《数据分析》课程

整体设计

一、课程基本信息

|  |  |  |  |
| --- | --- | --- | --- |
| 课程代码 | 250533 | 课程性质 | 必修 |
| 适用专业 | 云计算技术与应用 | 开设学期 | 第2学年第1学期 |
| 课程类别 | 专业平台课程 | 课程类型 | B类（理论+实践） |
| 学 分 | 6.5 | 总 学 时 | 114 |
| 学时分配 | 理论学时： 57 ；实践学时： 57 |
| 实施场所 | 云计算实训室 | 授课方式 | 理实一体化 |
| 执笔人 | 郭峰 |
| 审核人 |  |
| 制订时间 | 2020.8.9 |

二、课程概述

（一）课程定位

《数据分析》课程是云计算技术与应用专业的一门必修专业基础课，涉及科学计算库NumPy、数据分析工具Pandas、数据可视化、时间序列分析和文本数据分析。

通过对本课程的学习，学生能够熟悉数据分析的流程和思想，可以利用数据分析技术解决特定业务领域的问题。

（二）先修后续课程

|  |  |  |
| --- | --- | --- |
| 序号 | 先修课程名称 | 为本课程支撑的主要能力 |
| 1  | Python程序设计 | 提供计算机基础程序语言设计分析能力  |
| 2 | 数据库开发技术 | 提高计算机数据库应用能力 |
|  |  |  |
| 序号  | 后续课程名称  | 需要本课程支撑的主要能力  |
| 1 | 数据恢复 | 提供数据分析存储能力 |
| 2 | 软件项目综合开发 | 提供软件项目开发中数据分析应用能力 |

（三)本课程与中职、本科、培训班同类课程的区别。

|  |  |
| --- | --- |
| 层次 | 区别 |
| 本科 | 基础知识培养的基本要求不同，本科教育要求宽厚，宽广、扎实，本课程要求掌握基本就可以；专业知识培养的基本要求不同，本科要求了解就可以，本课程需要足够扎实；主要能力的培养要求不同，本科偏重研究能力、创新能力、创新创业能力，本课程偏重技能（操作能力、制作能力）。 |
| 中职 | 本课程理论知识比中职更加实用、有系统性，强调专业知识的来源并从中体验出基础知识在解决专业问题时的应用方法、途径和规律，强调专业知识的实用性、指导性。 |
| 培训班 | 培训班重点在操作应用能力方面的实践，本课程重点在提高技能（操作）与理论共同能力方面的实践。 |

三、课程目标

（一）总体目标：

该课程将使学生掌握数据分析管理的基本技术知识。学生在学习本课程的过程中将完成数据分析的建立与维护，数据分析表的建立与数据的录入与维护，在应用程序中访问数据分析，数据分析编程等工作任务。以通过本课程的学习，具备为各类应用程序提供数据分析数据存储的技能。

