MapReduce课程设计选题

- 课程设计1 - 体育赛事日志分析
- 课程设计2 - 人物关系挖掘
- 课程设计3 - 新闻自动分类

课程设计1—体育赛事日志分析

- 1. 课程设计目标

  本课程设计通过使用 MapReduce 实现比赛日志分析。

  通过本课程设计的学习，利用 MapReduce 工具实现大数据下的数据分析方法。

- 2. 学习技能

  本次课程设计可以掌握以下 MapReduce 编程技能：

  1. 海量日志数据的统计分析
  2. 基于 MapReduce 的预测模型设计

课程设计1—体育赛事日志分析

- 3. 题目描述
  - 各项体育赛事中，根据运动员在场上的具体表现情况，会产生大量的数据。在职业体育赛事中，对赛事过程中产生的日志进行分析，可以有效分析对手的技战术特点，从而可以帮助教练团队制定相应策略予以应对。
  - 本课程设计数据中记录了一系列篮球赛事的比赛日志，要求学生按照要求进行比赛日志的统计、分析，并根据已有的比赛日志预测后续赛事的比赛结果。

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 以现实数据为例

1st Q				
Time	Boston	Score		Golden State
12:00.0	Jump ball: R. Williams vs. K. Looney (M. Smart gains possession)			
11:42.0	J. Tatum misses 2-pt jump shot from 21 ft	0-0		
11:40.0		0-0		Defensive rebound by S. Curry
11:25.0		0-0		A. Wiggins misses 3-pt jump shot from 27 ft
11:23.0	Defensive rebound by A. Horford	0-0		
11:15.0	J. Tatum misses 3-pt jump shot from 26 ft	0-0		
11:15.0		0-0		Defensive rebound by Team
11:06.0		0-0		S. Curry misses 3-pt jump shot from 26 ft
11:03.0		0-0		Offensive rebound by K. Looney
11:02.0		0-3	+3	S. Curry makes 3-pt jump shot from 26 ft (assist by K. Looney)
10:30.0	J. Brown misses 2-pt jump shot from 17 ft	0-3		
10:28.0	Offensive rebound by Team	0-3		
10:19.0	J. Tatum makes 3-pt jump shot from 28 ft (assist by A. Horford) +3	3-3		
10:01.0		3-5	+2	A. Wiggins makes 2-pt layup from 6 ft (assist by K. Looney)
9:50.0	M. Smart makes 3-pt jump shot from 26 ft (assist by J. Tatum) +3	6-5		
9:35.0		6-5		K. Looney misses 2-pt jump shot from 16 ft
9:32.0	Defensive rebound by J. Brown	6-5		
9:23.0	M. Smart misses 3-pt jump shot from 23 ft	6-5		
9:21.0		6-5		Defensive rebound by D. Green
9:17.0		6-8	+3	K. Thompson makes 3-pt jump shot from 23 ft (assist by S. Curry)
9:00.0	Personal foul by A. Wiggins (drawn by J. Tatum)	6-8		

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 日志文件
  - 日志文件的结构如下：
    - Date：比赛日期
    - AwayTeam 和 HomeTeam：参与比赛的球队，区分主客场
    - PlayBy：产生该条日志的球队名称
    - Quarter：事件发生的节次（1~4节，加时赛从5开始递增）
    - SecLeft：事件发生时该节的剩余时间（按秒计算）
    - 其它字段：根据不同的日志类型，会有不同的字段被填入。
  - 日志中的元素
    - 日期：从 2000/1/1 至 2000/6/25 不等。
    - 球队：从 team001 - team030，共 30 支球队。
    - 球员姓名：由日志生成器随机生成。保证所有球员姓名不重复。

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 日志文件
  - 日志文件中不同类型的事件：
    - 投篮事件
      - Shooter*：投篮运动员姓名
      - ShotType*：投篮类型（2 分或 3 分）
      - ShotOutcome*：投篮结果（命中 make 或未命中 miss）
      - 注意：投篮事件可能和犯规事件一起被记录
      - 注意：投篮命中时，可能和助攻事件一起被记录
      - 注意：投篮不中时，可能和封盖事件一起被记录
      - 事件结果：投篮命中时，投篮运动员得分 +2 或 +3

