数据分析与挖掘

一、基本信息

课程代码：【2050544】
课程学分：【3】
面向专业：【软件工程】
课程性质：【院级选修课◎】
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 网络爬虫与数据采集，吕云翔著，人民邮电出版社，2021 年 9 月】

参考书目：

【人工智能基础（第二版），刘垚主编，华东师范大学出版社，2023 年 3 月第 2 版】

课程网站网址：
先修课程：【计算机导论】

二、课程简介

人工智能是一门极富挑战性的科学，从事这项工作的人必须懂得计算机知识，心理学和哲学等。人工智能是包括十分广泛的科学，它由不同的领域组成，如机器学习，计算机视觉等等。总的来说，人工智能研究的一个主要目标是使机器能够胜任一些通常需要人类智能才能完成的复杂工作。网络爬虫是一个自动提取网页的程序，它为搜索引擎从万维网上下载网页，是搜索引擎的重要组成。传统爬虫从一个或若干初始网页的 URL 开始，获得初始网页上的 URL，在抓取网页的过程中，不断从当前页面上抽取新的 URL 放入队列，直到满足系统的一定停止条件。聚焦爬虫的工作流程较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的 URL 队列。本课程高屋建瓴探讨数据分析与挖掘，内容深入浅出，简单易懂，适合计算机相关专业各个年级学生学习。课程内容包括人工智能基础、网络爬虫技术的基本原理等。

三、选课建议

本课程是适用于计算机类专业、软件工程专业的专业选修课，要求具有计算机导论的基础。

四、课程与专业毕业要求的关联性

专业毕业要求	关联
LO11：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	
LO21：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	●

LO31：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	
LO41：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	
LO51：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	
LO61：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	●
LO71：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	●
LO81：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	
LO91：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	
LO101：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	
LO111：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	
LO121：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	

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	LO211 具备对系统设计、软件开发等涉及到的复杂工程问题进行识别与判断，并结合专业知识进行有效分解的能力	引导学生步入人工智能时代，积极投身数据分析与挖掘的变革浪潮之中	案例教学 任务引领 练习实践	作业评价 课堂测试 作品展示
	LO214 在充分理解专业知识的基础上，能够运用所学知识开展文献检索和资料查询	熟悉人工智能各个环节的相关技术，为后续深入学习相关数据分析与挖掘技术奠定基础	自主学习 实践	资料汇总

2	LO612 熟悉计算机专业领域相关的技术标准、知识产权、产业政策和法律法规	了解数据分析专业知识体系，形成对数据挖掘的整体认知	自主学习 实践	资料汇总
	LO613 能客观评价计算机应用项目的实施对社会、健康、安全、 法律以及文化的影响	了解人工智能的概念，熟悉数据分析应用，培养大数据思维，养成数据安全意识	自主学习 实践	资料汇总
3	LO711/LO712 了解与本专业相关的职业和行业的生产、设计、研究与开发、环境保护和可持续发展等方面的方针、政策和法律、法规。能正确认识并评价计算机科学在现实社会中应用的影响	激发学生基于数据分析的创新创业热情	自主学习 实践	资料汇总

六、课程内容

第一单元 概述及 Python 基础

第 1 讲 基本数据类型
第 2 讲 组合数据类型及程序设计

Python 是一种结合了解释型、编译性和交互式的面向对象计算机编程语言， python 基本数据类型主要是以下几种: 1.数字类型 数字类型主要包括整数类型、浮点类型、和复数类型,整数类型有二进制、八进制、十进制、十六进制这几种表示形式。

组合数据类型更能够将多个同类或不同类型组织起来，通过单一的表示使数据更有序、更容易。根据数据之间的关系，组合数据类型可以分为 3 类：序列类型、集合类型和映射类型。

Python 编程包括三大控制结构，分别是顺序结构、分支结构（选择结构）以及循环结构，控制结构就是控制程序执行顺序的结构。

Python 函数就是一段封装好的，可以重复使用的代码，它使得我们的程序更加模块化，不需要编写大量重复的代码。函数还可以接收参数，并根据参数的不同做出不同的操作，最后再把处理结果返回给我们。函数的本质就是一段有特定功能、可以重复使用的代码。

理论课时数：4；实践课时数：0

第二单元 Python 数据分析及可视化基础

第 4 讲 Numpy
第 5 讲 Pandas
第 6 讲 Matplotlib

在 Python 数据分析领域，被称为"数据分析三剑客"的是 Numpy、Pandas 和 Matplotlib 这三个模块。它们组合在一起，可以让我们轻松地完成数据处理、分析和可视化工作。

