**《数据采集与预处理》课程教学大纲**

**一、课程简介**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **课程中文名** | **数据采集与预处理** | | | | | | |
| **课程英文名** | **Data collection and preprocessing** | | | | **双语授课** | | □是 ☑否 |
| **课程代码** | **10112164** | **课程学分** | **4** | **总学时数** | | 64（含实践32） | |
| **课程类别** | □通识教育课程  □公共基础课程  ☑专业教育课程  □综合实践课程  □教师教育课程 | **课程性质** | □必修  ☑选修  □其他 | **课程形态** | | □线上  ☑线下  □线上线下混合式  □社会实践  □虚拟仿真实验教学 | |
| **考核方式** | □闭卷 □开卷 □课程论文 ☑课程作品 □汇报展示 ☑报告  ☑课堂表现 □阶段性测试 ☑平时作业 □其他 （可多选） | | | | | | |
| **开课学院** | 大数据与智能工程 | | **开课**  **系(教研室)** | 数据科学与大数据技术 | | | |
| **面向专业** | 计算机科学与技术 | | **开课学期** | 第3学期 | | | |
| **课程负责人** | 黄金龙 | | **审核人** | 曾俊、张素兰、段江丽 | | | |
| **先修课程** | Python程序设计 | | | | | | |
| **后续课程** | 无 | | | | | | |
| **选用教材** | 1. 朱晓姝, 许桂秋. 大数据预处理[M]. 北京:人民邮电出版社, 2019. | | | | | | |
| **参考书目** | 1. [Matt Casters](http://book.jd.com/writer/Matt%20Casters_1.html)（[马特·卡斯特](http://book.jd.com/writer/马特·卡斯特_1.html)）, [Roland Bouman](http://book.jd.com/writer/Roland%20Bouman_1.html)（[罗兰·布曼](http://book.jd.com/writer/罗兰·布曼_1.html)）, [Jos van Dongen](http://book.jd.com/writer/Jos%20van%20Dongen_1.html)（[乔斯·凡·东恩](http://book.jd.com/writer/乔斯·凡·东恩_1.html)）. Pentaho Kettle解决方案：使用PDI构建开源ETL解决方案[M] 北京:[电子工业出版社](http://book.jd.com/publish/电子工业出版社_1.html), 2014. | | | | | | |
| **课程资源** | 1.数据获取与预处理. 网易云课堂.  <https://mooc.study.163.com/course/2001353005?tid=2001578010#/info> | | | | | | |
| **课程简介** | 数据采集与预处理是计算机科学与技术技术专业的选修课程，旨在让学生掌握主流的数据采集技术和大数据预处理方法和工具，具备采集大数据并对其进行预处理的能力。课程内容包括正则表达式、网络爬虫、数据预处理背景及其意义、数据预处理理论方法、数据预处理工具（Kettle）、Kettle的使用等。 | | | | | | |

**二、课程目标**

**表 2-1 课程目标**

|  |  |
| --- | --- |
| **序号** | **具体课程目标** |
| **课程目标 1** | 能够应用数学、计算机科学和数据采集与预处理的基本原理，在数据采集与预处理流程的构思与设计阶段，通过文献研究、实验试验、工程推理、数学建模等方法，识别、表达、分析复杂数据采集与预处理问题及其解决方法、识别和判断数据采集与预处理问题的关键环节和参数，已获得数学模型、工程知识库等有效结论。 |
| **课程目标 2** | 能够利用数据采集与预处理基础知识，以及计算机及软件工程等相关领域的专业知识、技能与工具，设计针对数据采集与预处理问题的解决方案，解决实际数据采集与预处理全过程中遇到的复杂工程问题，并能够在设计环节中体现创新意识，考虑社会、健康、安全、法律、文化以及环境等因素。 |
| **课程目标** 3 | 能够针对数据采集与预处理过程中涉及的复杂工程问题，选择和使用恰当的数据采集和预处理工具，进行信息的采集、预处理、表达、建模、设计、模拟、验证，并能够在实践中了解国内主流数据采集和预处理平台在不同应用场景中的优势和局限性；树立使用国产软件、保护版权的意识，激发民族自豪感；坚持自主创新，为建设世界科技强国而奋斗。 |