（二）素质目标：

1.在数据分析管理数据工作中，培养学生具有思维的严谨性；

2.在数据分析维护工作中，培养学生具有良好的文化修养；

3.在团体数据分析管理分工合作中，培养学生具有良好的团结协作精神、团队意识、组织协调能力；

4.在数据分析编程工作中，培养学生具有开拓创新精神、独立解决问题的能力；

（三）知识目标：

1.理解数据分析、数据分析系统、数据分析的体系结构结构及分类等基本概念。

2.掌握数据分析基本管理方法：表的操作、数据完整性以及表的索引和视图、数据分析查询和管理、数据分析备份与恢复。

3.了解SQL语言、了解SQL语言的组成、功能。

4.了解数据分析应用项目开发过程。

5.了解数据分析编程。

（四）能力目标：

1.具有根据系统需求分析绘制E-R图，并将E-R图转换为关系模型的能力；

2.具有对关系模型进行规范化能力；

3.具有创建数据分析和数据分析表的能力；

4.具有对数据分析表进行添加、修改和删除数据的能力；

5.具有对数据进行查询、统计汇总的能力；

6.具有对数据分析进行完整性维护的能力；

二、课程的主要内容及基本要求

第1章 数据分析概述

|  |  |  |  |
| --- | --- | --- | --- |
| 章名 | 数据分析概述 | 学时 | 12 |
| 学习目标 | 1. 了解数据分析的背景及应用场景
2. 掌握什么是数据分析以及数据分析的流程
3. 会创建Pyhton环境，使用Anaconda管理Python包
4. 会简单使用Jupyter Notebook
5. 认识常见的数据分析工具
 |
| 知识点 | 了解 | 熟悉 | 掌握 | 重点 |
| 数据分析的背景 | √ |  |  |  |
| 什么是数据分析 | √ |  |  |  |
| 数据分析的应用场景 | √ |  |  |  |
| 数据分析的流程 |  |  |  | √ |
| Python做数据分析的优势 | √ |  |  |  |
| Anconda的介绍 |  | √ |  |  |
| Anconda的安装及使用 |  |  |  | √ |
| 启用Jupyter Notebook |  |  |  | √ |
| Jupyter Notebook界面介绍及使用 |  |  |  | √ |
| 常见的数据分析工具 |  | √ |  |  |

第2章 科学计算库NumPy

|  |  |  |  |
| --- | --- | --- | --- |
| 章名 | 科学计算库NumPy | 学时 | 24 |
| 学习目标 | 1. 认识NumPy数组对象，会创建NumPy数组
2. 熟悉ndarray对象的数据类型，并会转换数据类型
3. 掌握数组运算方式
4. 掌握数组的索引和切片
5. 会使用数组进行数据处理
6. 熟悉线性代数模块和随机数模块的使用
 |
| 知识点 | 了解 | 熟悉 | 掌握 | 重点 |
| 认识NumPy数组对象 |  |  | √ |  |
| 创建NumPy数组 |  |  |  | √ |
| ndarray对象的数据类型 |  |  |  | √ |
| 数组运算 |  | √ |  |  |
| 整数索引和切片的基本使用 |  |  |  | √ |
| 花式（数组）索引的基本使用 |  |  |  | √ |
| 布尔型索引的基本使用 |  |  |  | √ |
| 数组的转置和轴对称 |  |  | √ |  |
| NumPy通用函数 |  | √ |  |  |
| 将条件逻辑转为数组运算 |  |  | √ |  |
| 数组统计运算 |  | √ |  |  |
| 数组排序 |  | √ |  |  |
| 检索数组元素 |  |  | √ |  |
| 线性代数模块 | √ |  |  |  |
| 随机数模块 |  |  | √ |  |
| 案例—酒鬼漫步 |  | √ |  |  |

第3章 数据分析工具Pandas

|  |  |  |  |
| --- | --- | --- | --- |
| 章名 | 数据分析工具Pandas | 学时 | 24 |
| 学习目标 | 1. 掌握Pandas的两种数据结构
2. 掌握Pandas索引的相关操作
3. 掌握Pandas的常见操作，包括算术运算、排序、统计计算
4. 掌握Pandas读写数据的方式
 |
| 知识点 | 了解 | 熟悉 | 掌握 | 重点 |
| Series |  |  |  | √ |
| DataFrame |  |  |  | √ |
| 索引对象 |  | √ |  |  |
| 重置索引 |  |  | √ |  |
| 索引操作 |  |  | √ |  |
| 算术运算与数据对齐 |  | √ |  |  |
| 数据排序 |  |  | √ |  |
| 统计计算与描述 | √ |  |  |  |
| 认识层次化索引 |  | √ |  |  |
| 层次化索引的操作 |  |  |  | √ |
| 读写文本文件 |  |  |  | √ |
| 读写Excel文件 |  |  |  | √ |
| 读写HTML表格数据 |  |  |  | √ |
| 读写数据分析 |  |  |  | √ |
| 案例—北京高考分数线统计分析 |  | √ |  |  |

第4章 数据预处理

|  |  |  |  |
| --- | --- | --- | --- |
| 章名 | 数据预处理 | 学时 | 12 |
| 学习目标 | 1. 掌握数据清洗常用操作，会检查和处理各类有问题的数据
2. 掌握数据合并的常用方法，会使用不同的方式合并数据
3. 掌握数据重塑的常见操作，会重塑Pandas对象的结构
4. 掌握数据转换的常见操作，可以实现离散化和哑变量处理
 |
| 知识点 | 了解 | 熟悉 | 掌握 | 重点 |
| 空值和缺失值的处理 |  |  | √ |  |
| 重复值的处理 |  |  | √ |  |
| 异常值的处理 |  |  | √ |  |
| 更改数据类型 |  |  | √ |  |
| 轴向堆叠数据 |  |  |  | √ |
| 主键合并数据 |  |  |  | √ |
| 根据行索引合并数据 |  |  |  | √ |
| 合并重塑数据 |  |  |  | √ |
| 重塑层次化索引 |  |  |  | √ |
| 轴向旋转 |  |  | √ |  |
| 重命名轴索引 |  |  | √ |  |
| 离散化连续数据 |  | √ |  |  |
| 哑变量处理类别型数据 | √ |  |  |  |
| 案例—预处理部分地区信息 |  | √ |  |  |

