**《海量数据预处理综合实训》实训课程教学大纲**

**字体、字号请参考范例**

**注意：**

**首字母大写**

**植物拉丁学名斜体**

**一、课程简介**

|  |  |
| --- | --- |
| **课程中文名** | 海量数据预处理综合实训 |
| **课程英文名** | Comprehensive Training of Massive Data Preprocessing | **双语授课** | □是 ☑否 |
| **课程代码** | 10114083 | **课程学分** | 2 | **周（学时）** | 48学时 |
| **课程类别** | □专业认知实习□专业见习☑工程实训□毕业实习□其他  | **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** | 大数据与智能工程学院 | **开课****系(教研室)** | 数据科学与大数据技术 |
| **面向专业** | 数据科学与大数据技术 | **开课学期** | 第5学期 |
| **课程负责人** | 黄金龙 | **审核人** | 张素兰 |
| **先修课程** | 数据采集与预处理 |
| **后续课程** | 无 |
| **选用教材** | 1. 斯夸尔（Megan Squire） 著、任政委译. 大数据策略. 人民邮电出版社，2016.
 |
| **参考书目** | 1. 朝乐门. 数据科学. 清华大学出版社，2016.
2. Matt Casters（马特·卡斯特），Roland Bouman（罗兰·布曼），Jos van Dongen（乔斯·凡·东恩） 著，初建军译. Pentaho Kettle解决方案：使用PDI构建开源ETL解决方案. 电子工业出版社，2014.
 |
| **课程资源** | 1. 山东大学公开课：数据预处理

http://open.163.com/newview/movie/free?pid=MELPPEPSH&mid=MEMB095TG |
| **课程简介** | 海量数据预处理综合实训是数据科学与大数据技术专业必修的综合实践课程。该课程内容包括数据采集与预处理的基础理论技术、Kettle作业的概念及其创建、子作业和子转换的创建与运行、数据集成、数据转换、数据分析等方法。使学生能够选择不同的数据预处理组件和技术方法，解决国内主流的大数据应用场景中存在的问题。 |

**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** | 能够应用计算机科学和软件工程、物联网等领域的知识，在数据采集与预处理的构思与设计阶段，通过文献研究、实验试验等方法，识别、表达数据导入与预处理的问题及其解决方法。 |
| **课程目标2** | 能够应用数据采集与预处理的理论与方法，以及物联网等相关领域的专业知识、技能与工具，发现数据采集与预处理过程中的问题，提出解决方案；能够在设计环节中体现创新意识、版权保护意识、家国情怀和社会责任，坚持自主创新和团队合作，为建设世界科技强国而奋斗；能够掌握国内大数据常用的大数据开发工具，了解数据采集技术的优势和局限性。 |
| **课程目标3** | 能够设计针对数据采集与预处理的解决方案，大数据系统开发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，进行信息采集、建模、设计、模拟、验证，并能够在实践中了解国内常用的数据采集和预处理工具在不同应用场景中的优势和局限性。 |

**表2-2 课程目标与毕业要求对应关系**

| **毕业要求** | **指标点** | **课程目标** |
| --- | --- | --- |
| **毕业要求5：**能够针对大数据应用领域的复杂工程问题，选择和使用恰当的技术、资源和现代工程工具，进行信息表达、建模、设计、模拟、验证，并能够在实践中了解这些工具使用的局限性。【M】 | 5.1了解大数据应用领域常用的软硬件平台、数据处理工具、现代工程工具和大数据框架的使用原理和方法，并理解其局限性。 | 课程目标3 |
| **毕业要求9：**具有良好的综合素质，能够在多学科背景下的团队中理解与承担个体、团队成员以及负责人角色，并发挥相应的作用。【M】 | 9.2能够适应多学科背景的团队合作方式，具备一定的组织管理能力，并能综合团队成员的意见，进行合理决策。 | 课程目标1 |
| **毕业要求11：**能够理解并掌握工程管理与经济决策方法，熟悉大数据应用领域工程项目管理的基本方法和技术，并能在多学科环境中应用。【H】 | 11.1理解从事大数据应用领域工程实践活动所需的管理与经济因素，掌握计算机工程管理和经济决策方法。 | 课程目标2 |

