数据挖掘、分析与应用

一、基本信息

课程代码：【2059337】
课程学分：【3】
面向专业：【计算机科学与技术】
课程性质：【专业实践】
开课院系：【信息技术学院计算机科学与技术系】
使用教材：

教材：

【Python 网络爬虫技术与实践，吕云翔著，机械工业出版社，2023 年 6 月】

参考书目：

【大数据导论（英文版），Thomas Erl, Wajid Khattak, Paul Buhler 著，机械工业出版社，2017 年 10 月】
【大数据导论，杨尊琦主编，机械工业出版社，2018 年 10 月】
【大数据导论，梅宏编著，高等教育出版社，2018 年 11 月】

课程网站网址：
先修课程：【计算机导论】

二、课程简介

数据挖掘、分析与应用是一门新兴的交叉性学科，是在信息技术领域和人工智能领域迅速兴起的计算机技术。数据挖掘技术面向应用，在很多重要的领域，数据挖掘都发挥着积极的作用。广大从事数据库应用与决策支持，以及数据分析等学科的科研工作者和工程技术人员迫切需要了解和掌握它。因此数据挖掘已经成为计算机专业及相关专业的重要课程之一。

本课程为计算机专业学生的专业实践课程。本课程主要介绍数据挖掘的基本概念，原理、方法和技术。旨在通过一学期的学习，使学生理解数据挖掘的基本流程，掌握数据挖掘的基本理论和技术，熟悉数据挖掘成果的显示；掌握数据挖掘的基本方法，能熟练地应用数据挖掘技术对现实数据进行有效的分析；结合相关软件能从大量数据中获取有价值的信息。

三、选课建议

本课程是适用于计算机类专业的专业实践课程，要求具有计算机导论的基础、基本编程能力和系统设计能力、对于数据库的基本操作能力和算法设计能力。

四、课程与专业毕业要求的关联性

专业毕业要求	关联
LO1：工程知识：能够将数学、自然科学、工程基础和专业知识用于解决复杂工程问题	
LO2：问题分析：能够应用数学、自然科学和工程科学的基本原理，识别、表达、并通过文献研究分析复杂工程问题，以获得有效结论	●

LO3：设计/开发解决方案：能够设计针对复杂工程问题的解决方案，设计满足特定需求的系统、单元（部件）或工艺流程，并能够在设计环节中体现创新意识	
LO4：研究：能够基于科学原理并采用科学方法对复杂工程问题进行研究，包括设计实验、分析与解释数据、并通过信息综合得到合理有效的结论	●
LO5：使用现代工具：能够针对复杂工程问题，开发、选择与使用恰当的技术、资源、现代工程工具和信息技术工具，包括对复杂工程问题的预测与模拟，并能够理解其局限性	
LO6：工程与社会：能够基于工程相关背景知识进行合理分析，评价专业工程实践和复杂工程问题解决方案对社会、健康、安全、法律以及文化的影响，并理解应承担的责任	
LO7：环境和可持续发展：能够理解和评价针对复杂工程问题的专业工程实践对环境、社会可持续发展的影响	
LO8：职业规范：具有人文社会科学素养、社会责任感，能够在工程实践中理解并遵守工程职业道德和规范，履行责任	
LO9：个人和团队：能够在多学科背景下的团队中承担个体、团队成员以及负责人的角色	●
LO10：沟通：能够就复杂工程问题与业界同行及社会公众进行有效沟通和交流，包括撰写报告和设计文稿、陈述发言、清晰表达或回应指令。并具备一定的国际视野，能够在跨文化背景下进行沟通和交流	
LO11：项目管理：理解并掌握工程管理原理与经济决策方法，并能在多学科环境中应用	
LO12：终身学习：具有自主学习和终身学习的意识，有不断学习和适应发展的能力	

