大数据综合处理实验
课 程 简 介

鸣谢：本课程得到Google与Intel公司
中国大学合作部精品课程计划资助

南京大学计算机科学与技术系

教师：黄宜华，顾荣

课程简介

教学内容简介

本课程将系统介绍大数据并行处理技术和编程方法。课程首先介绍并行计算技术的基本概念、原理、方法和技术，在此基础上，介绍基于集群的大数据并行处理技术原理和方法，重点介绍Hadoop MapReduce并行计算集群的构架、用于大数据存储和计算的分布式文件系统、以及基于MapReduce集群的大数据并行处理技术和编程方法，MapReduce并行化算法设计技术、并行化算法应用研究案例。

教学目标

课程的主要目标是通过介绍多处理器并行处理技术、以及基于集群的大数据并行处理技术和MapReduce并行编程模型和方法，要求学生理解和掌握并行处理技术的基本概念、原理和构架、以及基于集群的大规模海量数据并行处理与编程技术方法，并能够用MapReduce实际设计和编写具体的大数据处理应用问题的算法和程序。

选课要求

具有Java程序设计能力、Linux系统操作使用能力、以及基础性机器学习与数据挖掘算法知识；除课堂听课外需要完成编程实验，还要求在学期结束时完成一个综合性课程设计

选修本课程的重要性

并行处理成为计算技术的重大发展趋势

- 单处理器性能提升达到极限，多核/多处理并行计算成为计算技术发展必然趋势
- 并行计算技术将渗透到每个计算应用领域
- 并行计算技术将影响传统计算技术的各个层面，与传统计算技术相互结合产生很多新的研究热点和课题

IT行业和应用已进入“大数据(Big Data)”和“数据为王”的时代

IT行业应用规模急剧扩大，出现越来越多的超大规模数据处理应用需求，传统系统难以提供足够的存储和计算资源进行处理

- 2008年国际著名的《Nature》杂志出版一期专刊专门讨论未来大数据(Big Data)处理相关的技术问题和挑战
- 世界权威的IT数据分析公司IDC：全世界的数据量2009年为800EB，到2020年将增长44倍，达到35ZB(35,000EB)
- 未来的IT行业中，价值在于数据，未来是“数据为王”的时代

选修本课程的重要性

计算机专业人员面临挑战，市场迫切需要相应的专业技术人才

- 并行计算技术将从硬件到软件全面影响传统计算技术的各个层面，新的技术发展挑战和需求迫使我们软件开发和程序设计人员必须尽快掌握并行计算技术
- 20-30年前程序设计技术最大的革命是面向对象技术，而下一个程序设计技术的革命将是并行程序设计技术
- 今天绝大多数程序员不懂并行设计技术，就像20年前绝大多数程序员不懂面向对象技术一样
- 目前国内外的知名IT企业迫切需要大量掌握大规模数据并行处理技术的人才

课程内容

Ch.1 并行计算技术简介

简要介绍并行计算技术的概况，基本分类，主要技术问题，MPI并行程序设计，大规模并行数据处理技术

Ch.2 MapReduce简介

简要介绍MapReduce技术的由来，基本构思，编程模型，主要设计思想和技术特征，基本应用

Ch.3 Google 和Hadoop MapReduce的基本构架

介绍Google MapReduce并行计算框架的基本结构、工作原理，Google分布式文件系统GFS的基本构架与工作原理，Google结构化数据管理系统BigTable的基本结构与工作原理

