任务9-2 智能推荐系统

|  |  |
| --- | --- |
| 任务书 | |
| 一、任务描述与要求 | |
| 任务描述 | 从文件、数据库或者其他数据源中选择并加载需要进行复杂查询的数据集，同时检查数据的格式、完整性和规模等信息。 |
| 具体要求 | 1.通过Spark SQL进行多表连接；  2.应用窗口函数进行复杂的数据分析。 |
| 二、任务目标 | |
| 知识目标 | * 洞悉智能推荐系统的算法逻辑； * 熟知推荐系统的评价标准。 |
| 能力目标 | * 可以独立开发简易智能推荐系统； * 拥有提升推荐系统效能的能力。 |
| 三、知识储备 | |
| 1. **推荐系统的基本原理**   推荐系统的基本原理涉及多个方面，通过综合分析用户行为、物品特征和算法模型，提供个性化的推荐。以下是更详细的描述：   1. 数据收集   推荐系统的第一步是数据收集，包括：  用户行为数据：这包括用户的点击、浏览、搜索、收藏和购买等行为。这些数据能够揭示用户的兴趣和偏好。用户特征数据：用户的基本信息，如年龄、性别、地理位置等。这些信息帮助系统理解用户的背景，从而提供更符合其需求的推荐。  物品特征数据：物品的详细属性，这些特征数据帮助系统理解物品的内容和类别。   1. 用户和物品建模   在数据收集后，推荐系统会构建用户模型和物品模型：  用户模型：通过分析用户的行为数据（如购买记录、浏览历史）和特征数据，系统构建用户偏好的模型。物品模型：通过分析物品的特征数据，系统构建物品的描述模型。    **2.使用Spark MLlib实现协同过滤**  协同过滤是一种广泛使用的推荐系统技术，其核心思想是通过分析用户之间或物品之间的相似性来进行个性化推荐。以下是对协同过滤的重点介绍：  基本原理：协同过滤的基本原理是利用用户历史行为数据来预测用户对未见物品的兴趣。具体来说，如果两个用户对一组物品的评价相似，那么他们可能对其他物品的评价也会类似。这种方法可以分为用户协同过滤和物品协同过滤两种类型。 | |

|  |
| --- |
| 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. 准备数据集  在这个案例中，将构建一个简单的电影推荐系统，使用Apache Spark来处理数据和生成推荐。将使用协同过滤算法来推荐电影，这种算法通过分析用户的评分数据来预测用户对未评分电影的喜好。将使用一个简化的电影评分数据集。这里用python语言编写，PySpark框架完成。假设有一个CSV文件 ratings.csv 和 movies.csv，内容如下：   |  | | --- | | ratings.csv  userId,movieId,rating,timestamp  1,1,5,964982703  1,2,3,964981247  2,1,4,964982703  2,3,2,964982703  3,2,5,964981247  3,3,3,964982703  movies.csv  movieId,title,genre  1,Toy Story (1995),Animation|Children's|Comedy  2,Jumanji (1995),Adventure|Children's|Fantasy  3,Grumpier Old Men (1995),Comedy|Romance |  1. **环境准备**   确保你已安装Apache Spark和相关的Python库。你可以使用PySpark来运行Spark应用。完整代码如下：   |  | | --- | | pip install pyspark |   3.创建Spark应用  以下是一个基于Spark的推荐系统的简单Python代码示例，使用协同过滤算法（ALS算法）来生成推荐。完整代码如下：   |  | | --- | | from pyspark.sql import SparkSession  from pyspark.ml.recommendation import ALS  from pyspark.sql.functions import col  # 创建SparkSession  spark = SparkSession.builder \  .appName("MovieRecommendationSystem") \  .getOrCreate()  # 加载数据  ratings = spark.read.csv("ratings.csv", header=True, inferSchema=True)  movies = spark.read.csv("movies.csv", header=True, inferSchema=True)  # 显示数据  ratings.show()  movies.show()  # 划分训练集和测试集  (training, test) = ratings.randomSplit([0.8, 0.2])  # 创建ALS模型  als = ALS(  maxIter=10,  regParam=0.01,  userCol="userId",  itemCol="movieId",  ratingCol="rating",  coldStartStrategy="drop"  )  # 训练模型  model = als.fit(training)  # 进行预测  predictions = model.transform(test)  # 评估模型  from pyspark.ml.evaluation import RegressionEvaluator  evaluator = RegressionEvaluator(  metricName="rmse",  labelCol="rating",  predictionCol="prediction"  )  rmse = evaluator.evaluate(predictions)  print(f"Root-Mean-Square Error = {rmse}")  # 生成推荐  userRecs = model.recommendForAllUsers(10)  # 显示推荐结果  userRecs.show()  # 为特定用户生成推荐  user\_id = 1  user\_recs = userRecs.filter(col("userId") == user\_id)  user\_recs.show(truncate=False)  # 结束SparkSession  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 |
| 为了进一步提升电影推荐系统的性能和实用性，提高推荐系统的准确性，通过融合内容特征提升推荐质量。根据用户的上下文（如时间、地点、设备等）提供更准确的推荐。 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评价标准 | | | | | | | | |
| 一、基本信息 | | | | | | | | |
| 项目名称 | | 单元九 Spark项目实战 | | | | | | |
| 任务名称 | | 任务9-2 智能推荐系统 | | | | | | |
| 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 | ☆☆☆☆☆ | | | | | | | | | | |