厦门大学计算机科学系研究生课程

大数据处理技术 Spark

（2018-2019 学年春季学期）

期末作业说明

主讲教师：林子雨
授课地点：厦门大学海韵教学楼 307 教室
班级主页：http://dblab.xmu.edu.cn/post/12157/
二零一九年五月十五日

目录

大数据处理技术 Spark

2018-2019 学年春季学期期末作业说明

主讲教师：林子雨　ziyulin@xmu.edu.cn

一、 作业题目

基于 Spark 的数据处理与分析

二、 作业目的

综合运用大数据处理框架 Spark、Hadoop 及数据可视化技术，对数据进行存储、处理和分析。

三、 作业性质

必做。作为评定期末总成绩的重要参考。

四、 作业考核方法

作业成绩评定方法如下：

- 不按时交作业、所提交的作业无法打开或抄袭他人作业：零分
- 作业评分范围：0-100 分

温馨提示：作业必须自己独立完成（所有作业全部要求自己独立完成，没有采用团队合作的形式），不得抄袭他人作业，不得直接拷贝厦门大学数据库实验室网站上提供的大数据案例，否则，期末总成绩不及格。

五、 提交日期与方式

1、必须于 2019 年 5 月 28 日（周二）晚 24 时之前提交；

2、提交的内容为压缩文件 RAR 文件，最后把压缩包文件发送到林子雨老师邮箱：ziyulin@xmu.edu.cn（如果邮件太大，可以使用 QQ 邮箱超大附件功能发送）；

3、文件名命名为“姓名学号.rar”，例如“王小明 23020091152890.rar”；

4、文件夹中应该包含实验报告 WORD 文档（需要包含实验过程说明、代码和一些必要的实验过程截图）、软件版本号 TXT 文件（包含作业中用到的所有软件和编程语言的名称和版本号信息）、工程文件以及其他有必要提交的文档，使得老师可以根据这些信息在老师电脑上可以重现实验内容。

六、 作业工具和环境要求

（1） 必须在 Linux 系统下完成作业。

（2） 可以任意选择自己喜欢的开发工具，比如 Eclipse、IntelliJ IDEA 等。

（3） 相关软件的版本要求如下：

- Linux: Ubuntu16.04 或 14.04
- Hadoop: 2.7.1
- Spark：2.1.0

上述软件版本必须和要求的版本号一致，方便老师统一调试。如果同学使用了其他软件，请一定在软件版本号 TXT 文件中明确列出。

七、 作业内容和要求

完整实现数据分析全流程，具体如下：

（1） 从网络上下载一个数据集；

（2） 对数据集进行数据预处理（比如选取部分字段、进行格式转换等），然后保存到 HDFS 中；

（3） 使用 Spark 对数据进行分析，可以任意使用 Spark Core、Spark SQL、Spark Streaming 和 Spark MLlib 组件，只要使用了 Spark 编程知识即可；编程语言需要使用 Python；如果有需要，分析结果也可以选择保存到 MySQL 中；

（4） 对分析结果进行可视化呈现，可以任意选择可视化方法（比如 R 语言可视化、网页可视化以及其他可视化方法）。

八、 参考资料

（1） 大数据学习路线图（http://dblab.xmu.edu.cn/post/bigdataroadmap/）。

（2） 厦门大学林子雨编著《Spark 入门教程》http://dblab.xmu.edu.cn/blog/spark/

（3） 厦门大学数据库实验室制作《Spark 课程实验案例：Spark+Kafka 构建实时分析 Dashboard》（地址：http://dblab.xmu.edu.cn/post/8274/）。

（4） Spark 课程综合实验案例：淘宝双 11 数据分析与预测（http://dblab.xmu.edu.cn/post/8116/）。

（5） 用 Node.js 搭建一个简易的 Web 端文件词频统计动态网页（http://dblab.xmu.edu.cn/blog/1883-2/）。

（6） 厦门大学林子雨主讲《大数据处理技术 Spark》2019 班级主页，里面包含讲义 PPT。班级主页地址：http://dblab.xmu.edu.cn/post/12157/。

（7） 厦门大学林子雨主讲《大数据技术原理与应用》在线课程视频 http://dblab.xmu.edu.cn/post/bigdata-online-course/

（8） 厦门大学数据库实验室编写《大数据软件安装和编程实践指南》，详细介绍如何安装运行各种大数据软件以及如何进行初级编程实践，包括 Hadoop、HDFS、HBase、MapReduce、Spark、MongoDB 等安装、操作、编程指南。在线访问网址：http://dblab.xmu.edu.cn/post/5663/。

（9） 厦门租房信息分析展示 http://dblab.xmu.edu.cn/blog/2307/。

九、 附录 1:教师介绍

林子雨(1978－),男,博士,厦门大学计算机科学系 助理教授,主要研究领域为数据库,数据仓库,数据挖掘,大数据
主讲课程：大数据处理技术
办公地点：厦门大学海韵园科研 2 号楼
E-mail: ziyulin@xmu.edu.cn

林子雨，男，1978 年出生，博士（毕业于北京大学），现为厦门大学计算机科学系助理教授（讲师），曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国计算机学会数据库专业委员会委员，中国计算机学会信息系统专业委员会委员，荣获“2017 中国大数据创新百人”称号。中国高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013 年度和 2017 年度厦门大学奖教金获得者，荣获 2018 年厦门大学高等教育教学成果特等奖，主讲的《大数据技术原理与应用》课程荣获“2018 国家精品在线开放课程”。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》

和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括 1 项国家自然科学青年基金项目(No.61303004)、1 项福建省自然科学青年基金项目(No.2013J05099)和 1 项中央高校基本科研业务费项目(No.2011121049)；作为课题负责人主持的教学项目包括 1 项福建省教改课题和 1 项教育部产学合作育人项目。同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015 泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009 年至今，“数字教师”大平台累计向网络免费发布超过 500 万字高价值的研究和教学资料，累计网络访问量超过 500 万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，年访问量超过 200 万次，成为全国高校大数据教学知名品牌。具有丰富的政府和企业信息化培训经验，厦门大学管理学院 EDP 中心、浙江大学管理学院 EDP 中心、厦门大学继续教育学院、泉州市科技培训中心特邀培训讲师，曾给中国移动通信集团公司、福州马尾区政府、福建龙岩卷烟厂、福建省物联网科学研究院、石狮市物流协会、厦门市物流协会、浙江省中小企业家、四川泸州企业家、江苏沛县企业家等开展信息化培训，累计培训人数达 3000 人以上。