介绍开源大数据处理系统Hadoop 的基本组成结构和工作原理，MapReduce基本框架和工作原理，HDFS基本组成及工作原理，并介绍HDFS的基本编程

课程内容

Ch.4 Hadoop系统安装运行与程序开发

介绍单机和集群Hadoop系统安装方法和步骤，以及程序开发环境与开发过程

实验1：Hadoop系统安装与WordCount词频统计编程实验

Ch.5 MapReduce算法设计

介绍排序算法、文档倒排索引、文档共现算法、专利文献数据分析应用

实验2：搜索引擎文档倒排索引编程实验

Ch.6 Hadoop HBase与Hive原理与编程技术

介绍Hadoop 分布式数据管理系统HBase工作原理及其编程技术；介绍Hadoop数据仓库 Hive基本结构、工作原理及其编程技术

实验3：Hadoop HBase和Hive编程实验 （待定）

课程内容

Ch.7 高级MapReduce编程技术

介绍复杂I/O数据表示、用复合键值对完成特殊处理、程序员定制的I/O格式、Partitioner、Combiner，基于迭代的MapReduce求解方法、数据相关MapReduce任务计算、链式MapReduce计算、多数据源连接、访问关系数据库等高级技术

Ch.8 基于MapReduce的搜索引擎算法