第5章 数据聚合与分组运算

|  |  |  |  |
| --- | --- | --- | --- |
| 章名 | 数据聚合与分组运算 | 学时 | 12 |
| 学习目标 | 1. 理解分组与聚合的原理
2. 掌握groupby()方法，可以按照不同的规则进行分组
3. 掌握聚合操作，会使用统计方法和聚合方法聚合数据
4. 掌握其他分组级运算方法的使用
 |
| 知识点 | 了解 | 熟悉 | 掌握 | 重点 |
| 分组与聚合的原理 | √ |  |  |  |
| 通过groupby()方法将数据拆分成组 |  |  |  | √ |
| 使用内置统计方法聚合数据 |  |  | √ |  |
| 面向列的聚合方法 |  |  |  | √ |
| 数据转换 |  |  |  | √ |
| 数据应用 |  |  |  | √ |
| 案例—运动员信息的分组与聚合 |  | √ |  |  |

第6章 数据可视化

|  |  |  |  |
| --- | --- | --- | --- |
| 章名 | 数据可视化 | 学时 | 12 |
| 学习目标 | 1. 了解什么是数据可视化
2. 掌握Matplotlib绘制图表的常用方法
3. 了解seaborn的基本使用
4. 了解bokeh的基本使用
 |
| 知识点 | 了解 | 熟悉 | 掌握 | 重点 |
| 什么是数据可视化 | √ |  |  |  |
| 常见的图表类型 |  | √ |  |  |
| 数据可视化的工具 | √ |  |  |  |
| 通过figure()函数创建画布 |  | √ |  |  |
| 通过subplot()函数创建单个子图 |  |  | √ |  |
| 通过subplots()函数创建多个子图 |  |  | √ |  |
| 通过add\_subplot()方法添加和选中子图 |  |  | √ |  |
| 添加各类标签 |  |  | √ |  |
| 绘制常见图表 |  |  | √ |  |
| 本地保存图形 |  |  | √ |  |
| 可视化数据的分布 | √ |  |  |  |
| 用分类数据绘图 |  | √ |  |  |
| 认识Bokeh库 | √ |  |  |  |
| 通过Plotting绘制图形 |  | √ |  |  |
| 案例—画图分析某年旅游景点数据 |  | √ |  |  |

第7章 时间序列数据分析

|  |  |  |  |
| --- | --- | --- | --- |
| 章名 | 时间序列数据分析 | 学时 | 12 |
| 学习目标 | 1. 了解什么是时间序列，会创建时间序列对象
2. 会使用时间戳索引和切片选取子集
3. 学会创建固定频率的时间序列，能够调整时间序列的频率
4. 掌握Period（时期），转换时期的频率
5. 掌握重采样，实现降采样和升采样
6. 熟悉滑动窗口的使用
 |
| 知识点 | 了解 | 熟悉 | 掌握 | 重点 |
| 创建时间序列 |  |  | √ |  |
| 通过时间戳索引选取子集 |  | √ |  |  |
| 创建固定频率的时间序列 |  |  | √ |  |
| 时间序列的频率、偏移量 |  |  | √ |  |
| 时间序列的移动 |  |  | √ |  |
| 创建时期对象 |  |  | √ |  |
| 时期的频率转换 |  |  | √ |  |
| 重采样方法（resample） |  |  | √ |  |
| 降采样 |  |  | √ |  |
| 升采样 |  |  | √ |  |
| 数据统计—滑动窗口 |  |  | √ |  |
| 时序模型—ARIMA | √ |  |  |  |
| 案例—股票收盘价分析 |  | √ |  |  |

