任务5-1 使用MLlib进行数据预处理

|  |  |
| --- | --- |
| 任务书 | |
| 一、任务描述与要求 | |
| 任务描述 | 深入了解Apache Spark MLlib库的架构、功能和优势。掌握MLlib在大规模数据处理和机器学习任务中的应用场景。 |
| 具体要求 | 1.高效地加载和转换不同格式的大规模数据集；  2.使用MLlib进行有效的特征选择。 |
| 二、任务目标 | |
| 知识目标 | * 掌握不同模型的构建原理和方法； * 了解模型评估的指标和意义。 |
| 能力目标 | * 能够独立构建有效的数据分析模型； * 具备准确评估模型性能的能力。 |
| 三、知识储备 | |
| 1.MLlib支持的数据类型  Spark MLlib支持多种数据类型,其中最为核心和常用的是向量、矩阵和DataFrame。具体介绍如下：  IMG_256   1. Vector   Vector是MLlib中最基本的数据结构，用于表示特征或数学向量。它分为DenseVector（稠密向量）和SparseVector（稀疏向量）。DenseVector存储所有元素，适用于大多数元素非零的情况；SparseVector只存储非零元素及其索引，适用于大多数元素为零的情况。   1. Matrix   Matrix（矩阵）用于表示二维数据结构。本地矩阵存储在单个机器上，而分布式矩阵跨多个机器存储，适用于大规模数据处理。常见的分布式矩阵类型包括RowMatrix、IndexedRowMatrix、CoordinateMatrix和BlockMatrix。   1. DataFrame   DataFrame是一种结构化数据集，类似于关系数据库表。它具有命名列和已知数据类型，支持SQL查询，并能与Spark生态系统无缝集成。DataFrame是MLlib中进行数据预处理、特征工程和模型训练的主要数据结构。   1. LabeledPoint   LabeledPoint用于监督学习算法，combining特征向量和标签（用于分类或回归）。它的格式为LabeledPoint(label, features)，其中label是预测目标，features是特征向量。  这些数据类型构成了Spark MLlib的基础，使其能够高效地处理各种机器学习任务，从数据预处理到模型训练和评估。选择合适的数据类型对于优化性能和资源利用至关重要。    **2. 特征提取技术**  Spark MLlib提供了多种特征提取技术，用于将原始数据转换为机器学习算法可用的特征向量。以下是几种常用的特征提取方法：   1. TF-IDF   这是一种用于文本分析的重要技术。它衡量一个词在文档集中的重要性。TF-IDF值增加比例与词在文档中出现的次数成正比，同时与词在整个文档集中出现的频率成反比。Spark MLlib提供了HashingTF和IDF类来实现这一技术。   1. Word2Vec   这是一种将单词转换为向量的技术。它基于单词的上下文来学习单词的向量表示，能够捕捉单词之间的语义关系。Spark的Word2Vec实现可以处理大规模文本数据。   1. CountVectorizer   这个特征提取器可以将一组文本文档转换为标记计数的向量。它能够帮助建立一个词汇表，并生成文档-词频矩阵。  IMG_256 | |

|  |
| --- |
| 操作指南 |
| 一、制定计划 |
| |  |  |  |  | | --- | --- | --- | --- | | **序号** | **作业项目** | **序号** | **作业项目** | | 1 |  | 5 |  | | 2 |  | 6 |  | | 3 |  | 7 |  | | 4 |  | 8 |  | | 计划审核 | 审核意见： | | | |
| 二、实施方案 |
| 1. 环境设置和数据加载  首先，需要配置环境设置，并且把数据加载进来：   |  | | --- | | import org.apache.spark.sql.SparkSession  import org.apache.spark.ml.feature.\_  val spark = SparkSession.builder()  .appName("SimpleMLlibPreprocessing").getOrCreate()  // 读取包含文本评论和评分的CSV文件  val data = spark.read.format("csv")  .option("header", "true")  .option("inferSchema", "true")  .load("path/to/reviews.csv") |   2. 文本特征处理  接着处理文本特征，具体代码如下：   |  | | --- | | // 将文本转换为词频向量  val tokenizer = new Tokenizer().setInputCol("review").setOutputCol("words")  val countVectorizer = new CountVectorizer()  .setInputCol("words").setOutputCol("features") |   3.