介绍网页排名算法PageRank，搜索引擎文档倒排索引算法，以及全文检索系统的设计实现

实验4： Wikipedia网页PageRank实验（待定）

Ch.9 基于MapReduce的数据挖掘基础算法

介绍机器学习和数据挖掘中的聚类算法、分类算法、频繁项集挖掘等算法的MapReduce并行化设计技术方法

实验5：待定

课程内容

Ch.10 Spark系统和编程技术介绍

介绍基于内存计算的Spark系统及其基本编程技术

课程设计大作业

自选或由老师指定具有一定难度和工作量的题目，完成一个综合性大数据课程设计

课时安排

2020年春季学期

学期中： 课堂讲授，课程实验，复习，期末考试

每周2课时，共计18次（32课时）

暑假7月份：分组完成综合课程设计

考核方法

期末考试

笔试，占50%

课程实验

5次，共计占25%

课程设计

占25%

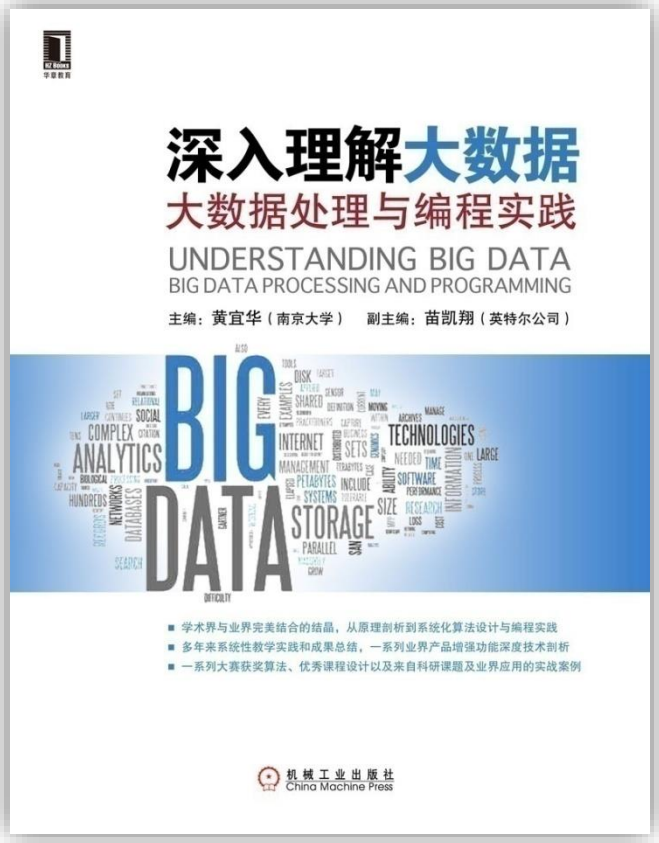
课件与参考书目、文献

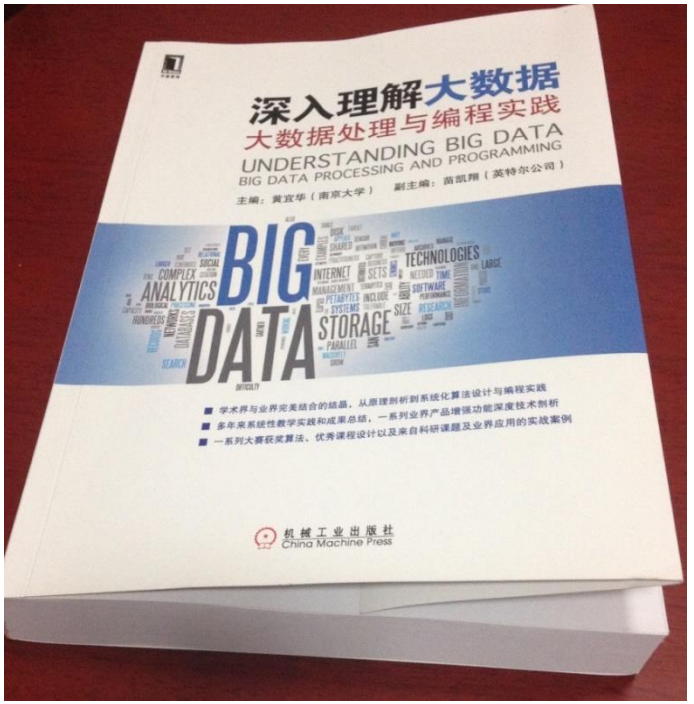
课件参考资料下载：计算机系本科教学服务平台

教材（安装配置，编程API部分版本较老，最新的参考相关软件官网）：

《深入理解大数据—大数据编程技术与实践》

黄宜华，苗凯翔主编，机械工业出版社，2014

