

《数据科学导论》课程实验教学大纲

课 程 名 称: 数据科学导论

英 文 名 称: Introduction to Data Science

课 程 编 号: sd03031390-0

实 验 课 性 质: 非独立设课

课 程 负 责 人: 崔立真

开放实验题目数:

大 纲 主 撰 人: 崔立真

大 纲 审 核 人:

一、学时、学分

课程总学时: 32

实验学时:

课程总学分: 2

实验学分:

二、适用专业及年级

软件工程 本科 3 年级

三、实验教学目的与基本要求

本课程面向计算机、软件工程的高年级学生, 目的是介绍大数据基本概念, 数据驱动的思维方法、工具, 通过一些典型算法和工具的使用, 训练学生综合运用数据科学的数据探索思维、数据分析和处理的方法和工具, 来解决以数据驱动为核心的实际应用问题。

四、主要仪器设备

Python 语言等数据分析开发和应用工具

五、实验课程内容和学时分配

序号	实验题目名称	实 验 内 容	学时分配	实验属性	实验类型	每组人数	实验要求	指导教师	已开/未开
1	数据科学家项目	能够按照数据分析的流程, 从大量数据中分析用户行为, 为用户进行分类, 最后输出洞察报告。		专业类	综合性	5	必做	崔立真	已开
2	文本分析与实体识别	利用文本分析技术进行关于两个商品数据集 (Amazon 和 Google 的众多商品的数据记录文件与) 的实体识别。		专业类	设计性	1	必做	崔立真	已开

3	安装 Hadoop	掌握 Linux 虚拟机安装方法或者双操作系统安装方法； 掌握 Hadoop 的伪分布式安装方法。		专业类	设计性	1	必做	崔立真	已开
4	HDFS-Java-API 编程实践	理解 HBase 在 Hadoop 体系结构中的角色； 熟练使用 HBase 操作常用的 Shell 命令； 熟悉 HBase 操作常用的 Java API。		专业类	设计性	1	选做	崔立真	已开
5	编写 MapReduce 程序实现自然连接	掌握基本的 MapReduce 编程方法； 理解 Hadoop 中 MapReduce 模块的处理逻辑； 掌握用 MapReduce 解决一些常见的数据处理问题，包括词频统计、数据排序、自然连接等。		专业类	设计性	1	必做	崔立真	已开
6	市场营销数据科学实践	从市场营销数据科学练习包中选择 2 个问题的数据集进行试验。		专业类	设计性	1	必做	崔立真	已开

7	Tabular data processing with Pandas	能够使用存储在表格形式中的数据并进行数据分析；能够使用python 数据分析库Pandas。		专业类	设计性	1	选做	崔立真	已开
8	Natural Language Parsing	使用从web service 中获取半结构化数据，浏览数据并解析自由文本内容。		专业类	设计性	1	选做	崔立真	已开
9	图数据建模和可视化	使用 Gephi 工具，基于给定的 Twitter 数据集，进行群体网络结构和群体演化可视分析。		专业类	设计性	1	必做	崔立真	已开

六、考核方式

实验报告+上机程序检查

七、实验教科书、参考书

实验指导+ipython notebook