**表2-2 课程目标与毕业要求对应关系（计算机科学与技术专业）**

| **毕业要求** | **指标点** | **课程目标** |
| --- | --- | --- |
| **毕业要求3.设计/开发解决方案：**能够针对计算机复杂工程问题的解决方案，设计并实现满足特定需求的软件系统，并能够在设计和实现环节中体现创新意识，考虑社会、健康、安全、法律、文化以及环境等因素。【H】 | 3.2 能够识别计算机软件系统设计与实现中的关键问题，并对方案进行持续优化，体现创新性。 | 课程目标1 |
| **毕业要求5.使用现代工具：**能够针对计算机复杂工程问题，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，包括对计算机复杂工程问题的预测与模拟，并能够理解其局限性。【M】 | 5.1 掌握常用软硬件开发环境和开发工具的性能、适用范围以及对开发、分析的适用性，并能正确使用。 | 课程目标2 |
| **毕业要求6.工程与社会：**能够基于工程背景知识进行合理分析，评价计算机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。【L】 | 6.1 具有社会、健康、安全、法律以及文化意识，能够认识到计算机软件、硬件、网络和相应的复杂工程问题在这几方面的影响，并能够基于工程相关背景知识进行合理分析。 | 课程目标3 |

**三、课程学习内容与方法**

**（一）理论学习内容及要求**

**表3-1 课程目标、学习内容和教学方法对应关系**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **序号** | **课程模块** | **学习内容** | **学习任务** | **课程目标** | **学习重点难点** | **教学方法** | **学时** |
| 1 | 数据采集 | 1.数据采集概述 | 1.拓展阅读：阅读数据采集相关的文献，查阅网络爬虫相关案例，学习国内主流的数据采集工具 | 课程目标1 | 重点：  1.数据采集技术算法的实现  2.网络爬虫算法实现  难点：  3.正则表达式的应用 | 1.课堂讲授：能够引导学生对数据采集的兴趣促进学生对正则表达的理解 | 4 |
| 2.数据采集技术算法 |
| 3.正则表达式 | 课程目标2 |
| 4.网络爬虫 |
| 5.国内主流的数据采集工具 | 课程目标3 |
| 2 | 数据预处理综述 | 1.数据预处理背景 | 1.拓展阅读：阅读预处理相关的文献，学习国内主流的数据预处理工具 | 课程目标1 | 重点：  1.数据预处理的目的  2.数据与处理流程 | 1.讲授法：能够引导学生理解数据预处理概念，掌握数据预处理的目的与流程 | 4 |
| 2.数据预处理的目的 |
| 3.数据预处理的流程 |
| 4.国内主流的数据预处理的工具 | 课程目标3 |
| 3 | 数据预处理技术 | 1.数据清理 | 1.个人作业：完成课后布置的数据归约、数据变换等作业 | 课程目标1 | 重点：  1.数据清理  2.数据集成  难点：  1.数据规约  2.数据变换 | 1.讲授法：能够引导学生熟悉并掌握数据预处理相关技术方法 | 5 |
| 2.数据集成 |
| 3.数据归约 | 课程目标2 |
| 4.数据变换 |
| 4 | Kettle工具的初步使用 | 1.Kettle简介及其安装 | 1.拓展阅读：阅览kettle相关数据，熟悉其使用方法及应用场景 | 课程目标3 | 重点：  1.Kettle使用方法及应用场景 | 1.实验指导：能够促使学生更快的熟悉Kettle工具 | 2 |
| 2.Kettle使用方法及应用场景 |
| 5 | 数据导入与导出 | 1.基于文件的数据导入与导出 | 1.个人作业：完成课后布置的不同数据源的数据导入与导出任务 | 课程目标2 | 重点：  1.基于文件的数据导入与导出  难点：  2.基于数据库的数据导入与导出 | 1.实验指导：  能够促使学生更快的使用Kettle进行数据的导入与导出 | 4 |
| 2.基于数据库的数据导入与导出 |
| 3.基于Web的数据导入与导出 |
| 4.基于CDC变更数据的导入与导出 |
| 6 | Kettle数据清洗 | 1.Kettle数据清洗概述 | 1.个人作业：完成课后布置的利用Kettle清洗数据的任务 | 课程目标1 | 重点：  1.字符串处理  2.数据排重  难点：  1.数据校验  2.Kettle的脚本使用 | 1.实验指导：能够促使学生更快的掌握Kettle争对不同问题的解决方法 | 6 |
| 2.Kettle字符串处理 | 课程目标2 |
| 3.Kettle数据排重 |
| 4.Kettle数据校验 |
| 6.Kettle使用脚本进行数据清理 | 课程目标3 |
| 7 | 构建数据仓库 | 1.构建维度表 | 1.拓展阅读：查阅数据仓库相关概念，理解维度表和事实表 | 课程目标2 | 重点：  1.构建维度表  2.构建事实表 | 1.课堂讲授：能够引导学生掌握数据仓库及维度表和事实表的概念 | 3 |
| 2.构建事实表 |
| 3.数据剖析 |
| 8 | Kettle作业设计 | 1.作业的概念及组成 | 1.个人作业：完成课后布置的Kettle作业创建及执行等任务 | 课程目标1 | 重点：  1.作业的执行方式  难点：  1.变量与监控 | 1.实验指导：能够促使学生更快的掌握Kettle作业的概念及其执行方式 | 4 |
| 2.作业的执行方式 | 课程目标2 |
| 3.作业创建及常用作业项 |
| 4.变量与监控 | 课程目标3 |

