视化分析，帮助用户快速找到自己想要的小说。 2. 数据采集:完成数据获取的思路设计，爬虫设计以及数据存储的形式。 3. 数据清洗与分析:完成无效数据的清洗，确定关键数据字段进行数据分析。 4. 数据可视化分析:完成数据分析的可视化展示并给出具体的分析内容。 5. 解决问题的方法及建议:针对本项目提出的问题给出具体的解决方法或可行的建议。					
答辩组提问及学生回答情况简述：（5 分钟） 问题 1：数据采集时用的是什么 python 哪个第三方库，使用过程中遇到了什么问题？ 答：用的是 python 的 xpath 库进行网页的数据爬取，使用过程中遇到某些字段格式不统一，出现爬取多余的数据问题，后续进行 if 判断特定情况进行排查解决问题。 问题 2：数据处理用的是 python 哪个第三方库，对这个第三方库有什么心得体会？ 答：用的是 python 的 pandas 第三方库，pandas 里面的函数非常多，提供了许多常见问题解决方案，可以很实用的运用到程序里面，数据清洗能力很强，快速的讲表格内容进行规范化，专业化，以便后续的可视化呈现。 问题 3：数据可视化里面你觉得那个库最方便，最好用？ 答：python 里面的 PVEcharts 库最好用，能呈现出非常漂亮的网页可视化界面，多种多样的呈现格式能自主搭配。					
记录员（签名）			答辩小组长签名		

湖南网络工程职业学院毕业设计答辩记录表

学生姓名	谢伟枝	学号	20216511003 1	指导老师	曹守富
二级学院	信息工程学院	专业	大数据技术	班级	21 大数据 2 班
毕业设计题目	基于无忧招聘网的计算机类招聘数据分析与可视化报告				
答辩日期	2024 年 5 月 21 日		答辩地点	终教楼 1107	
答辩小组成员	刘芳、刘艳、曹守富				
学生自述要点：（9 时 15 分—— 9 时 20 分 ） 在发布招聘信息的过程中，企业逐渐放弃了传统的线下招聘方式，为了广大的求职者准确快速地找到符合自身需求并满足职位要求的工作，基于此本毕业设计以无忧招聘资源数据为背景，对招聘网站数据爬取、解析、处理并进行可视化展示，为应聘者提供参考。 数据获取：通过 selenium 库向目标站点发起请求，获取源代码。得到的内容用 XPath 网页解析库进行解析。获取网站中岗位名称、公司名称、公司地区、工作经验、学历、薪资、公司行业、公司性质、招聘发布时间的字段数据，并存储到 csv 中。 数据处理：删除重复值和缺失值、岗位名称的数据只保留了岗位，删除无关数据、公司地区统一将市进行删除，只保留省份、公司性质中去除了公司规模的人数数据，只保留了民营、合资、国企、上市公司等数据。 数据可视化：该项目中绘制了关于公司行业类型的词云图，关于学历、薪资的饼图和折线图，关于河南岗位招聘数前十的条形图，关于工作经验与薪资的雷达图，关于公司性质与学历的桑基图。					
答辩组提问及学生回答情况简述：（ 9 时 20 分—— 9 时 25 分 ） 问题 1：你为什么会选择这个题目？ 近年来随着我国计算机水平的发展，计算机行业的热门，越来越多的计算机人才涌入社会，对于求职者来说，想要搜集到有效的数据，在网上找到一份适合的工作，是一件非常困难的事情。临近毕业，我想要更深入的了解计算机行业的招聘情况，为之后的职业规划做参考。 问题 2：使用了哪几种主要技术？ 网络爬虫技术、Python 数据处理技术、数据可视化技术 问题 3：为什么使用 Echarts 进行数据可视化？ Echarts 是一个基 Javascript 的数据可视化库, 它能提供直观、生动、可交互、可高度个性化定制的数据可视化图表。Echarts 支持多种图表类型，包括常见的折线图、柱状图、饼图、热力图、散点图、雷达图等，并且支持图与图之间的混搭。同时提供了丰富的配置选项，用户可以根据自迹己的需求对图表进行定制。也提供了数据区域缩放、数据视图、动态数据更新等功能，这些功能使用户能够更好地探索和分析数据。					
记录员（签名）	曹守富		答辩小组长签名	刘芳	

