厦门大学计算机科学与技术系研究生课程

大数据处理技术

（2024-2025 学年春季学期）

期末作业说明

主讲教师：林子雨
授课地点：厦大翔安校区西部片区 2 号楼 106
班级主页：https://dblab.xmu.edu.cn/post/spark2025/
二零二五年五月

目录

大数据处理技术

2024-2025 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 和大模型的大数据处理与分析

二、 作业目的

综合运用大数据处理框架 Spark、Hadoop、数据可视化技术和大模型，对数据进行爬取、存储、处理、分析和可视化。

三、 作业性质

必做。作为评定期末总成绩的依据。

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。

作业由林子雨老师亲自批改，本学期无助教。

五、 提交日期与方式

1、必须于 2025 年 5 月 17 日（第 13 周周六）早 8 时到晚 24 时之间提交，不要在其他时间提交；教学周第 14 周（5 月 19 日）到第 16 周（6 月 2 日）的课堂（3 次课），进行作业答辩，学生需要到讲台讲解作业，接受老师的提问。

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的电子邮箱：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 23020191152890.rar”；

4、文件夹中需要包含数据集、实验报告（WORD 文档）、工程文件（含代码）以及其他有必要提交的文档（其中，数据集文件，必须单独保存在一个名称为“数据集”的子目录下，数据集不要太大，建议控制在 50MB 以内）。实验报告必须详细，里面需要包含一些实验过程截图和步骤说明，可以把代码粘贴到 WORD 文档中，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（编程可以在 Windows 下完成，但是，代码必须在 Linux 中运行成功，包括采集、存储、处理、分析和可视化所有环节）。

（2） 必须用到大模型工具（比如 DeepSeek 或 AI 编程工具 Cursor、Cline 等）。

（3） 可以任意选择自己喜欢的编程开发工具，比如 Eclipse、IntelliJ IDEA、VSCode 等，必须使用 AI 编程功能。

（4） 相关软件的版本要求如下（必须严格遵循版本要求）：

- Linux：Ubuntu16.04 及以上版本
- Hadoop：3.3.5
- Spark：3.4.0
- Python3.X
- JDK1.8

（4）安装教程：

- Hadoop3.3.5 安装教程_单机/伪分布式配置
  _Hadoop3.3.5/Ubuntu22.04(20.04/18.04/16.04)
- Spark 安装和编程实践（Spark3.4.0）https://dblab.xmu.edu.cn/blog/4322/

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） 从网络寻找一个网站，使用 Python 语言编写网络爬虫采集数据集（数据集不要太大，建议控制在 50MB 以内）；

（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、去除空值和重复值等），必须使用 Python 语言（必须包含使用 pandas，去除空字段，去除重复值等）。清洗以后的数据保存到文本文件中，并且上传到 HDFS 中（必须使用到 HDFS）。

（3） 使用 Spark（不限定编程语言，可以使用 Python、Scala 或者 Java）对 HDFS 中的数据进行处理和分析，分析结果可以保存到文件或者 MySQL 数据库中。（备注：该步骤考察对 Spark 技术的综合运用能力，该步骤完成的质量，对老师评分高低有重要影响）

（4） 对分析结果进行可视化呈现，可以任意选择可视化方法（比如网页可视化、ECharts 以及其他可视化方法），可以使用任意语言（包括 Python、Java 等在内的任意语言）。

（5） 整个作业，必须至少有一个环节中用到了大模型工具，并且必须给出大模型工具使用过程的详细截图以及使用了什么提示词等，证明大模型在做作业中发挥了重要作用。作业会考察学生对大模型辅助自己编程开发的能力。

八、 参考资料

（1） Spark 数据处理分析案例 https://dblab.xmu.edu.cn/blog/2738/

特别注意：上面参考案例可能不包含 pandas 数据清洗的环节，本次作业必须包含 pandas 数据清洗环节。

附录 1:教师介绍

林子雨(1978－),男,博士,厦门大学计算机科学与技术系副教授,主要研究领域为数据库,数据仓库,数据挖掘,大数据,人工智能
主讲课程：大数据处理技术
办公地点：厦大翔安校区西部片区 5 号楼
E-mail: ziyulin@xmu.edu.cn

林子雨（1978－），男，博士（毕业于北京大学），国内高校知名大数据教师，厦门大学计算机科学与技术系副教授，厦门大学数据库实验室负责人，中国计算机学会数据库专委会委员，中国计算机学会信息系统专委会委员，全国工业大数据行业产教融合共同体特聘专家，入选“2021 年高校计算机专业优秀教师奖励计划”，荣获“2022 年福建省高等教育教学成果奖特等奖（个人排名第一）”和“2018 年福建省高等教育教学成果奖二等奖（个人排名第一）”，编著出版 15 本大数据系列教材，被国内 1000 多所高校采用，建设了国内高校首个大数据课程公共服务平台，平台累计网络访问量超过 2500 万次，成为全国高校大数据教学知名品牌，主持的课程《大数据技术原理与应用》获评“2018 年国家精品在线开放课程”和“2020 年国家级线上一流本科课程”，主持的课程《Spark 编程基础》获评“2021 年国家级线上一流本科课程”。建设的大数据系列 MOOC 课程入选“2023 年教育部国家智慧教育公共服务平台应用典型案例”。撰写的大模型科普报告，被国内广泛传播，全网浏览量超过 1000 万。

推荐阅读：《林子雨老师教学创新成果报告：服务全国高校的大数据教学创新实践》，阅读地址：https://dblab.xmu.edu.cn/post/bigdata-report/