标签处理  接着进行标签处理，具体代码如下：   |  | | --- | | // 将评分转换为数值标签  val labelIndexer = new StringIndexer().setInputCol("rating")  .setOutputCol("label") |   4.创建和应用预处理流水线  接着，进行创建和预处理流水线的操作。具体代码如下：   |  | | --- | | import org.apache.spark.ml.Pipeline  val pipeline = new Pipeline().setStages(Array(tokenizer, countVectorizer, labelIndexer))  val model = pipeline.fit(data)  val processedData = model.transform(data) |   5.查看结果并保存  最终，查看处理结果并保存数据。具体代码如下：   |  | | --- | | // 显示处理后的数据  processedData.select("review", "rating", "features", "label").show(5)  // 保存处理后的数据  processedData.write.parquet("path/to/processed\_reviews")  spark.stop() | |
| 三、实施记录 |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | **序号** | **关键步骤记录** |  | **序号** | **解决问题记录** | | 1 |  |  | 1 |  | | 2 |  |  | 2 |  | | 3 |  |  | 3 |  | | 4 |  |  | 4 |  | | 5 |  |  | 5 |  | | 6 |  |  | 6 |  | | 7 |  |  | 7 |  | | 8 |  |  | 8 |  | | 9 |  |  | 9 |  | |
| 四、任务拓展 |
| 目前使用的Spark MLlib数据预处理任务已完成了基本的文本处理和标签编码。为了进一步提升数据的质量和模型的性能，需要对现有任务进行拓展。此次拓展任务旨在通过增强的数据处理技术，提供更丰富的特征和改进的数据质量。  对现有的数据预处理任务进行以下拓展：文本处理阶段，添加停用词移除步骤，以去除无意义的常用词汇。使用正则表达式 tokenizer 替代现有的 tokenizer，以提高文本分词的精确性。 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评价标准 | | | | | | | | |
| 一、基本信息 | | | | | | | | |
| 项目名称 | | 单元五 Spark机器学习库（MLlib） | | | | | | |
| 任务名称 | | 任务5-1 使用MLlib进行数据预处理 | | | | | | |
| 班级 |  | | 学号 | |  | 姓名 | |  |
| 组名 |  | | 学时 | |  | 日期 | |  |
| 组员  分工 |  | | |  | | |  | |
|  | | |  | | |  | |
| 二、任务检查评价单 | | | | | | | | |
| |  |  |  | | --- | --- | --- | | **评价内容** | **评价标准** | **得分** | | 知识运用  （20分） | 掌握相关理论知识，理解本次任务要求，制定详细计划，计划条理清晰，逻辑正确（20分） |  | | 理解相关理论知识，能根据本次任务要求、制定合理计划（15分） | | 了解相关理论知识，有制定计划（10分） | | 无制定计划（0分） | | 专业技能  （40分） | 结果验证全部满足。（40分） |  | | 结果验证只有一个功能不能实现，其它功能全部实现（30分） | | 结果验证只有一个功能实现，其它功能全部没有实现（20分） | | 结果验证功能均未实现（0分） | | 核心素养  （20分） | 具有良好的自主学习能力、分析解决问题的能力、整个任务过程中有指导他人（20分） |  | | 具有较好的学习能力和分析解决问题的能力，任务过程中无指导他人（15分） | | 能够主动学习并收集信息，有请教他人进行解决问题的能力（10分） | | 不主动学习（0分） | | 课堂纪律  （20分） | 设备无损坏、设备摆放整齐、工位区域内保持整洁、无干扰课堂秩序（20分） |  | | 设备无损坏、无干扰课堂秩序（15分） | | 无干扰课堂秩序（10分） | | 干扰课堂秩序（0分） | | 得分（满分100） | |  | | 综合评价 | ☆☆☆☆☆ | | | | | | | | | | |