《大数据分析计算机基础》教学大纲

课程编号：
课程类型：专业必修课
总 学 时：48　　**讲课学时：**　　**实验（上机）学时：**
学　　分：3
适用对象：应用统计专业硕士
先修课程：计算机基础、数据库基础

一、课程的教学目标

大数据分析计算机基础是大数据分析应用统计专业硕士学生的专业必修课，通过本课程的学习使学生能够掌握大数据分析领域所需要具备的数据处理语言以及在此基础上的数据获取、数据存储、数据管理、数据分析、数据可视化、数据智能以及数据应用等。

二、教学基本要求

本课程的教学内容包括六个部分：数据处理语言、数据获取（网络爬虫及其相关技术）、数据存储管理与分析（数据库和数据仓库及其相关技术），数据处理、数据可视化以及数据智能等。

每一部分知识结构的设计面向当前主流的大数据相关技术，有针对性地向学生介绍大数据分析所应具备的理念、知识和技能。在授课过程中力求重点突出，教授与启发相结合，既传授基础知识和技能，又培养学生的动手能力和动脑习惯，培养学生利用信息技术独立解决问题的能力。授课方式采用 PPT 课堂授课，操作演示，案例设计，任务驱动相结合的模式，在课堂教学的同时设计实践性任务，通过作业或大作业检验学生知识技能掌握装填。

学生作业和大作业作为阶段性测试，计入期末总成绩。

大作业是综合检验学生学生学习状态的重要措施，可单独或组队完成，可自定题目或选择备选题目。

三、各教学环节学时分配

教学课时分配

序号	章节内容	讲课	作业
1	课程介绍以及大数据主题报告	3 课时	
2	数据处理语言：Python 快速入门	3 课时	作业 1：基础算法作业
3	数据处理语言：列表、字符串及文件	3 课时	
4	数据处理语言：字典、集合和其他	3 课时	作业 2：应用性作业
5	数据处理语言：函数及扩展库	3 课时	
6	数据处理语言：异常、面向对象	3 课时	作业 3：综合作:1
7	数据获取：网络爬虫	3 课时	作业 4：数据获取
8	数据获取：内容解析	3 课时	
9	数据存储管理：数据库基础 1	3 课时	发布大作业选题
10	数据存储管理：数据库基础 2	3 课时	
11	数据存储管理：数据仓库与数据分析	3 课时	作业 5：综合作业 2
12	数据处理：Python 相关扩展库	3 课时	
13	数据可视化	3 课时	
14	数据智能：技术框架	3 课时	
15	数据只能：应用实践	3 课时	
16	大作业展示	3 课时	大作业结题
合计		48 课时	

【注：教学课时可能根据教学情况适当调整】

四、教学内容

导引部分

课程介绍，包括：课程内容组成、课程考核以及计分等。

主题报告：数据集大数据的定义、大数据特征、大数据分析技术、人工智能与大数据、案例化大数据应用。

由授课教师为学生作一个关于大数据及大数据分析技术的主题报告，介绍大数据的产生过程，涉及的关键技术，应用需求，主流解决技术方案，关键技术及未来的技术发展趋势。

第一部分 Python 程序基础

第一章 Python 快速入门

以对比法了解 Python 语言特征，掌握 Python 安装以初步程序设计。

1. 从 Hello, world!开始
2. 认识 print()函数
3. 认识数据类型
4. 认识 input()函数
5. 标识符、变量、字面量
6. 认识运算符、表达式
7. 流程控制语句：分支
8. 流程控制语句：循环

重点、难点：循环以及从问题到求解。

第二章 列表、字符串及文件

理解顺序容器（sequential container）的基本概念，即顺序稳定的容器，包括 Python 中的 List、Tuple、String，文件是存储在外存储器中的内容，可以视为顺序容器。掌握顺序容器的操作，包括：下标、切片、循环、字符串函数、

enumerate、文本文件打开、读写等。

1. List 的定义及其使用
2. Tuple 的定义及其使用
3. String 的定义以及相关函数
4. 下标与切片
5. 循环以及 enumerate()函数的使用
6. 文本文件的读写

重点、难点：熟练使用切片以及文件操作。文件是数据存储的重要方式，可以基于此完成各种真实数据处理。

第三章 字典、集合及其他

熟练掌握字典和集合，字典是最为常用的数据结构。

1. 字典的定义及其使用
2. 集合的定义、运算及其使用
3. 数据类型转换
4. 构造复杂数据结构

重点、难点：数据类型转换以及复杂数据结构的构造

第四章 内置函数、自定义函数和扩展库

函数是程序设计的重要理念和概念，有助于任务分解复杂度控制等，对软件工程极为重要。本章包括内置函数的使用、自定义函数以及自定义扩展库和常用扩展库的使用；

1. 函数综述
2. Python 内置函数（built-in functions）
3. 自定义函数（和其他语言相比，Python 的函数很有特色）
4. 常用 Python 标准库
5. 扩展库的意义
6. 自定义扩展库