1. **实验学习内容及要求**

**表3-2 课程目标、学习内容和教学方法对应关系**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **序号** | **项目名称** | **项目来源** | **教学目标（观测点、重难点）** | **学时数** | **项目类型** | **要求** | **每组人数** | **教学方法** | **课程目标** |
| 1 | 实验1：复杂表头的Excel数据源处理 | 实验教材 | 1.掌握转换的创建方法 | 2 | 设计研究 | 必做 | 1 | 实验指导 | 课程目标1 |
| 2.掌握步骤的参数配置 |
| 3.运行转换并输出结果 |
| 2 | 实验2：合并多个excel文件 | 实验教材 | 1.完成Excel文件输入输出步骤的配置 | 2 | 设计研究 | 必做 | 1 | 实验指导 | 课程目标2 |
| 2.理解多个文件合并的原理 |
| 3. 完成Excel文件输出步骤的配置 |
| 3 | 实验3：用Kettle生成测试数据 | 实验教材 | 1.掌握测试数据随机生成方法 | 2 | 设计研究 | 必做 | 1 | 实验指导 | 课程目标3 |
| 2.完成测试数据步骤参数配置 |
| 4.完成测试数据转换输出 |
| 4 | 实验4：数据全量、增量、比较更新 | 实验教材 | 1.理解数据全量、增量概念 | 4 | 设计研究 | 必做 | 1 | 实验指导 | 课程目标1 |
| 2.完成数据全量、增量及比较更新操作 |
| 3.完成转换运行输出 |
| 5 | 实验5：字符串操作 | 实验教材 | 1.掌握字符串去空方法 | 6 | 设计研究 | 必做 | 1 | 实验指导 | 课程目标3 |
| 2.完成字符串值替换操作 |
| 3.掌握字符串补位操作 |
| 4.完成字符串操作的转换运行输出 |
| 6 | 实验6：字段的拆分、合并、值映射 | 实验教材 | 1.完成字段拆分步骤的参数配置 | 4 | 设计研究 | 必做 | 1 | 实验指导 | 课程目标2 |
| 2.完成字段合并步骤的参数配置 |
| 3.完成字段值映射步骤的参数配置 |
| **7** | 实验7：去除重复的数据 | 实验教材 | 1.掌握去除重复数据的原理方法 | 4 | 设计研究 | 必做 | 1 | 实验指导 | 课程目标2 |
| 2.理解数据对象之间相似度的计算方法 |
| 3.完成数据去重转换的运行输出 |
| 8 | 实验8：异常数据分流 | 实验教材 | 1.理解异常数据的概念 | 4 | 设计研究 | 必做 | 1 | 实验指导 | 课程目标1 |
| 2.掌握异常数据分流的转换设计 |
| 3.完成数据分流转换的运行输出 |
| 9 | 实验9：数据质量统计 | 实验教材 | 1.理解数据质量的衡量方法 | 4 | 设计研究 | 必做 | 1 | 实验指导 | 课程目标3 |
| 2.掌握常用的数据质量标准 |
| 3.完成数据质量统计的转换运行输出 |