**三、教学内容及要求**

**（一）学习内容**

A.通过本次课程设计，进一步掌握和应用数据预处理的知识，需掌握的内容包括：

1）数据属性与度量，每一个数据对象用一组属性描述，数据集是用结构化数据表表示。

2）数据集类型，从数据对象之间的结构关系角度进行划分，比较常见的有记录数据、有序数据、图形数据。

3）数据清洗，主要包括：测量误差、数据收集错误、噪声、离群点（outlier）、缺失值、不一致值、重复数据等问题。

5）数据集成，将若干个分散的数据源中的数据，逻辑地或物理地集成到一个统一的数据集合中。

5）数据规范化，数据规范化主要包括数据同趋化处理和无量纲化处理两个方面。

6）数据约简，在尽可能保持原始数据集完整性的前堤下，最大限度地精简数据量。

7）数据离散化，要求数据属性是标称类别，当数据中包含数值属性时，为了使用这些算法需要将数值属性转换成标称属性。

8）数据变换，将数据变换或统一成适合挖掘的形式。

B.从下列项目名中选择，或根据情况设定项目名称，按表中所列步骤完成该课程：

1）公司财务数据预处理

2）超市销售数据预处理

3）网站流量数据预处理

4）学生成绩数据预处理

5）楼盘数据分析预处理

6）大数据职位数据预处理

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 项目名称 | 学习目标 | 学习内容 | 主要仪器设备 |
| 1 | 数据类型认识 | 掌握数据属性与度量方法及数据集的类型和常用表示方法 | 1. 数据属性与度量
2. 数据集类型
 | 计算机，及以下软件：Linux虚拟机、Hadoop及其相关组件等 |
| 2 | 数据清洗 | 了解数据清洗常用方式方法，及具体操作。 | 1. 测量误差
2. 数据收集错误
3. 噪声
4. 离群点（outlier）
5. 缺失值
6. 不一致值
7. 重复数据
 | 计算机，及以下软件：Linux虚拟机、Hadoop及其相关组件等 |
| 3 | 数据集成 | 掌握将若干个分散的数据源中的数据，逻辑地或物理地集成到一个统一的数据集合中。 | 1. 整合加工形成可用的数据
2. 将数据以符合规范的方式发布
 | 计算机，及以下软件：Linux虚拟机、Hadoop及其相关组件等 |
| 4 | 数据规范化 | 了解数据同趋化处理和无量纲化处理两个方面的方法。 | 1. 数据同趋化处理
2. 数据无量纲化处理
 | 计算机，及以下软件：Linux虚拟机、Hadoop及其相关组件等 |
| 5 | 数据变换 | 掌握精简数据量方法，了解数据数据属性转换。 | 1. 精简数据量方法
2. 数据属性转换
 | 计算机，及以下软件：Linux虚拟机、Hadoop及其相关组件等 |
| 6 | 数据预处理分析 | 掌握不同的数据分析方法 | 1. 探索性数据分析
2. 模型选定分析
3. 推断分析
 | 计算机，及以下软件：Linux虚拟机、Hadoop及其相关组件等 |

**（二）时间安排**

|  |  |  |
| --- | --- | --- |
|  | 上午 | 下午 |
| 周一 | 教师讲解 | 认识数据集类型 |
| 周二 | 数据清洗（测量误差、数据收集错误） | 数据清洗（噪声、离群点（outlier）） |
| 周三 | 数据清洗（缺失值、不一致值） | 数据清洗（重复数据） |
| 周四 | 教师讲解 | 整合加工形成可用的数据 |
| 周五 | 教师讲解 | 将数据以符合规范的方式发布 |
| 周一 | 教师讲解 | 数据同趋化处理 |
| 周二 | 教师讲解 | 数据无量纲化处理 |
| 周三 | 教师讲解 | 精简数据量方法 |
| 周四 | 数据属性转换 | 探索性数据分析 |
| 周五 | 模型选定分析 | 推断分析 |

**（三）工作流程**

1)技术讲解：教师应用讲授法、演示法、实验法、任务驱动法、讨论法、自主学习法，讲解综合实训应用的知识点,及各个软件的使用要点。

2)问题分析：通过对项目需要解决的问题进行分析，确定各个步骤的技术方案，书写对问题分析的文档。

3）技术方案设计：对项目所需要的总体技术方案进行设计。

4) 工具使用：通过VT完成对二维数据、三维数据的可视化操作；通过mongo shell、Robomongo完成对数据的可视化操作；利用excel的api完成各种类型图形的开发；利用ECharts完成对数据的可视化及挖掘。

