# 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 - 哈利波特的魔法世界
- 课程设计3 - 新闻自动分类

# 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 - 哈利波特的魔法世界
- 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

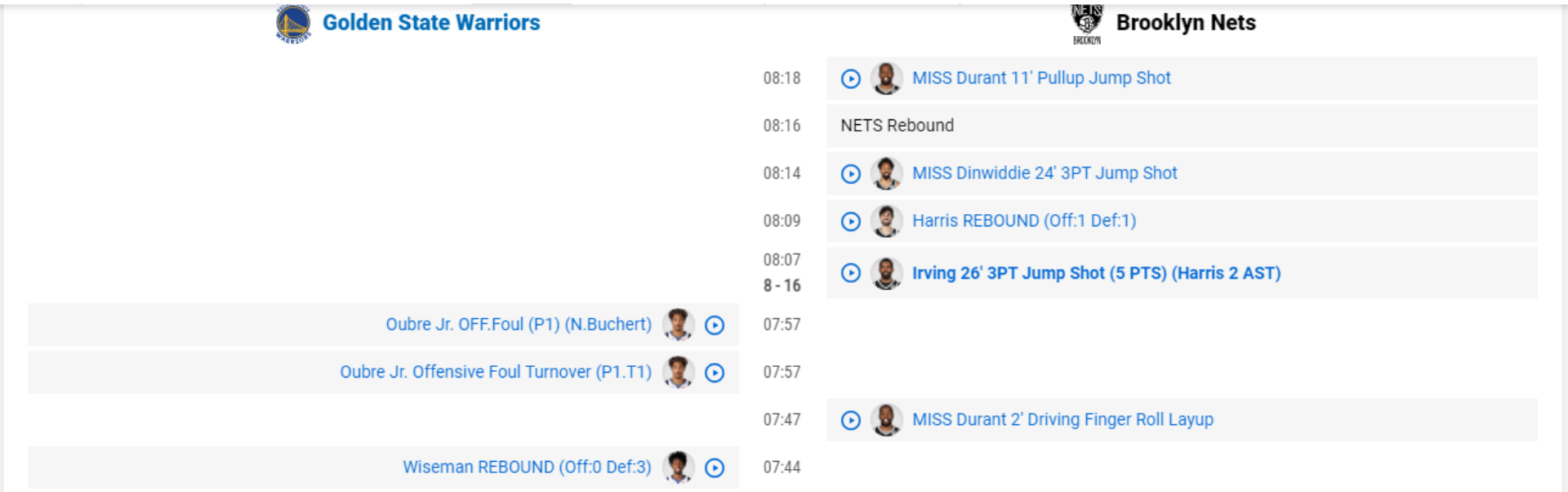
# 课程设计1—体育赛事日志分析

- 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

| Date | Quarter | Sec Left | Away Team | Home Team | Shooter | ShotType | Shot Outcome | Assister | Blocker | FoulType | Fouler | Rebounder | Rebound Type | Violation Player | Violation Type | Timeout Team | FreeThrow Shooter | FreeThrow Outcome | Turnover Player | Turnover Type | Turnover Causer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| December 22 2020 | 1 | 498 | GSW | BRK | K. Durant | 2-pt jump shot | miss | | | | | | | | | | | | | | |
| December 22 2020 | 1 | 496 | GSW | BRK | | | | | | | | Team | offensive | | | | | | | | |
| December 22 2020 | 1 | 494 | GSW | BRK | S. Dinwiddie | 3-pt jump shot | miss | | | | | | | | | | | | | | |
| December 22 2020 | 1 | 489 | GSW | BRK | | | | | | | | J. Harris | offensive | | | | | | | | |
| December 22 2020 | 1 | 487 | GSW | BRK | K. Irving | 3-pt jump shot | make | J. Harris | | | | | | | | | | | | | |
| December 22 2020 | 1 | 477 | GSW | BRK | | | | | | offensive | K. Oubre | | | | | | | | | | |
| December 22 2020 | 1 | 477 | GSW | BRK | | | | | | | | | | | | | | | K. Oubre | offensive foul | |
| December 22 2020 | 1 | 467 | GSW | BRK | K. Durant | 2-pt layup | miss | | | | | | | | | | | | | | |
| December 22 2020 | 1 | 464 | GSW | BRK | | | | | | | | J. Wiseman | defensive | | | | | | | | |



# 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 11 月至次年 3 月不等，格式例如 11-1、3-31。
    - 球队：从 team001 - team030，共 30 支球队。
    - 球员姓名：随机生成的球员姓名，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”