湖南网络工程职业学院毕业设计答辩记录表

学生姓名	杨炯健	学号	202165110033	指导老师	曹守富
二级学院	信息工程学院	专业	大数据技术	班级	21 大数据 2 班
毕业设计题目	基于中国非物质文化遗产网·中国非物质文化遗产数字博物馆的数据分析与可视化报告				
答辩日期	2024 年 5 月 21 日		答辩地点	终教楼 1107	
答辩小组成员	刘芳、刘艳、曹守富				
学生自述要点：（9 时 45 分—— 9 时 50 分 ） 该毕业设计的题目为《基于中国非物质文化遗产网·中国非物质文化遗产数字博物馆的数据分析与可视化报告》，非物质文化遗产与物质文化遗产共同承载着人类社会的文明，是世界文化多样性的体现。 数据获取：在该项目中使用了 Python 的第三方库 request 等进行的数据获取，获取爬取的字段有：序号，项目序号、编号、名称、类别、公布时间、类型、申报地区或单位、保护单位等字段数据，将获取的数据保存为 csv 文件。 数据处理：对重复、空白值数据进行处理，删除，存储成可用的数据类型。 数据可视化：中国地图，折线图，条形图，环形图，柱状图。 总结与建议：根据提出来的问题，结合可视化展示出来的数据，总结出当前我国非物质文化遗产时间和空间的分布规律，非物质文化遗产的种类和分布情况，为我国推进非物质文化遗产的保护和传承工作提供参考和建议。					
答辩组提问及学生回答情况简述：（ 9 时 50 分—— 9 时 55 分） 问题 1：对数据可视化技术，请简单说明？ 答：数据可视化技术是将数据转换成图表、图形或其他视觉元素的过程，以便更直观地理解和分析数据。常见的数据可视化技术包括柱状图、折线图、饼图、散点图、热力图等，它们可以帮助人们发现数据中的模式、趋势和关系。数据可视化在各行各业都有广泛的应用，从市场营销到科学研究都可以借助数据可视化技术来解释和传达数据。 问题 2：此次毕业设计你收获和体会是什么？ 答：通过这次毕业设计，让我见识到了每个非物质文化遗产都有自己独特的特点和特色，通过对不同类型非物质文化遗产的可视化分析，可以更加深入地了解它们的文化内涵、历史渊源、地域特色等信息，从而更好地呈现出中国非物质文化遗产的丰富性和多样性。然后将这些宝贵的遗产与现代的大数据技术相结合，一定可以更好地保护和传承传统文化。 问题 3：你是如何清洗和处理数据的？ 答：通过 pandas 库对数据进行清洗，包括删除异常值、去除重复值、转换数据格式等，以提升数据质量。					
记录员（签名）	曹守富		答辩小组长签名	刘芳	