备注：LO=learning outcomes（学习成果）

五、课程目标/课程预期学习成果

序号	课程预期学习成果	课程目标（细化的预期学习成果）	教与学方式	评价方式
1	LO23 具备对复杂工程问题进行分析和求解的能力	1. 掌握获取数据的方式； 2. 掌握存储数据的手段； 3. 掌握大数据场景下，数据处理的基本流程； 4. 掌握数据可视化方式；	案例教学 任务引领 练习实践	作业评价 课堂测试 作品展示
2	LO41 能够基于科学原理，结合智能制造行业，具有将智能制造中	1. 掌握数据的分析手段； 2. 具备结果的分析能力； 3. 并根据分析结果，改进现有流程；	自主学习 实践	实验报告 大作业评审

	关于应用系统开发各方面知识集成的能力，并根据实际对系统设计进行优化			
3	LO91 能够理解团队合作的意义，能与团队成员有效沟通，用人单位评价好	1. 掌握团队分工与协作； 2. 熟悉团队运作方式； 3. 理解自身定位；	自主学习 实践	资料汇总

六、课程内容

第一单元 Python 基础及网络爬虫

介绍大数据的概念、大数据时代到来的背景及课程要求。Python 是一种通用的高级编程语言，具有简洁而易读的语法，适用于多种应用场景，包括 Web 开发、数据分析、人工智能等。结合本课程重点介绍在爬虫技术中用到的基础知识，并加以实践。

理论课时数：0；实验课时数：6

第二单元 数据采集与预处理

数据采集（Data Collection）和预处理（Data Preprocessing）是数据科学和分析中的两个关键步骤，它们涉及从不同来源获取数据并使其适用于分析的过程。介绍数据采集是从不同来源获取数据的过程，其目的是为了构建一个能够支持分析和决策的数据集。数据来源包含但不限于开放数据集、API（应用程序接口）、爬虫技术、传感器数据。数据预处理是为了清理和转换原始数据，以便更好地适应分析或机器学习模型的过程。其主要任务包含缺失值处理、异常值处理、数据转换、数据合并与重塑、特征工程、数据分割、时间序列处理等种类。

理论课时数：0；实验课时数：6

第三单元 静态网页采集

静态网页采集涉及从不包含动态内容或 JavaScript 渲染的网页中获取信息。介绍常见的静态网页采集技术和工具的：1）HTTP 请求库，在进行网页采集时，首先需要向目标网页发送 HTTP 请求。使用 HTTP 请求库可以方便地实现这一步骤。Requests 库：Python 中最常用的 HTTP 请求库之一。它简单而功能强大，可以发送 GET、POST 等请求，并获取服务器的响应。2）HTML 解析库，解析 HTML 代码是从网页中提取数据的关键步骤。HTML 解析库帮助将 HTML 文档转化为易于操作的数据结构。Beautiful Soup：用于解析 HTML 和 XML 文档的 Python 库。它提供了灵活的 API，可以通过标签、类名、ID 等选择器提取数据。lxml：另一个用于解析 HTML 和 XML 的库，通常比 Beautiful Soup 更快速，尤其对于大型文档。

理论课时数：0；实验课时数：6

第四单元 数据存储

大数据存储方式是为了有效地存储、管理和处理大规模数据而设计的系统。介绍分布式文件系统 Hadoop Distributed File System (HDFS)，HDFS 是 Apache Hadoop 项目中的分布式文件系统，用于存储大量数据。它将数据切分成块并分布在集群的多个节点上，提供高可靠性和容错性。列式存储 Apache HBase，HBase 是一个基于 Hadoop 的开源、分布式、列式存储数据库。它适用于快速读取和写入大量数据，支持随机和顺序访问。NoSQL 数据库 Apache Cassandra、MongoDB。列队系统 Apache Kafka，Kafka 是一个分布式的流处理平台，用于实时数据传输和处理。它提供高可用性、持久性和可扩展性，广泛用于日志和事件流处理。上述大数据存储方式通常用于构建大规模、高性能的数据存储和处理系统，以满足现代数据分析和应用的需求。选择合适的存储方式取决于数据的特性、工作负载、性能需求以及云服务的选择。