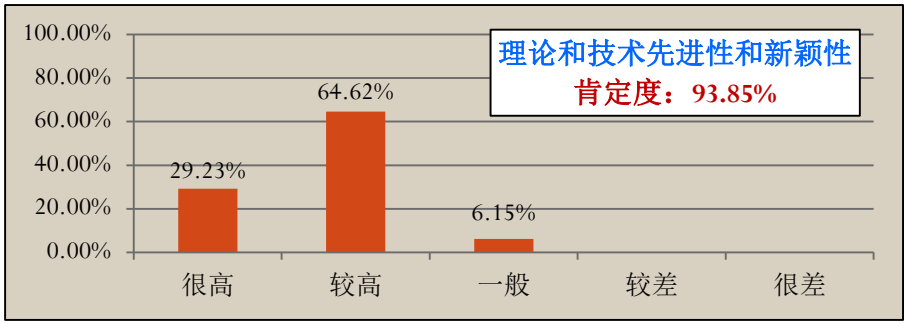


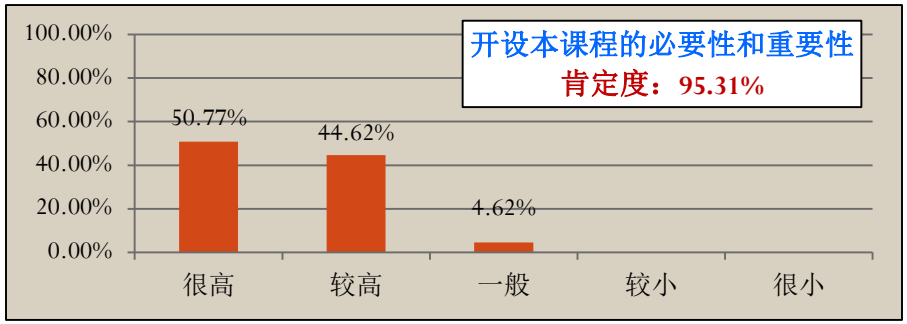


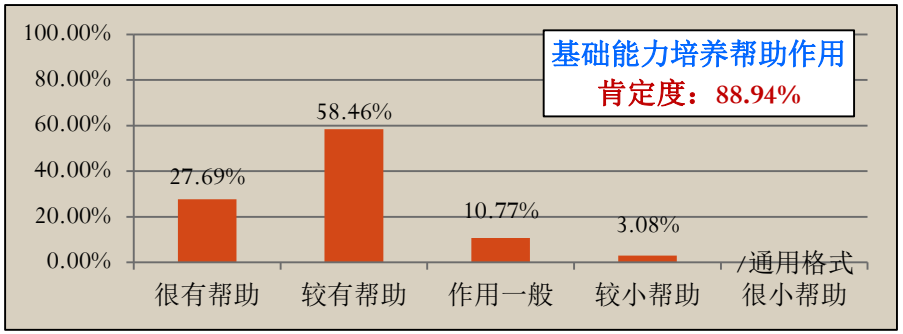
课程开课情况

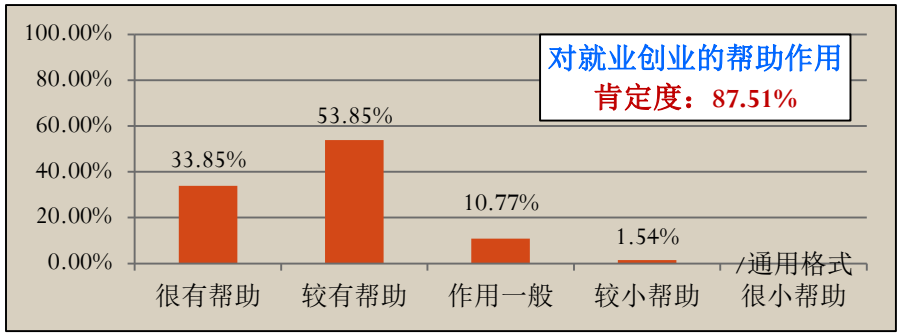
- 2011-2018学年已经为研究生开设了8个学期，每周2学时，研究生和本科生，本系选修人数900多人（研究生为主）
- 2016年首次为本科生开设大数据综合实验课程，120人选修，课程情况良好，绝大多数同学一次通过，仅有6位同学因笔试成绩不好或因课程设计综合实验未做，后补考通过
- 安排5次又简到难的课程实验，从基本工具平台使用到编程实验
- 课程结束后，要求分小组完成一个具有一定难度的课程项目设计，每年都会出现一批相当出色的课程设计项目。

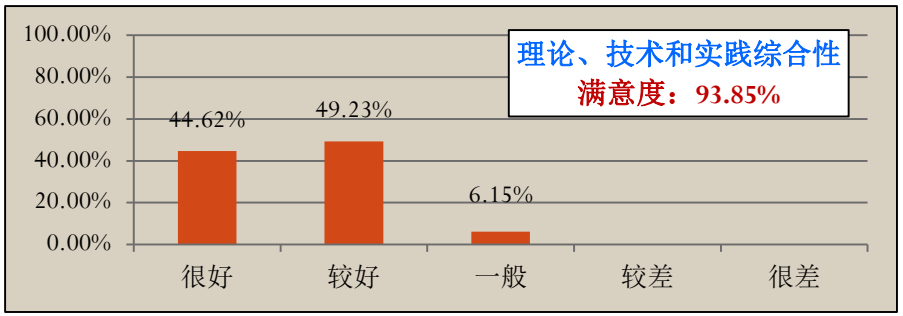
课程教学评估

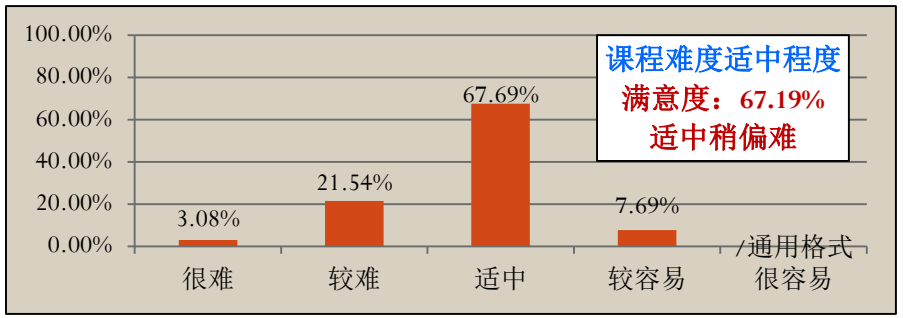

理论和技术先进性和新颖性
肯定度：93.85%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
29.23%
64.62%
6.15%
很高
较高
一般
较差
很差