# 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：计算每条日志对比赛比分的影响
    - 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，输出“XXX 队得 X 分”或“不得分”

# 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：统计每场比赛的比赛结果
    - 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
  - 任务3：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

# 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：预测给定比赛中，主队和客队的胜率
    - 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。
    - 比较各个预测值与数据生成模型本身预测胜率的差值，判断模型的准确度。
    - 注意：本实验对预测算法设计的评估是主要的，对于预测结果是否准确的评估是次要的。

# 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：设计合理的评价标准，评选出赛季最有价值球员（选做）
    - 可以考虑的因素：
      - 球员场均数据
      - 球员所在球队的战绩
      - 合理设计现成的或新的高阶指标，以评价球员表现
    - 根据评价标准，选出最好的 5 名球员。其中，第 1 名为最有价值球员，第 2 至第 5 名为最有价值球员候选人。

# 课程设计1—体育赛事日志分析


- 4. 提交作业
  - 源代码
  - 可执行 JAR 包及 JAR 包的执行方式
  - 程序设计报告，包括方法设计、执行流程、算法优化、实验结果、代码运行过程截图等
  - 任务 4 中胜率预测的预测结果
  - 任务 5 中得到的五名最有价值球员候选人（选做）

# 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 - 哈利波特的魔法世界
- 课程设计3 - 新闻自动分类

# 课程设计2—哈利波特的魔法世界


## 1　课程设计目标

通过一个综合数据分析案例：”哈利波特的魔法世界——哈利波特系列小说中的人物关系挖掘“，来学习和掌握 MapReduce 程序设计。通过本课程设计的学习，可以体会如何使用 MapReduce 完成一个综合性的数据挖掘任务，包括全流程的数据预处理、数据分析、数据后处理等。

# 课程设计2—哈利波特的魔法世界


## 2 学习技能

通过本课程设计，可以熟悉和掌握以下 MapReduce 编程技能：

1. 在 Hadoop 中使用第三方的 Jar 包来辅助分析；
2. 掌握简单的 MapReduce 算法设计：
   a) 单词同现算法；
   b) 数据整理与归一化算法；
   c) 数据排序（选做）；
3. 掌握带有迭代特性的 MapReduce 算法设计：
   a) PageRank 算法；
   b) 标签传播（Label Propagation）算法（选做）。

# 课程设计2—哈利波特的魔法世界

## 3　任务描述

本课程设计包括如下的若干任务。这些任务组合起来，就构成了一个完整的人物关系分析流程。

### 任务 1 数据预处理

本任务的主要工作是从原始的哈利波特系列小说的文本中，抽取出与人物互动相关的数据，而屏蔽掉与人物关系无关的文本内容，为后面的基于人物共现的分析做准备。

#### 输入输出

数据输入：1.哈利波特系列小说文集（未分词）；2. 哈利波特系列小说中的人名列表。

数据输出：分词后保留人名。

#### 示例

输入：哈利波特与魔法石中的某一段内容

“**哈利**！”**纳威**一看见他们两个，就脱口而出，“我一直在找你们，想给你们提个醒儿，我听见**马尔福**说他要来抓**哈利**你，他说你有一条龙——”

输出：哈利 纳威 马尔福 哈利

# 课程设计2—哈利波特的魔法世界

## 任务 2 特征抽取：人物同现统计

本任务的重要完成基于单词同现算法的人物同现统计。在人物同现分析中，如果两个人在原文的同一段落中出现，则认为两个人发生了一次同现关系。我们需要对人物之间的同现关系次数进行统计，同现关系次数越多，则说明两人的关系越密切。

### 输入输出

输入：任务 1 的输出；

输出：在哈利波特系列小说中，人物之间的同现次数。

### 示例

输入：

穆迪 哈利 **哈利波特** 赫敏 罗恩
哈利 罗恩 罗恩

输出：

<穆迪，哈利> 1
<穆迪，赫敏> 1
<穆迪，罗恩> 1
<哈利，穆迪> 1
<哈利，赫敏> 1
<哈利，罗恩> 2
<赫敏，穆迪> 1
<赫敏，哈利> 1
<赫敏，罗恩> 1
<罗恩，穆迪> 1
<罗恩，哈利> 2
<罗恩，赫敏> 1

# 课程设计2—哈利波特的魔法世界

注：小说中全文中对于人物名称的使用并不统一。例如部分章节使用 Given Name，部分章节使用 Given Name + Family Name，或者部分章节使用名称，部分章节使用外号等。为了提高分析结果的准确性，请将小说中的**主要人物**的名称进行统一，次要人物可不进行处理。例如 将哈利、哈利·波特、哈利波特的结果进行统一。