7. 第三方库（包括：自然语言分词库 jieba 等）

重点、难点：自定义函数、自定义扩展库以及第三方库

第五章 异常处理和面向对象程序设计

异常处理是程序设计的重要组成部分，本部分要求掌握异常处理在 Python 程序设计中的应用。

Python 是面向对象的程序设计语言，掌握面向对象概念及其在 Python 程序设计中的实践。

1. 异常处理的基本概念
2. Python 中的异常处理
3. 面向对象程序设计的概念
4. Python 中的面向对象程序设计

重点、难点：异常处理和面向对象程序设计

第二部分 数据获取

第六章 数据获取：网络爬虫

数据获取对大数据很重要，涉及多种技术。网络爬虫是获取数据的重要技术。本章将讲解多种形式的网络爬虫技术。

1. 搜索引擎与网络爬虫的基本概念
2. BS 架构与网络爬虫
3. 登录后数据爬取
4. Selenium 与数据获取
5. 反爬虫技术及其对抗
6. 了解 JavaScript 和 AJAX；

重点难点：BS 架构、爬虫与反爬虫技术、JavaScript 与 AJAX 技术。

第七章 数据获取：内容解析

1. 认识正则表达式
2. 正则表达式与网页解析
3. BeautifulSoap 与网页解析

重点难点：深入了解网页结构，解析网页动态数据。

第三部分 数据存储与管理

外存能持久存储更多数据，管理外存中的大量数据一直是计算技术的热点和重点，其中多种数据库技术是其重要组成部分。

第八章 数据库基础知识

1. 认识外存数据及其相关技术和挑战
2. 认识数据库，重点关系数据库和 Key-Value 数据库
3. 关系数据库的模型
4. 关系数据库系统结构
5. SQL 的基本概念和语法
6. 数据定义 SQL
7. 数据查询 SQL
8. 数据更新 SQL
9. 视图的定义和使用
10. 触发器的定义和使用
11. 在 Python 程序中使用关系数据库

重点、难点：关系模型的概念和理解，SQL 命令的使用，连接 SQL 命令的使用

第九章 数据库分析查询命令实践

1. SQL 中数据处理函数的使用
2. 统计函数
3. 分组聚集 SQL 命令的使用

4. 初识 Key-value 数据库 MongoDB
5. MongoDB 与关系型数据的比较
6. 在 Python 中使用 MongoDB。

重点、难点：聚集函数的使用，分组聚集操作，应用于分析的函数使用

第十章 数据仓库基础知识

1. 数据仓库基本概念
2. 数据仓库体系结构
3. 数据仓库中的数据及组织
4. 数据仓库与关系型数据库的比较
5. 在线分析技术概要
6. 多维数据模型
7. 多维分析操作
8. 数据仓库与在线分析技术事件

重点、难点：数据仓库的概念和理解，数据仓库与数据库的区别，构建数据仓库的技术路线，OLAP（在线分析技术）的基本概念及理解，维、层次的概念及理解，多维分析操作。

第四部分 数据处理

Python 提供了大量优秀的数据处理扩展库，包括但不限于：Pandas、NumPy、StatsModels 等。掌握这些工具，可以很好地助力大数据处理。

第十一章 数据处理：Python 的扩展库

1. Python 的数据处理扩展库概述
2. Python 结合 Excel 的数据处理
3. 相关 Python 扩展库（Pandas、NumPy、StatsModels 等）

重点、难点：数据处理工具较多，更多了解有助于提升数据处理效率。

第五部分 数据可视化

数据可视化是以图形化（或视觉化）技术，以更加清晰有效地传达与沟通信息，是技术与艺术的结合。

第十二章　数据可视化

Python 有较多的数据可视化组件，以 Matplotlib 较为常用。另外 PyEcharts 也较为不错。

1. 时空数据可视化
2. 动态数据可视化
3. 相关扩展库的使用

教学重点、难点：时空数据可视化和动态可视化。

第六部分 数据智能

数据是本期人工智能的基础，而人工智能算法又大大地促进了数据科学的发展，二者相辅相成。通过大规模机器学习和深度学习等技术，对海量数据进行处理、分析和挖掘，能提取数据中所包含的有价值的信息和知识，从而使数据具有“智能”，并在此基础上寻求现有问题的求解预测等。

第十三章　数据智能：技术框架

1. 人工智能发展历程
2. 大数据与人工智能的关系
3. Python 中人工智能和大数据扩展库

重点、难点：理解各种算法的典型应用场景。

第十四章　数据智能：应用实践

1. 框架在大数据中的意义
2. 国外流行框架
3. 国产流行框架

4. 人工智能和大数据技术的应用示例

重点、难点：将大型人工智能库应用到具体场景。

五、主要参考书

1. 唐大仕 著，《Python 程序设计》，北京，电子工业出版社
2. 陈允杰 著，《Python 网络爬虫与数据可视化应用实战》，北京，中国水利水电出版社
3. 沈兆阳 编著，《SQL Server 2000 OLAP 解决方案数据仓库与 Analysis Services》，北京，清华大学出版社
4. 王珊，萨师煊著，《数据库系统概论（第四版）》，北京：高等教育出版社，2007
5. 王珊，李翠平，李盛恩等编著，《数据仓库与数据分析教程》，北京：高等教育出版社，2012