第8章 文本数据分析

|  |  |  |  |
| --- | --- | --- | --- |
| 章名 | 文本数据分析 | 学时 | 4 |
| 学习目标 | 1. 了解文本分析的工具NLTK与jieba，会安装和使用这些工具
2. 掌握文本预处理的流程
3. 掌握文本情感分析，可以用NLTK分析情感倾向
4. 掌握文本相似度，可以结合NLTK与余弦相似度实现相似度分析
5. 掌握文本分类，可以结合NLTK与算法对文本进行分类
 |
| 知识点 | 了解 | 熟悉 | 掌握 | 重点 |
| NLTK与jieba概述 | √ |  |  |  |
| 安装NLTK和下载语料库 |  | √ |  |  |
| jieba库的安装 |  | √ |  |  |
| 预处理的流程 |  |  | √ |  |
| 分词 |  |  | √ |  |
| 词性标注 |  |  | √ |  |
| 词形归一化 |  |  | √ |  |
| 删除停用词 |  |  | √ |  |
| 文本情感分析 |  | √ |  |  |
| 文本相似度 |  | √ |  |  |
| 文本分类 |  | √ |  |  |
| 案例—商品评价分析 |  | √ |  |  |

第9章 实战—北京租房数据统计分析

|  |  |  |  |
| --- | --- | --- | --- |
| 章名 | 实战—北京租房数据统计分析 | 学时 | 2 |
| 学习目标 | 1. 掌握数据预处理的方法
2. 掌握使用Matplotlib绘制图像的方法
3. 了解百度地图API的使用
4. 会基于数据进行独立分析
 |
| 知识点 | 了解 | 熟悉 | 掌握 | 重点 |
| 数据读取 |  |  | √ |  |
| 重复值和空值处理 |  |  | √ |  |
| 数据转换类型 |  |  | √ |  |
| 房源数量、位置分布分析 |  |  | √ |  |
| 户型数量分析 |  |  | √ |  |
| 平均租金分析 |  |  | √ |  |
| 面积区间分析 |  |  | √ |  |

三、学时分配

|  |  |  |  |
| --- | --- | --- | --- |
| 章目 | 讲课 | 上机 | 合计 |
| 第1章 数据分析概述 | 6学时 | 6学时 | 12学时 |
| 第2章 科学计算库NumPy | 12学时 | 12学时 | 24学时 |
| 第3章 数据分析工具Pandas | 12学时 | 12学时 | 24学时 |
| 第4章 数据预处理 | 6学时 | 6学时 | 12学时 |
| 第5章 数据聚合与分组运算 | 6学时 | 6学时 | 12学时 |
| 第6章 数据可视化 | 6学时 | 6学时 | 12学时 |
| 第7章 时间序列数据分析 | 6学时 | 6学时 | 12学时 |
| 第8章 文本数据分析 | 2学时 | 2学时 | 4学时 |
| 第9章 实战—北京租房数据统计分析 | 1学时 | 1学时 | 2学时 |
| 合计 | 57学时 | 57学时 | 114学时 |

四、重点难点

## 第1章 数据分析概述

|  |  |
| --- | --- |
| 知识点 | 重要程度 |
| 数据分析的背景 | 了解 |
| 什么是数据分析 | 了解 |
| 数据分析的应用场景 | 了解 |
| 数据分析的流程 | 重点 |
| Python做数据分析的优势 | 了解 |
| Anconda的介绍 | 熟悉 |
| Anconda的安装及使用 | 重点 |
| 启用Jupyter Notebook | 重点 |
| Jupyter Notebook界面介绍及使用 | 重点 |
| 常见的数据分析工具 | 熟悉 |

## 第2章 科学计算库NumPy

|  |  |
| --- | --- |
| 知识点 | 重要程度 |
| 认识NumPy数组对象 | 掌握 |
| 创建NumPy数组 | 重点 |
| ndarray对象的数据类型 | 重点 |
| 数组运算 | 熟悉 |
| 整数索引和切片的基本使用 | 重点 |
| 花式（数组）索引的基本使用 | 重点 |
| 布尔型索引的基本使用 | 重点 |
| 数组的转置和轴对称 | 掌握 |
| NumPy通用函数 | 熟悉 |
| 将条件逻辑转为数组运算 | 掌握 |
| 数组统计运算 | 熟悉 |
| 数组排序 | 熟悉 |
| 检索数组元素 | 掌握 |
| 唯一化及其它集合逻辑 | 掌握 |
|  线性代数模块 | 了解 |
| 随机数模块 | 掌握 |