## 四、课程考核

**（一）考核内容与考核方式**

**表4-1 课程目标、考核内容与考核方式对应关系**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **课程目标** | **考核内容** | **所属**  **学习模块/项目** | **考核占比** | **考核方式** |
| 课程  目标 1 | 1.数据采集与预处理的目的、流程 | 模块1、2、3、6、8  项目1-4 | 46% | 课堂交互  实验报告  课后练习  闭卷考试 |
| 2.数据清洗的基本原理、目的、流程和经典方法 |
| 3.作业的概念、组成及其执行方式 |
| 4.数据集的合并、重塑和转换 |
| 课程  目标 2 | 1.文件数据、数据库数据、web数据等的导入与导出操作 | 模块1、3、5、6、7、8  项目5-7 | 35% | 课堂交互  实验报告  课后练习  闭卷考试 |
| 2.维度表和事实表的构建，数据的剖析及捕获 |
| 3.数据排重，脚本组件进行数据清理 |
| 4.作业创建及常用作业项，变量与监控 |
| 5.文本文件、二进制文件以及数据库文件的导入导出 |
| 课程  目标 3 | 1.国内主流的数据采集与预处理工具的优势和局限性 | 模块1、2、4、6、8  项目8-9 | 19% | 课堂交互  实验报告  课后练习  闭卷考试 |
| 2.Kettle工具的安装、操作 |
| 3.Kettle使用方法及应用场景 |

**表4-2 课程目标与考核方式矩阵关系**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课程  目标 | 考核方式 | | | | | | | | 考核占比 | |
| 平时考核（40%） | | | | | | 期末考核  （60%） | |
| 课堂交互（10%） | | 实验报告  （20%） | | 课后练习  （10%） | |
| 课程目标1 | 60% | 60% | | 40% | | 40% | | 46%=10%\*60%+20%\*60%+10%\*40%+60%\*40% | |
| 课程目标2 | 30% | 20% | | 40% | | 40% | | 35%=10%\*30%+20%\*20%+10%\*40%+60%\*40% | |
| 课程目标3 | 10% | 20% | | 20% | | 20% | | 19%=10%\*10%+20%\*20%+10%\*20%+60%\*20% | |

## （二）成绩评定

**1.平时成绩评定**

**（1）课堂交互（25 分）**：通过学生在课堂上发言、回答提问情况，评价学生的课程参与能力。（包括理想信念、家国情怀、社会责任、学习态度、学习兴趣、团队合作与终身学习意识等）

**（2）实验报告（50分）**：围绕课程的学习目标进行作业的设计。如让学生简述大数据的认识，考核学生对于大数据相关概念的理解情况，帮助学生将定义转化为自己的理解。

**（3）课后练习（25分）**：学生收集课程知识相关资料的能力，研究并提出问题的能力，解决实际问题能力和合作研究能力；

**2.期末成绩评定**

期末考试主要考察学生对数据采集与预处理知识体系、研究范畴和研究方向，数据采集与预处理领域的认识和构架的基本概念等，要求学生掌握使用数据采集与预处理工作平台Kettle进行数据处理的方法、各种类型数据在实践中的采集与处理能力等，掌握专业相关复杂工程问题与社会可持续发展的客观评价能力和反思能力。方式为闭卷考试。

**3.总成绩评定**

总成绩由平时成绩、期末开卷试卷考试成绩构成。

总成绩（100%）=平时成绩（40%）+期末闭卷试卷考试成绩（60%）

## （三）评分标准

以试卷参考答案及评分细则为准。

## 五、其它说明

本课程大纲依据2023版数据科学与大数据计算专业人才培养方案，由大数据与智能工程学院（部）数据科学与大数据技术教学系（教研室）讨论制定，大数据与智能工程学院（部）教学工作委员会审定，教务处审核批准，自2023级开始执行。