# 课程设计2—哈利波特的魔法世界


## 任务 3 特征处理：人物关系图构建与特征归一化

当获取了人物之间的共现关系之后，我们就可以根据共现关系，生成人物之间的关系图了。人物关系图使用邻接表的形式表示，方便后面的 PageRank 计算。在人物关系图中，人物是顶点，人物之间的互动关系是边。人物之间的互动关系靠人物之间的共现关系确定。如果两个人之间具有共现关系，则两个人之间就具有一条边。两人之间的共现次数体现出两人关系的密切程度，反映到共现关系图上就是边的权重。边的权重越高则体现了两个人的关系越密切。

为了使后面的方便分析，还需要对共现次数进行归一化处理：将共现次数转换为共现概率，具体的过程见后面的示例。

### 输入输出

输入：任务 2 的输出

输出：归一化权重后的人物关系图

# 课程设计2—哈利波特的魔法世界

示例

输入：

<穆迪，哈利> 1
<穆迪，赫敏> 1
<穆迪，罗恩> 1
<哈利，穆迪 > 1
<哈利，赫敏 > 1
<哈利，罗恩 > 2
<赫敏，穆迪 > 1
<赫敏，哈利 > 1
<赫敏，罗恩 > 1
<罗恩，穆迪> 1
<罗恩，哈利> 2
<罗恩，赫敏> 1

输出：

穆迪 [哈利,0.33333|赫敏，0.333333|罗恩 0.333333]
哈利 [穆迪,0.25 |赫敏，0.25|罗恩 0.5]
赫敏 [穆迪,0.33333|哈利，0.333333|罗恩 0.333333]
罗恩 [穆迪 0.25|哈利,0.5|赫敏，0.25]

首先是将统计出的人物共现次数结果，转换成邻接表的形式表示：每一行表示一个邻接关系。

“哈利 [穆迪,0.25 |赫敏，0.25|罗恩 0.5] ”表示了顶点“哈利”，有三个邻接点，分别是“穆迪”、“赫敏”和“罗恩”，对应三条邻接边，每条邻接边上分别具有不同的权重。这个邻接边的权重是依靠某个人与其他人共现的“概率”得到的，以“哈利”为例，他分别与三个人（“穆迪”共现 1 次、“赫敏”，共现 1 次、“罗恩”共现 2 次）有共现关系，则哈利与三个人共现的“概率”分别为 1/(1+1+2) = 0.25，1/(1+1+2) = 0.25， 2/(1+1+2) = 0.5。这三个“概率”值对应与三条边的权重。通过这种归一化，我们确保了某个顶点的出边权重的和为 1。

# 课程设计2—哈利波特的魔法世界


## 任务 4 数据分析：基于人物关系图的 PageRank 计算

在给出人物关系图之后，我们就可以对人物关系图进行一个数据分析。其中一个典型的分析任务是：PageRank 值计算。通过计算 PageRank，我们就可以定量地分析出哈利波特系列小说的"主角"们是哪些。

## 输入输出

输入：任务 3 的输出

输出：人物的 PageRank 值

该任务默认的输出内容是杂乱的，从中无法直接的得到分析结论。可以对人物的 PageRank 值进行全局排序，从而很容易地发现 PageRank 值最高的几个人物。排序工作可以通过一个排序 MapReduce 程序完成，也可以将 PageRank 值导入 Hive 中，然后利用 Hive 完成排序。

# 课程设计2—哈利波特的魔法世界


## 任务 5 数据分析：在人物关系图上的标签传播（选做）

标签传播（Label Propagation）是一种半监督的图分析算法，他能为图上的顶点打标签，进行图顶点的聚类分析，从而在一张类似社交网络图中完成社区发现（Community Detection）。图 1 中人物顶点的颜色就是根据标签传播的结果进行的染色。

## 参考资料

1. 英文资料：标签传播算法英文原始文献可参考原始英文论文中的 *III. COMMUNITY DETECTION USING LABEL PROPAGATION* 一节内容。