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 日志文件
  - 日志文件中不同类型的事件：
    - 助攻事件
      - Assister*：助攻运动员姓名
      - 注意：只会和投篮事件一起出现，不会独立出现
      - 事件结果：助攻运动员助攻数 +1
    - 封盖事件
      - Blocker*：封盖运动员姓名
      - 注意：只会和投篮事件一起出现，不会独立出现
      - 事件结果：封盖运动员封盖数 +1

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 日志文件
  - 日志文件中不同类型的事件：
    - 篮板事件
      - ReboundPlayer*：抢得篮板的运动员姓名或“Team”
      - ReboundType*：失误类型（防守篮板或进攻篮板）
      - 注意：若不能确定具体抢到篮板的个人，则标记为“Team”
    - 罚球事件
      - FreeThrowShooter*：罚球运动员姓名
      - FreeThrowOutcome*：罚球结果（命中或不中）
      - 事件结果：罚球命中时，罚球运动员得分 +1

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 日志文件
  - 日志文件中不同类型的事件：
    - 犯规事件
      - Fouler*：犯规运动员姓名或“Team”
      - FoulType*：犯规类型
    - 违例事件
      - ViolationPlayer*：违例运动员姓名“Team”
      - ViolationType*：违例类型
    - 注意：若不是某运动员的犯规或违例，则标记为“Team”

数据集中没有

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 日志文件
  - 日志文件中不同类型的事件：
    - 失误事件
      - TurnoverPlayer*：失误运动员姓名或“Team”
      - TurnoverType*：失误类型
      - 注意：失误事件可能和抢断事件一起被记录
    - 抢断事件
      - TurnoverCauser：抢断运动员
      - 注意：只会和失误事件一起出现，不会独立出现
      - 事件结果：抢断运动员抢断 +1

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 日志文件
  - 日志文件中不同类型的事件：
    - 换人事件
      - EnterGame*：被换上场的球员
      - LeaveGame*：被换下场的球员
    - 其它事件：

      未结构化的其它场上事件，用空白行表示。

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 实验任务
  - 任务1：统计每场比赛的比赛结果
    - FreeThrowMade 为 make 时，PlayBy 球队得 1 分
    - ShotOutcome 为 make 时：
      - 若 ShotType 为 2-pt ***，PlayBy 球队得 2 分
      - 若 ShotType 为 3-pt ***，PlayBy 球队得 3 分
    - 针对每条日志计算得分情况后按比赛计算总得分
    - 提示：各支球队每天只会有一场比赛
    - 输出格式：日期，主队，主队比分，客队，客队比分

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 实验任务
  - 任务2：计算数据集中各项技术统计的前五名球员
    - 提示：根据每条日志产生行为的制造队员进行统计
    - 要求得到得分、篮板、助攻、抢断、盖帽最多的 5 名球员
  - 任务3：预测给定比赛的各队胜率
    - 根据已有的数据作为训练数据，设计预测算法，预测给定几组对阵中主队和客队的胜率。
    - 根据比较数据生成模型本身预测胜率的差值，判断模型的准确度，但本任务更看重算法设计部分。

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 实验任务
  - 任务4：设计合理的评价标准，评选出表现最好的 5 名球员
    - 可以考虑的因素：
      - 球员技术统计数据
      - 球员所在球队的战绩
      - 实现或定义高阶指标，以评价球员表现
    - 根据评价标准，选出最好的 5 名球员（排名分先后）。

课程设计1—体育赛事日志分析

- 3. 题目描述 —— 实验任务
  - 任务5：分析 team025 和 team028 的比赛特点（选做）
    - 可以考虑的分析方向：
      - 分析两队球员的出场时间、投篮方式分布
      - 分析两队的轮换策略（各节偏好的出场球员）
      - 分析两队的关键球打法（比分接近时，比赛最后时刻由谁出手）
    - 根据分析出的比赛特点，从其中一队教练的角度出发，尝试提出对抗另一只球队的策略。

课程设计1—体育赛事日志分析

- 4. 提交作业
  - 程序源代码，要求提供包含完整目录结构的 src 代码包，并提供编译和执行方法说明
  - 程序可执行 jar 包以及 jar 包的执行方式。本课程设计的运行环境为 hadoop-2.7、jdk-1.7 或以上环境
  - 程序设计报告。报告内容包括程序设计的主要流程、程序采用的主要算法、进行的优化工作、优化取得的效果、程序的性能分析以及程序运行截图等。