理论课时数：0；实验课时数：6

第五单元 JavaScript 与动态内容

在爬虫过程中，如果目标网站使用 JavaScript 来渲染动态内容，传统的静态爬虫可能无法获取到完整的页面信息。为了应对这种情况，可以使用一些特定的技术和工具来处理 JavaScript 渲染的页面。JavaScript 在客户端执行，可以在页面加载后修改 DOM（文档对象模型）和进行异步数据请求。传统的静态爬虫只获取页面的初始 HTML 内容，无法获取经 JavaScript 处理后的 DOM 结构和异步加载的数据。介绍处理 JavaScript 渲染的页面的方法无头浏览器是一种没有图形用户界面的浏览器，可以在后台运行并执行 JavaScript。爬虫可以使用无头浏览器模拟用户行为，获取动态生成的内容。

理论课时数：0；实验课时数：6

第六单元 模拟登陆与验证码

模拟登录和处理验证码是爬虫过程中常见的挑战之一。介绍通用的方法和技术，可以帮助处理模拟登录和验证码的情况。模拟登录，使用 Requests 库：对于简单的登录页面，可以使用 Requests 库来发送 POST 请求模拟登录。使用 Session 保持登录状态：使用 Session 对象可以保持登录状态，避免在每个请求中重新登录。处理验证码，介绍手动输入验证码、使用第三方验证码识别服务、使用机器学习模型和规律性验证码处理手段。

理论课时数：0；实验课时数：6

第七单元 爬虫数据的分析与处理

介绍并使用爬虫数据的分析与处理涉及多种技术和工具，数据清理与预处理过程，Pandas，提供了强大的数据结构和数据分析工具，用于处理缺失值、异常值和数据转换。数据分析与统计过程 NumPy、Matplotlib 和 Seaborn 用于数据分析和可视化，并创建各种图表。特征工程过程中 Scikit-learn 提供了丰富的工具和算法，用于特征选择和提取。建模与预测过程中 Scikit-learn 提供了各种机器学习算法和工具。结果解释与报告可使用 Jupyter Notebooks 进行交互式数据分析和报告撰写。

理论课时数：0；实验课时数：12

七、课内实验名称及基本要求

实验序号	实验名称	主要内容	实验时数	实验类型
1	金融数据库设计与实现	金融数据库作为金融机构至关重要的信息支柱，不仅仅存储着海量的金融信息，同时也为金融机构提供信息挖掘、风险控制、运营优化等关键功能。随着金融行业的快速发展，金融数据库的架构和金融数据库应用系统的设计也日趋复杂和精密。本实验将结合课程相关内容，设计一个数据库，确保量化系统的正常运行。	9	设计型
2	并发系统设计与实现	在现代互联网系统中，用户数量和请求量的急剧增加对系统的性能和稳定性提出了更高的要求。高并发架构设计可以有效地解决这些问题，提供良好的用户体验并支持高可用性的运行。本实验在理解高并发架构设计要素的同时，运用常用技术栈，完成一个简易的并发系统。	12	设计型

3	爬取金融数据	作为一种采集和理解网络上海量信息的方式，网页抓取技术变得越来越重要。而编写简单的自动化程序（网络爬虫），一次就可以自动抓取上百万个网页中的信息，实现高效的数据采集和处理，满足大量数据需求应用场景。本实验将利用爬虫技术，实现对互联网中的公开金融信息进行收集、清洗和存储。	18	设计型
4	TA-Lib 库的运用	TA-Lib 是一个 Python 金融指数处理库。包含了很多技术分析里的常用参数指标，例如 MA、SMA、WMA、MACD、ATR 等，可应用于组装量化策略，以求达到收益最大化。本实验将使用 TA-Lib 模块计算相关的指标。	9	设计型

八、评价方式与成绩

总评构成（X）	评价方式	占比
X1	课堂学习 （签到、听讲、讨论、随堂练习等）	30%
X2	实验报告	30%
X3	课程大作业	40%

撰写人：董辛酉　　系主任审核签名：戴智明

审核时间：2024 年 3 月 4 日