## 第3章 数据分析工具Pandas

|  |  |
| --- | --- |
| 知识点 | 重要程度 |
| Series | 重点 |
| DataFrame | 重点 |
| 索引对象 | 熟悉 |
| 重置索引 | 掌握 |
| 索引操作 | 掌握 |
| 算术运算与数据对齐 | 熟悉 |
| 数据排序 | 掌握 |
| 统计计算与描述 | 了解 |
| 认识层次化索引 | 熟悉 |
| 层次化索引的操作 | 重点 |
| 读写文本文件 | 重点 |
| 读写Excel文件 | 重点 |
| 读写HTML表格数据 | 重点 |
| 读写数据库 | 重点 |

## 第4章 数据预处理

|  |  |
| --- | --- |
| 知识点 | 重要程度 |
| 空值和缺失值的处理 | 掌握 |
| 重复值的处理 | 掌握 |
| 异常值的处理 | 掌握 |
| 更改数据类型 | 掌握 |
| 轴向堆叠数据 | 重点 |
| 主键合并数据 | 重点 |
| 根据行索引合并数据 | 重点 |
| 合并重叠数据 | 重点 |
| 重塑层次化索引 | 重点 |
| 轴向旋转 | 掌握 |
| 重命名轴索引 | 掌握 |
| 离散化连续数据 | 熟悉 |
| 哑变量处理类别型数据 | 了解 |

## 第5章 数据聚合与分组运算

|  |  |
| --- | --- |
| 知识点 | 重要程度 |
| 分组与聚合的原理 | 了解 |
| 通过groupby()方法将数据拆分成组 | 重点 |
| 使用内置统计方法聚合数据 | 掌握 |
| 面向列的聚合方法 | 重点 |
| 数据转换 | 重点 |
| 数据应用 | 重点 |

## 第6章 数据可视化

|  |  |
| --- | --- |
| 知识点 | 重要程度 |
| 什么是数据可视化 | 了解 |
| 常见的图表类型 | 熟悉 |
| 数据可视化的工具 | 了解 |
| 通过figure()函数创建画布 | 熟悉 |
| 通过subplot()函数创建单个子图 | 掌握 |
| 通过subplots()函数创建多个子图 | 掌握 |
| 通过add\_subplot()方法添加和选中子图 | 掌握 |
| 添加各类标签 | 掌握 |
| 绘制常见图表 | 掌握 |
| 本地保存图形 | 掌握 |
| 可视化数据的分布 | 了解 |
| 用分类数据绘图 | 熟悉 |
| 认识Bokeh库 | 了解 |
| 通过Plotting绘制图形 | 熟悉 |

## 第7章 时间序列数据分析

|  |  |
| --- | --- |
| 知识点 | 重要程度 |
| 创建时间序列 | 掌握 |
| 通过时间戳索引选取子集 | 熟悉 |
| 创建固定频率的时间序列 | 重点 |
| 时间序列的频率、偏移量 | 重点 |
| 时间序列的移动 | 掌握 |
| 创建时期对象 | 掌握 |
| 时期的频率转换 | 掌握 |
| 重采样方法（resample） | 掌握 |
| 降采样 | 掌握 |
| 升采样 | 掌握 |
| 数据统计—滑动窗口 | 掌握 |
| 时序模型—ARIMA | 了解 |

## 第8章 文本数据分析

|  |  |
| --- | --- |
| 知识点 | 重要程度 |
| NLTK与jieba概述 | 了解 |
| 安装NLTK和下载语料库 | 掌握 |
| jieba库的安装 | 掌握 |
| 预处理的流程 | 重点 |
| 分词 | 重点 |
| 词性标注 | 重点 |
| 词形归一化 | 重点 |
| 删除停用词 | 重点 |
| 文本情感分析 | 熟悉 |
| 文本相似度 | 熟悉 |
| 文本分类 | 熟悉 |

## 第9章 实战—北京租房数据统计分析

|  |  |
| --- | --- |
| 知识点 | 重要程度 |
| 数据读取 | 掌握 |
| 重复值和空值处理 | 掌握 |
| 数据转换类型 | 掌握 |
| 房源数量、位置分布分析 | 掌握 |
| 户型数量分析 | 掌握 |
| 平均租金分析 | 掌握 |
| 面积区间分析 | 掌握 |