湖南网络工程职业学院毕业设计答辩记录表

学生姓名	李欣茜	学号	202165110064	指导老师	曹清清
二级学院	信息工程学院	专业	大数据技术	班级	21 大数据一班
毕业设计题目	基于土流网的土地流转数据分析与可视化报告				
答辩日期	2024 年 5 月 14 日		答辩地点	终教楼 1107	
答辩小组成员	曹清清、马萌、刘芳、曹首富				
学生自述要点：（ 时 分—— 时 分 ）					
本毕业设计以湖南土流网数据为背景，利用大数据分析技术，对湖南各地区土流资源数据中土地名称、土地价格、土地面积、年份、转让方式、土地地址、土地类型等进行采集，并对无效数据，如重复数据、错误数据、残缺数据进行清洗处理，再对土地面积、转让方式、土地地址、土地类型等数据进行数据分析及可视化展示，最后给各区域购买者提供更多合适的土地信息，帮助购买者解决找地、租地需求，并协助解决各地区土地资源、资产不集中、供需不对称的问题。					
答辩组提问及学生回答情况简述：（ 时 分—— 时 分 ）					
1. 你为什么要选这个题目？ 我选择这个基于土流网的土地流转数据分析与可视化报告作为毕业设计题目，主要是因为土地流转在农村经济发展中具有重要意义，通过分析土地流转数据可以更好地了解农村土地资源的利用情况、流转趋势以及对农村经济的影响，为农村土地政策的制定和农村发展的规划提供重要参考。土流网作为专门提供土地流转信息的平台，其数据来源广泛、权威可信，包括政府部门、企业以及个人的信息，因此分析基于这样的数据可以保证数据的可靠性和准确性。通过数据分析和可视化技术，可以将庞大的土地流转数据转化为直观、易懂的图表和图像，帮助决策者和研究人员更直观地了解土地流转的情况和特点，发现其中的规律和问题，从而更好地指导土地资源的管理和利用。 2. 你的数据采集使用的什么技术实现的？最后有多少条数据？ 在数据采集过程中，我采用了网络爬虫技术来实现数据的自动采集，使用的是爬虫模块中的 selenium，通过加载本地浏览器驱动，访问目标站点，获取目标站点的数据。这种技术能够有效地遍历目标网页的各个页面，实现对数据的全面收集，而且由于使用了本地浏览器驱动，可以避免被目标网站的反爬虫机制检测到。在爬取过程中，有遇到了重复缺失的数据，经过数据处理和筛选，最终获得了 1214 条有效数据，为后续的数据分析和可视化提供了充分的支持。通过网络爬虫技术的应用，我们能够快速、高效地获取大量的数据，为土地流转领域的研究和决策提供了有力的数据支持。 3. 这些发现或结论对于土地资源利用和土地购买者有什么实际意义？ 对于土地资源利用：通过分析土地流转数据，可以更好地了解土地利用情况，包括土地流转类型、流转方向、地块大小等信息，有助于合理规划土地利用方式，提高土地资源的利用效率。借助这些数据，相关部门和决策者可以更准确地评估土地利用状况，制定相应的土地资源管理政策，促进土地资源的可持续利用。 对于土地购买者：这些数据可以提供有关土地流转市场的信息，包括土地价格趋势、土地供需状况等，有助于他们做出更明智的土地投资决策。通过分析土地流转数据，购买者可					

以了解不同地区土地流转的特点，有助于选择合适的投资地点，并更好地把握土地市场的变化。 综上所述，通过对土地流转数据的挖掘和分析，可以为土地资源的合理利用和土地购买者的决策提供重要参考。这些数据不仅可以帮助相关部门制定政策，优化土地资源管理，也可以为投资者提供信息支持，提升他们的投资效益。因此，深入研究土地流转数据的发现和结论对于推动土地资源利用的可持续发展、促进土地市场的规范化和健康发展具有积极意义。			
记录员（签名）	曹宇富	答辩小组长签名	曹清