5)数据挖掘：通过对可视化工具的使用完成对数据的挖掘。

6)撰写总结材料：总结实验过程，并形成实验报告，分析收获和不足。

7)评分：根据评分标准对每个同学的项目的设计、完成度、项目演示、文档撰写等结果进行综合评分。

**（四）业务指导**

在综合课程设计期间，由校内老师1名和企业老师1名，每天深入机房，随时指导解答问题。

**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** | **所属环节** | **考核占比** | **考核方式** |
| 课程目标 1 | 认识数据集类型 | 数据属性与度量方法及数据集的类型 | 46% | 课堂表现实验报告团队协作期末考核 |
| 数据清洗（测量误差、数据收集错误） | 数据清洗方法 |
| 数据清洗（缺失值、不一致值） | 数据清洗方法 |
| 数据清洗（噪声、离群点（outlier）） | 数据清洗方法 |
| 数据清洗（重复数据） | 数据清洗方法 |
| 课程目标 2 | 探索性数据分析 | 数据预处理分析 | 35% | 课堂表现实验报告团队协作期末考核 |
| 推断分析 | 数据预处理分析 |
| 模型选定分析 | 数据预处理分析 |
| 课程目标 3 | 整合加工形成可用的数据 | 数据集成 | 19% | 课堂表现实验报告团队协作期末考核 |
| 将数据以符合规范的方式发布 | 数据集成 |
| 数据同趋化处理 | 数据规范化 |
| 数据无量纲化处理 | 数据规范化 |
| 精简数据量方法 | 数据变换 |
| 数据属性转换 | 数据变换 |

**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%） |
| 课堂表现（10%） | 实验报告（20%） | 团队协作（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.平时成绩评定**

1）课堂表现（25 分）：通过学生在课堂上的表现情况、提问与回答、与教师互动情况，评价学生相关的能力。（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）：学生在实训过程中协调、沟通等的情况。

3）团队协作（25分）：学生在实训过程中相关任务的完成情况。

**2.期末成绩评定**

期末成绩的评定为作品设计成绩，即系统代码实现和作业报告。

**3.总成绩评定**

总成绩（100%）=平时成绩（40%）+期末成绩（60%）

**（三）评分标准**

1)优秀（90-100分）

按设计任务书要求圆满完成规定任务；综合运用知识能力和实践动手能力强，软件设计方案合理，实验效果好；设计态度认真，独立工作能力强，并具有良好的团队协作精神。设计报告条理清晰、论述充分、图表规范、符合设计报告文本格式要求。 答辩过程中，思路清晰、论点正确、对设计方案理解深入，问题回答正确。

2)良好（80-89分）

按设计任务书要求完成规定设计任务；综合运用知识能力和实践动手能力较强，软件设计方案较合理，实验效果较好；设计成果质量较高；设计态度认真，有一定的独立工作能力，并具有较好的团队协作精神。设计报告条理清晰、论述正确、图表较为规范、符合设计报告文本格式要求。 答辩过程中，思路清晰、论点基本正确、对设计方案理解较深入，主要问题回答基本正确。

3)中等（70-79分）

按设计任务书要求完成规定设计任务；能够一定程度的综合运用所学知识，软件设计基本合理，有一定的实践动手能力，设计成果质量一般；设计态度较为认真，设计报告条理基本清晰、论述基本正确、文字通顺、图表基本规范、符合设计报告文本格式要求，但独立工作能力较差；答辩过程中，思路比较清晰、论点有个别错误，分析不够深入。

4)及格（60-69分）

在指导教师及同学的帮助下，能按期完成规定设计任务；综合运用所学知识能力及实践动手能力较差，设计方案基本合理，设计成果质量一般；独立工作能力差；或设计报告条理不够清晰、论述不够充分但没有原则性错误、文字基本通顺、图表不够规范、符合设计报告文本格式要求；或答辩过程中，主要问题经启发能回答，但分析较为肤浅。

5)不及格（60分以下）

未能按期完成规定设计任务。不能综合运用所学知识，实践动手能力差，设计方案存在原则性错误，计算、分析错误较多；或设计报告条理不清、论述有原则性错误、图表不规范、质量很差；或答辩过程中，主要问题阐述不清，对设计内容缺乏了解，概念模糊，问题基本回答不出。

## 五、其他说明

本课程大纲依据2023版数据科学与大数据技术专业人才培养方案，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。

**字体、字号请参考范例**

**注意：**

**首字母大写**

**植物拉丁学名斜体**

**字体、字号请参考范例**

**注意：**

**首字母大写**

**植物拉丁学名斜体**

**字体、字号请参考范例**

**注意：**

**首字母大写**

**植物拉丁学名斜体**