开设本课程的必要性和重要性
肯定度：95.31%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
50.77%
44.62%
4.62%
很高
较高
一般
较小
很小

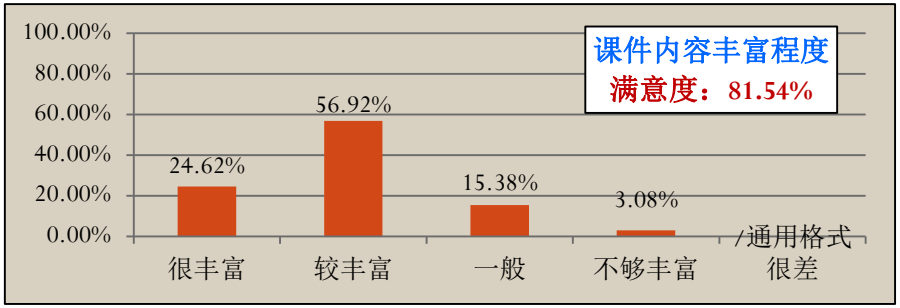

基础能力培养帮助作用
肯定度：88.94%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
27.69%
58.46%
10.77%
3.08%
很有帮助
较有帮助
作用一般
较小帮助
很小帮助
/通用格式

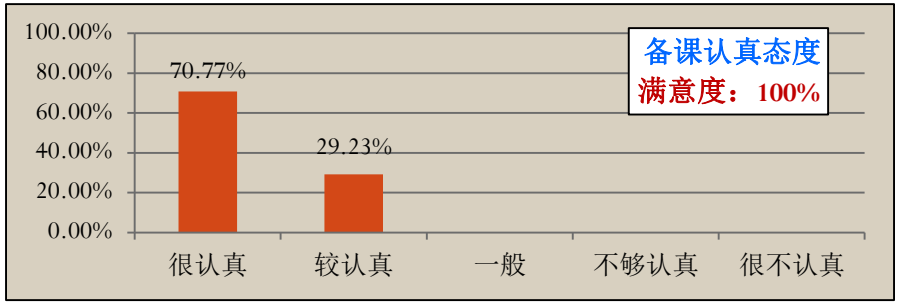

对就业创业的帮助作用
肯定度：87.51%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
33.85%
53.85%
10.77%
1.54%
很有帮助
较有帮助
作用一般
较小帮助
很小帮助
/通用格式

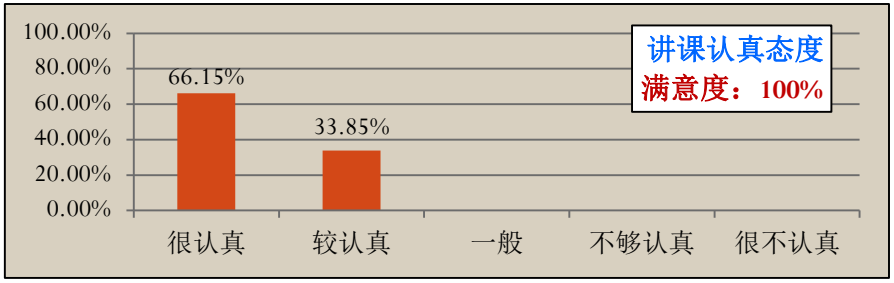

理论、技术和实践综合性
满意度：93.85%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
44.62%
49.23%
6.15%
很好
较好
一般
较差
很差