理论课时数：6；实践课时数：0

第三单元 Python 智能语音信号处理

第 7 讲 初识智能语音
第 9 讲 绘制语谱图
第 10 讲 智能小义之学会聆听
第 12 讲 智能客服之对话机器人
第 13 讲 智能门锁之声纹识别机器人

智能语音是人工智能领域的一项重要技术，作为新时代的代名词具有很广的应用范围，在社会中具有较高的认知度和发展潜力。智能语音包含语音识别、语义理解、自然语言处理、语音交互等。在历史发展长河中，将智能语音定义为一种人机语言交互技术，它是以语音信号识别为基础，自然语言处理和对话管理技术为辅，将语言输入信息进行提取、分析、整理，最终通过语音合成或文字展示等方式输出并完成响应。

随着科学技术日渐成熟，人类进入了 5G 通信时代。5G 通信技术是具有高速率、低延时和大连接特点的一种通讯技术。它促使智能语音的应用更加广泛，也为语音和语义识别提供了更有利的数据环境，进而丰富了有关智能语音的功能和产品。

理论课时数：10；实践课时数：0

第四单元 Python 网络爬虫

第 14 讲 分析网站及第一个爬虫
第 16 讲 数据采集

第 17 讲 文件和数据存储
第 18 讲 Javascript 与动态内容
第 20 讲 表单与模拟登录
第 21 讲 数据的进一步处理
第 22 讲 更灵活的爬虫
第 24 讲 总复习与展示

Python 网络爬虫是一种使用 Python 编程语言编写的程序，也被称为网页蜘蛛或网络机器人。它的主要功能是按照一定的规则自动浏览和检索网页信息，并能够将所需的数据抓取下来。通过对抓取的数据进行处理和分析，Python 网络爬虫可以提取出有价值的信息。

Python 网络爬虫可以模拟人类使用浏览器上网的行为，自动抓取互联网中的数据。这种技术被广泛应用于搜索引擎、数据挖掘、信息监测等领域。例如，搜索引擎如百度、搜狗、谷歌等都有自己的爬虫程序，用于抓取互联网上的网页信息，以便为用户提供更准确的搜索结果。

Python 作为一种脚本语言，具有语法优美、代码简洁、开发效率高等特点，同时支持多个爬虫模块，如 urllib、requests、Bs4 等。Python 的请求模块和解析模块丰富成熟，还提供了强大的 Scrapy 框架，使得编写爬虫程序变得更加简单和高效。因此，Python 网络爬虫在实际应用中得到了广泛的应用和认可。

需要注意的是，在使用 Python 网络爬虫抓取数据时，需要遵守相关法律法规和网站的 robots.txt 协议，以避免侵犯他人权益和造成不必要的法律纠纷。

理论课时数：12；实践课时数：4

第五单元 上机实践

第 3 讲 Python 的安装及程序设计（线上）
第 8 讲 数据可视化（线上）
第 11 讲 智能小义之学会说话（线上）
第 15 讲 分析网站及第一个爬虫（线上）
第 19 讲 京东用户评论抓取（线上）
第 23 讲 更强大的爬虫（线上）

Python 机器学习实践包含学习 Python 的语法、数据结构、函数、面向对象编程等基础知识；使用 Python 进行数据预处理，包括数据清洗、特征提取、特征选择等；各种常见的机器学习算法，如线性回归、逻辑回归、决策树、随机森林、SVM、神经网络等学习使用 Scikit-learn、Pandas、NumPy、Matplotlib 等常用的 Python 机器学习库。

Python 网络爬虫实践包含了解网络爬虫的基本概念、工作原理和分类；学习如何使用 Python 的库（如 requests、urllib）发送 HTTP 请求，并处理响应内容；使用正则表达式、BeautifulSoup、lxml 等工具解析网页内容，提取所需数据；学习将爬取的数据保存到文件、

数据库或其他存储介质中；实践爬取特定网站的数据，如新闻、社交媒体、电商网站等，并进行数据分析和可视化等。

理论课时数：0；实践课时数：12

七、课内实验名称及基本要求（选填，适用于课内实验）

列出课程实验的名称、学时数、实验类型（演示型、验证型、设计型、综合型）及每个实验的内容简述。

序号	实验名称	主要内容	实验时数	实验类型	备注
1	Python 安装	Python 的安装及程序设计，掌握 Python 程序设计与调试方法	2	上机实践	
2	数据可视化	数据可视化常用方法的使用	2	上机实践	
3	语音数据处理	掌握语音数据处理，掌握智能小义之学会说话	2	上机实践	
4	网站爬取	分析网站及第一个爬虫	2	上机实践	
5	网站数据爬取分析	京东用户评论抓取	4	上机实践	
6	爬虫综合	综合运用，更强大的爬虫	4	上机实践	

八、评价方式与成绩

总评构成（1+X）	评价方式	占比
1	期末展示	50%
X1	课堂学习	10%
X2	课后小论文	20%
X3	在线学习	20%

撰写人：彭青松　　系主任审核签名：戴智明

审核时间：2024 年 2 月 29 日