湖南网络工程职业学院毕业设计答辩记录表

学生姓名	李晶	学号	202165110063	指导老师	曹清清
二级学院	信息工程学院	专业	大数据技术	班级	21级大数据技术一班
毕业设计题目	基于链家网的长沙二手房数据分析与可视化报告				
答辩日期	2024年5月14日		答辩地点	终教楼1107	
答辩小组成员	曹清清 马萌 刘芳 曹首富				
学生自述要点：（ 时 分—— 时 分 ）					
基于链家网的长沙二手房数据分析与可视化报告，是一项针对长沙市二手房房地产市场的深度研究。报告通过收集链家网上的二手房数据，运用统计学方法和数据可视化技术，对长沙市的二手房市场进行了深入的剖析。 报告首先概述了长沙二手房市场的总体情况，包括房源数量、价格走势、交易热度等关键指标。接着，通过对数据的细致挖掘，报告揭示了不同区域、房型、价格区间的二手房供需状况，以及市场中的热点和趋势。此外，报告还特别关注了影响房价的关键因素，如地理位置、交通便利性、周边配套设施等。 为了使分析结果更加直观易懂，报告中采用了多种图表形式，如折线图、柱状图、散点图等，将复杂的数据转化为视觉化的信息。这不仅有助于读者快速把握市场动态，也为投资者、购房者提供了决策支持。 总体而言，这份报告是对长沙二手房市场的一次全面体检，它不仅提供了丰富的市场信息，也展示了数据分析在房地产市场研究中的应用价值。通过这份报告，相关利益方可以更准确地理解市场现状，做出更为明智的投资或购买选择。					
答辩组提问及学生回答情况简述：（ 时 分—— 时 分 ）					
1. 你为什么选这个题目？					
答：是因为链家网作为中国知名的房地产平台，其提供的数据具有较高的代表性和可靠性。通过对链家网上的长沙二手房数据进行爬取、清洗、分析和可视化，可以揭示出长沙二手房市场的特点和趋势，为购房者、投资者以及政策制定者提供有价值的参考信息。此外，链家网的数据通常较为丰富和更新及时，这有助于确保分析结果的时效性和准确性。					
2. 你在做实现的过程中遇到了哪些困难？你是怎么解决的？					
答：反爬虫机制：如果爬虫程序运行过快，可能会触发链家网的反爬虫机制，导致后续请求被重定向到验证页面，影响数据采集的完整性。					
解决方法：控制爬虫速度，通过设置合理的延时和随机睡眠策略，避免触发链家网的反爬虫机制。					
3. 你在完成毕业设计过程中收获了什么？					
答：在完成毕业设计过程中，我获得了宝贵的实践经验和对数据分析流程的深刻理解。首先，我学会了如何有效地收集和清洗大规模的网络数据，这对于确保数据质量和后续分析的准确性至关重要。其次，通过运用统计学和数据挖掘技术，我对房价趋势、影响因素以及市场动态有了更深入的认识。此外，掌握数据可视化工具使我能够将复杂的数据集转化为直观、易懂的图表，从而提高了我的沟通和展示能力。这次经历不仅增强了我的专业技能，也锻炼了我的问题解决能力和项目管理能力，为我未来的职业生涯奠定了坚实的基础。					

记录员（签名）	曹宇富	答辩小组长签名	曹清清
---------	-----	---------	-----

湖南网络工程职业学院毕业设计答辩记录表

学生姓名	管智琛	学号	202165110059	指导老师	曹清清
二级学院	信息工程学院	专业	大数据技术	班级	21 大数据一班
毕业设计题目	基于聚汇数据网的世界 GDP 数据分析与可视化报告				
答辩日期	2024 年 5 月 14 日		答辩地点	终教楼 1107	
答辩小组成员	曹清清、马萌、刘芳、曹首富				
学生自述要点：（ 时 分—— 时 分 ） 本次毕设旨在通过利用互联网上公开的 GDP 数据资源，并结合数据分析与可视化技术，对世界各国的 GDP 数据进行深入研究与分析。通过对 GDP 的时序变化、地理分布、行业结构等方面的分析与可视化，可以更全面地了解不同国家的经济发展情况与特点，发现潜在的经济规律，为政府、企事业单位以及投资者提供决策参考，总的来说 GDP 数据分析的意义在于可以帮助决策者了解经济趋势、制定经济政策和企业投资决策。通过深入挖掘 GDP 数据，可以发现经济增长的主要驱动因素、产业结构变化等，为政府和企业提供更有科学依据的经济发展路径和投资决策方向					
答辩组提问及学生回答情况简述：（ 时 分—— 时 分 ） 1. 你为什么要选这个题目？ 随着全球经济的快速发展，了解和分析各国的 GDP（国内生产总值）数据对于研究经济趋势、制定经济政策以及企业投资决策至关重要。然而，由于数据庞杂、统计方式不一致等问题， 2. 你的数据采集使用的什么技术实现的？最后有多少条数据？ 数据的爬取，使用的是 python 里面的 requests 库，这个库可以对网站发起请求获取到对应的数据，再利用 lxml 库使用 xpath 技术解析出我们要的数据，然后存储到本地 3. 反思与总结 (1)对于 matplotlib 这个库的使用，一开始是奔着用起来比较简单去的，但是画出来的图可能比较丑，所以想把图画好看还是得花一些功夫的，在这期间我对这个库也更加了解了，自己可以调节很多方面的东西。 (2)这次实验对于 python 的使用让我学到了很多，也了解到了原来 python 能干这么多事，而且出来的效果很不错，体会到了大数据的魅力。使用 python 制作出一些东西，完成一次实验让我很有成就感，本次实验让我收获良多。					
记录员（签名）	曹首富		答辩小组长签名	曹清清	