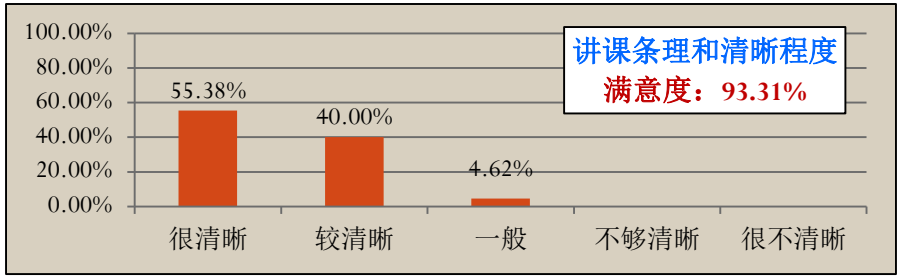

课程难度适中程度
满意度：67.19%
适中稍偏难
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
3.08%
21.54%
67.69%
7.69%
很难
较难
适中
较容易
很容易
/通用格式

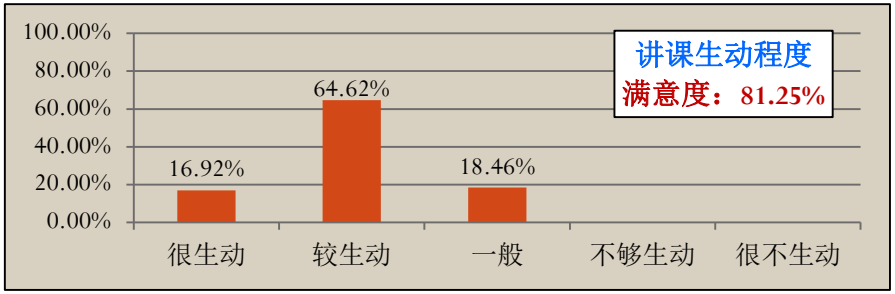
课程教学评估

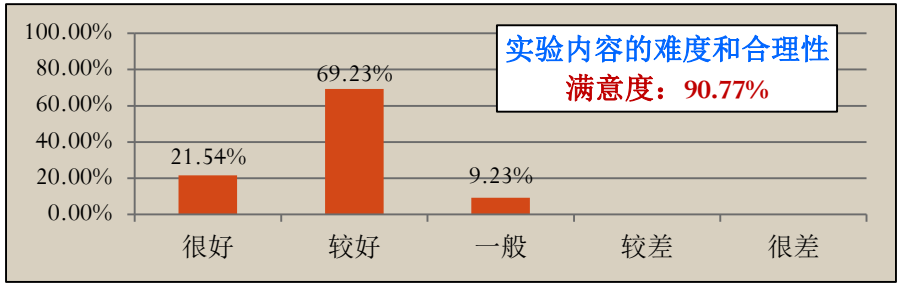

课件内容丰富程度
满意度：81.54%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
24.62%
56.92%
15.38%
3.08%
/通用格式
很丰富
较丰富
一般
不够丰富
很差

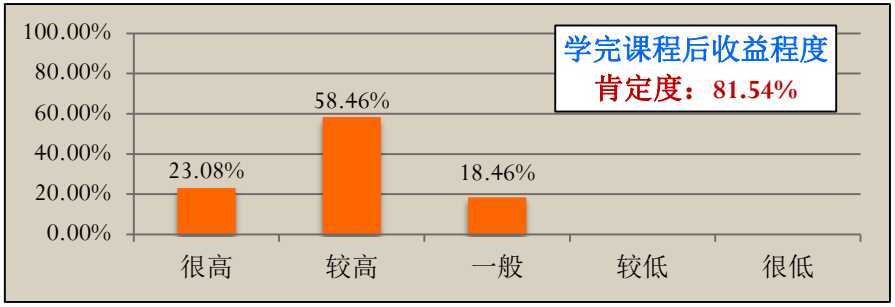

备课认真态度
满意度：100%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
70.77%
29.23%
很认真
较认真
一般
不够认真
很不认真

