大数据导论

一、基本信息

课程代码：【2050268】
课程学分：【2】
面向专业：【计算机科学与技术】
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：

【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

大数据时代已经全面开启，带来了信息技术发展的巨大变革，并深刻影响着社会生产和人民生活的方方面面。了解大数据概念、具备大数据思维，熟悉大数据技术是新时代对人才的新要求。本课程高屋建瓴探讨大数据，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。课程内容包括大数据爬虫技术、爬虫技术的基本原理、常见的爬虫框架等。

三、选课建议

本课程是适用于计算机类专业的专业选修课，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	引导学生步入大数据时代，积极投身大数据的变革浪潮之中	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	熟悉大数据各个环节的相关技术，为后续深入学习相关大数据技术奠定基础	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	了解大数据专业知识体系，形成对大数据专业的整体认知	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	了解大数据概念，熟悉大数据应用，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	激发学生基于大数据的创新创业热情	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 单元第 1 讲 概述

数据的概念、大数据时代到来的背景、课程要求

第 1 单元第 2 讲 Python 函数

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。

第 1 单元第 3 讲 Python 的安装与扣叮（线上）

Python 作为一种流行的高级编程语言，被广泛应用于数据分析、机器学习、网络编程等方向。如果你是一名想要学习 Python 的新手，那么第一步就是要在你的计算机上安装这个编程语言。对于开始安装有困难的同学，建议使用腾讯扣叮。

理论课时数：6；实验课时数：0

第二单元 静态网页抓取

第 2 单元第 1 讲 正则表达式

正则表达式对于程序编写而言是一个复杂的话题，它为了更好地“匹配”或者“寻找”某一种字符串而生。正则表达式常常用来描述一种规则，而通过这种规则，我们就能够更方便地查找邮箱地址或者筛选文本内容。

第 2 单元第 2 讲 BeautifulSoup

BeautifulSoup 是一个可以从 HTML 或 XML 文件中提取数据的 Python 库.它能够通过你喜欢的转换器实现惯用的文档导航,查找,修改文档的方式.BeautifulSoup 会帮你节省数小时甚至数天的工作时间。

第 2 单元第 3 讲 BeautifulSoup（线上）

通过程序实例，实现 BeautifulSoup 程序设计。

第 2 单元第 4 讲 XPath 与 XML

一般使用 lxml 这个库来处理 XPath，如果机器上没有安装 lxml，首先还是得用 pip install lxml 命令来进行安装，安装时可能会出现一些问题（这是由于 lxml 本身的特性造成的），另外，lxml 还可以使用 easy install 等方式安装，这些都可以参照 lxml 官方的说明。

第 2 单元第 5 讲 遍历页面与 API

爬虫程序所要面对的任务经常是根绝某种抓取逻辑，重复遍历多个页面甚至多个网站。这可能也是爬虫(蜘蛛)这个名字的由来——就像蜘蛛在网上爬行一样。在处理当前页面时，爬虫就应该考虑确定下一个将要访问的页面,下一个页面的链接地址有可能就在当前页面的某个元素中，也可能是通过特定的数据库读取（这取决于爬虫的爬取策略），通过从“爬取当前页”到“进入下一页”的循环，实现整个爬取过程。

第 2 单元第 6 讲 XPath 实战（线上）

通过程序实例，实现 XPath 程序设计。

理论课时数：12；实验课时数：0

第三单元 数据存储

第 3 单元第 1 讲 文件读写

Python 在文件读写操作中,会使用「内置函数」和「Pandas 库」两种方式。先来看内置函数,包括 open()、read()、readline()、readlines()、write()、writelines()、close()等方法。

第 3 单元第 2 讲 数据库应用

在 Python 中进行数据库操作需要通过特定的程序模块（API）来实现，其基本逻辑是，首先导入接口模块，然后通过设置数据库名、用户、密码等信息来连接数据库，接着执行数据库操作（可以通过直接执行 SQL 语句等方式），最后关闭与数据库的连接。

第 3 单元第 3 讲 文件与数据库实战（线上）

通过实例，讲解文件与数据库实战。

理论课时数：6；实验课时数：0

第四单元 表单与模拟登录

第 4 单元第 1 讲 模拟登录

在各式各样的网页中，有些网站页面是基于注册登录功能的，很多内容对于尚未登录的游客并不开放。目前的趋势是，各式网站都在朝着更社交、更注重用户交互的方向发展，因此，在爬虫编写中考虑账号登录的问题就显得很有必要。

第 4 单元第 2 讲 验证码

目前的网站在验证用户身份这个问题上总是精益求精，不惜下大力气防范非人类的访问，对于大型商业性网站而言尤其如此——最大的障碍在于验证码，验证码问题始终是程序模拟登录过程中最为头疼的一环，也可能是所有爬虫程序所要面对的最大问题之一。

第 4 单元第 3 讲 复习

线下总复习。

第 4 单元第 4 讲 验证码实战（线上）

通过实例，讲解验证码实战。

理论课时数：8；实验课时数：0

总评构成（1+X）	评价方式	占比
X1	课堂学习 （签到、听讲、讨论、随堂练习等）	10%

七、评价方式与成绩

X2	随堂测试	20%
X3	在线学习	20%
X4	期末大作业 （开卷当堂完成）	50%

撰写人：彭青松　　系主任审核签名：戴智明

审核时间：2023 年 9 月 4 日