学号 专业 姓名

实验日期 教师签字 成绩

实验报告

【实验名称】 基于Spark MLlib的开源软件项目流行度预测

【实验目的】

了解GitHub软件项目流行度预测的背景

了解Spark基本原理、Spark程序的运行模式

掌握特征工程的思路，并会使用SparkSQL进行特征提取

掌握Spark Pipeline使用方法

掌握Spark MLlib包中常用回归模型的使用方法

【实验原理】

见操作步骤文档

【实验内容】

**题目**：使用Spark对GitHub中PHP软件项目流行度进行预测，即预测软件项目的watch数。

**要求**：1.使用实验3提供的数据集；

2.可以使用实验3中通过Hive抽取的软件项目事实表作为输入特征，也可以使用SparkSQL进行事实表构建，抽取尽可能多的软件项目特征；

3.划分测试集和训练集，构建Spark机器学习Pipeline，使用回归模型进行软件项目流行度预测；

4.对模型预测效果进行评估，对超参数进行调优；

5.附件提交zeppelin notebook文件。

**提高（选做）：**

1.使用实验数据集以外的数据增强预测效果，例如使用抓取的segmentfault数据、StackOverflow数据等。