南京大学软件学院

2019-2020 学年本科三年级《云计算》课程作业说明

任桐炜 李传艺

实践内容上课计划：

2019 年 9 月 18 日——Spark、MongoDB 简介、Spark RDD 等

2019 年 9 月 23 日——Machine Learning 简介

2019 年 9 月 25 日——汇报作业设计

2019 年 10 月 14/16/21/23 日——汇报作业完成情况

课程目标：熟悉 Spark 平台的架构及功能，掌握 Spark Streaming、GraphX 和 MLlib 的基础编程能力，找到 Spark 能够发挥作用的业务场景，设计业务问题、通过编程解决问题，并汇报。

组队：各组在 2019 年 9 月 18 日中午 12:00 前提交组队信息，包括小组成员姓名和学号；各组人数推荐为 4 人，最多 6 人（小组人数越多对作业创新性要求越高，设计的业务问题应该越多）。

作业 1. 设计

描述：根据 Spark Streaming、GraphX 和 MLlib 的功能描述及适用场景，寻找相关业务场景并设计相关业务问题。可在同一个业务场景下，对 Streaming、GraphX 和 MLlib 分别设计业务问题；也可以在不同的业务场景下，分别对三种技术设计业务问题。

要求：数据要基于真实存在的业务场景；数据可以是从网络上爬取的数据，可以是一个网站，也可以是从多个网站爬取的；数据也可以是直接从一些大数据竞赛上下载的；一种技术对应的业务问题可以是 1 个或多个；要能够说明使用 Spark 各种技术解决所提业务问题的可行性。

完成过程:

(1) 2019 年 9 月 24 日 23:59 前在课程管理系统中上传介绍业务场景、业务问题、相关数据和可行性的 PPT;

(2) 2019 年 9 月 25 日在课上各组介绍自己的业务场景与研究问题设计，及其价值、意义;

(3) 2019 年 9 月 30 日 23:59 前每组在课程管理系统提交 3 个组内成员最喜欢其他小组的业务设计;

评分标准：被其他小组喜欢的次数与教师评价综合评分。

后续作业选题说明：

各组继续完成自己设计的选题；如果想要做其他组提出的想法，可以向李传艺老师申请；每组提出的想法至多有两个组可以同时做。

作业 2. Spark Streaming 实践

描述：根据业务场景、探究问题的设计，结合 Spark Streaming 技术进行实践。

汇报时间：2019 年 10 月 14 日、2019 年 10 月 16 日

汇报内容：

(1) 回顾业务场景、研究问题；

(2) 介绍数据获取、预处理、存储等工作；

(3) 介绍开发过程中遇到的困难及解决方案；

(4) 展示数据处理结果，回答相关研究问题；

(5) 总结、反思、展望。

评分标准：2019 年 10 月 16 日 23:59 前每组提交组内成员最喜欢 3 个其他小组的作业；综合被其他小组喜欢的次数与教师评价进行作业评分。

作业 3. Spark GraphX 实践

描述：根据业务场景、探究问题的设计，结合 Spark GraphX 技术进行实践。

汇报时间：2019 年 10 月 16 日、2019 年 10 月 21 日

汇报内容：

(1) 回顾业务场景、研究问题；

(2) 介绍数据获取、预处理、存储等工作；

(3) 介绍开发过程中遇到的困难及解决方案；

(4) 展示数据处理结果，回答相关研究问题；

(5) 总结、反思、展望。

评分标准：2019 年 10 月 21 日 23:59 前每组提交组内成员最喜欢 3 个其他小组的作业；综合被其他小组喜欢的次数与教师评价进行作业评分。

作业 4. Spark MLlib 实践

描述：根据业务场景、探究问题的设计，结合 Spark MLlib 技术进行实践。

汇报时间：2019 年 10 月 23 日

汇报内容：

(1) 回顾业务场景、研究问题；

(2) 介绍数据获取、预处理、存储等工作；

(3) 介绍开发过程中遇到的困难及解决方案；

(4) 展示数据处理结果，回答相关研究问题；

(5) 总结、反思、展望。

评分标准：2019 年 10 月 23 日 23:59 前每组提交组内成员最喜欢 3 个其他小组的作业；综合被其他小组喜欢的次数与教师评价进行作业评分。