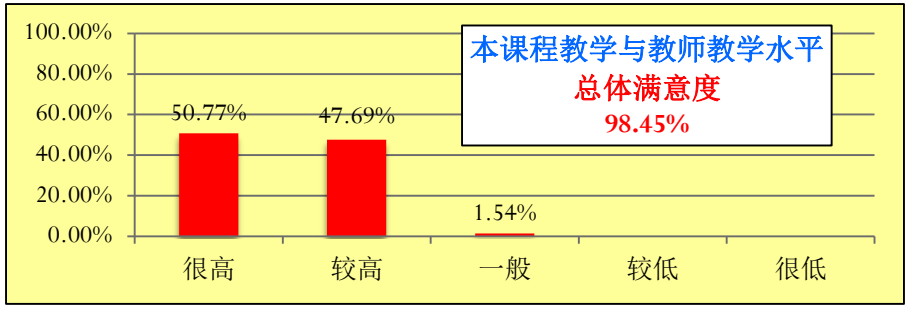

讲课认真态度
满意度：100%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
66.15%
33.85%
很认真
较认真
一般
不够认真
很不认真


讲课条理和清晰程度
满意度：93.31%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
55.38%
40.00%
4.62%
很清晰
较清晰
一般
不够清晰
很不清晰


讲课生动程度
满意度：81.25%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
16.92%
64.62%
18.46%
很生动
较生动
一般
不够生动
很不生动


实验内容的难度和合理性
满意度：90.77%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
21.54%
69.23%
9.23%
很好
较好
一般
较差
很差


学完课程后收益程度
肯定度：81.54%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
23.08%
58.46%
18.46%
很高
较高
一般
较低
很低


本课程教学与教师教学水平
总体满意度
98.45%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
50.77%
47.69%
1.54%
很高
较高
一般
较低
很低

教学效果

第一届“中国云/移动互联网创新大奖赛”

本课程开设后，我系机器学习与数据挖掘研究所和云计算与大数据并行计算课题组学习了MapReduce技术的同学组织了4支研究生代表队在“中国云产业联盟”组织的首届“中国云·移动互联网创新大奖赛”中参赛并荣获9项优胜奖（一等奖2项，二等奖4项，三等奖3项）和4项优秀领队奖，并获得大赛奖金20万元！占据大赛全部30个奖项中的9项，4道大数据赛题全部17个奖项中的8项！共有来自全国高校的230多支队伍参赛。

教学效果

第一届“中国云/移动互联网创新大奖赛”

技术类赛题 1：调色板搜图—在百万图片中搜索与指定调色板相近的图片

技术类赛题 2：多快好省的速递员 — 动态路况环境下的物流规划

技术类赛题 3：你不知道我知道 — 互联网问答系统用户行为分析

技术类赛题 4：难舍难分 — 大规模搜索关键字（短文本）分类

技术类赛题 5：麻雀级云数据中心 — 规定时间内在小规模硬件环境上部署大量虚拟机

创意类竞赛说明： 创意类赛题没有具体的问题约束。

教学效果

第一届“中国云/移动互联网创新大奖赛”

我系4支研究生代表队荣获9项优胜奖和4项优秀领队奖，获得奖金20万

教学效果

第一届“中国云/移动互联网创新大奖赛”

“中国云·移动互联网创新大奖赛”是由“中国云产业联盟”和百度、阿里巴巴、腾讯、用友等国内著名企业和北航、北大等著名高校于2012年5月联合发起组织的第一届全国云计算和互联网创新技术大赛。这次大赛由北航的怀俊鹏院士与中国云产业联盟联合倡议并发起，来自国内多所著名高校和著名企业的十多位专家学者共同参与，是目前为止国内规模和影响最大、级别最高的云计算和互联网创新技术大赛。颁奖仪式上，中国科学院院士怀俊鹏教授、微软集团副总裁陆奇博士到会做了关于云计算的主题报告，百度总裁李彦宏、宽带资本董事长田溯宁、用友软件董事长兼总裁王文京、中国联通总裁陆益民等嘉宾也到会并做了云计算主题对话

教学效果

“全国高校云计算创新应用大赛”大数据技能赛

- 2015-2018年，荣获教育部主办的“全国高校云计算创新应用大赛”大数据技能赛冠军，实现该项赛事全国四连冠！

本科课程

今年是本系第四次作为本科生准出核心平台选修课开设本课程

希望大家积极配合以便为大家开好本课程，圆满完成本课程的教学工作，达到预定的教学目标！

谢谢大家！