2. 中文资料：http://www.cnphp6.com/archives/24136

## 输入输出

输入：任务 3 的输出

输出：人物的标签信息

对于该任务的输出，可以通过写一个 MapReduce 程序，将属于同一个标签的人物输出到一起，便于人来查看标签传播的结果。

原始英文论文：https://journals.aps.org/pre/abstract/10.1103/PhysRevE.76.036106

# 课程设计2—哈利波特的魔法世界

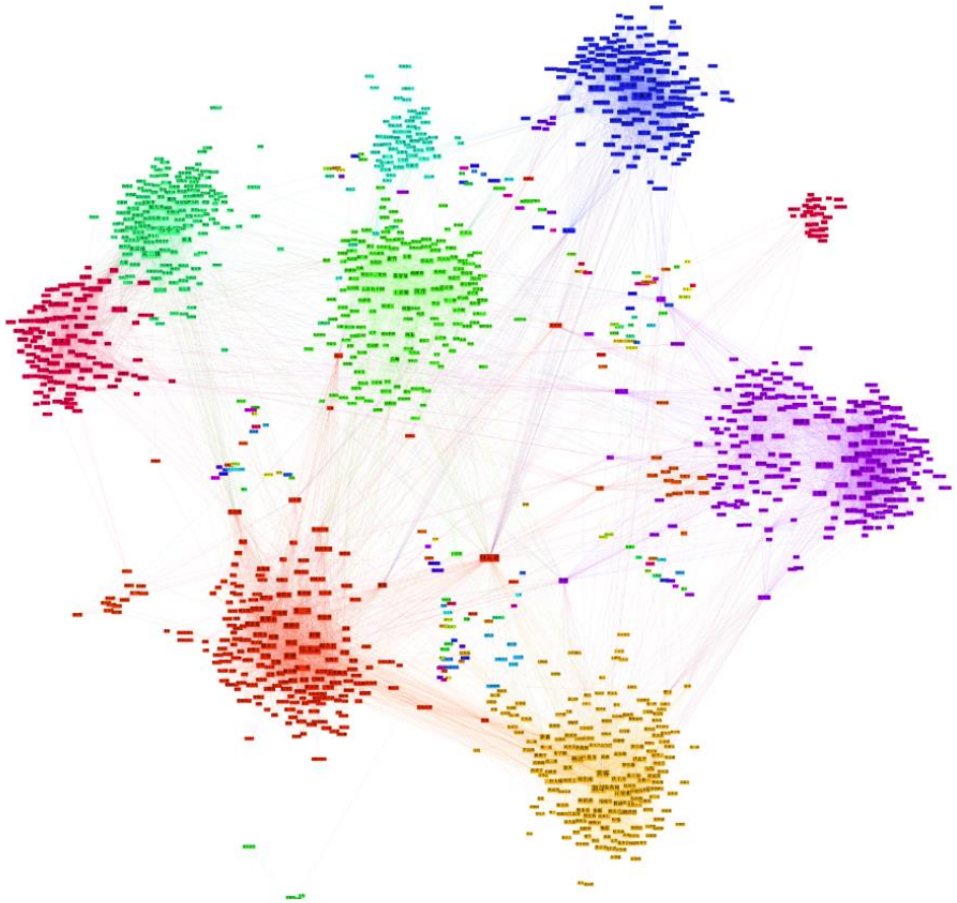


图 1 标签传播的结果展示

注：人物名字的大小由人物顶点的度数确定,人物标签的颜色根据标签传播算法的分析结果确定。

# 课程设计2—哈利波特的魔法世界


## 4　提交材料

请各位同学提交如下材料。

1、程序源代码，要求提供包含完整目录结构的 src 代码包，并且提供编译方法说明。

2、程序可执行 jar 包以及 jar 包的执行方式。本题目的运行环境在 hadoop-2.7、jdk-1.7 及以上的环境下，必须采用 MapReduce 编程模型。

3、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。

# 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 - 哈利波特的魔法世界
- 课程设计3 - 新闻自动分类

# 课程设计3—新闻自动分类


## 1. 课程设计目标

本课程设计的目标是通过 MapReduce 和基本的机器学习方法来实现对新闻的自动分类。通过本课程设计，可以学习如何使用 MapReduce 完成一个综合的数据挖掘任务，包括数据预处理，机器学习建模、样本预测等。

## 2. 学习技能

通过本课程设计，可以熟悉或掌握以下 MapReduce 编程技巧：

- 在 Hadoop 中使用第三方的 Jar 包来辅助分析
- MapReduce 算法设计
  - 文本特征选择算法
  - 文本特征表示算法
  - 文本分类算法

# 课程设计3—新闻自动分类


**3. 任务描述**

在日常生活中，我们所看到的新闻通常伴随着相应类别，例如政治、经济、科教等等。不同的新闻包含不同的主题特征。本课程设计的任务是通过 MapReduce 技术实现新闻文本的自动分类。具体包含如下若干任务，这些任务组合起来就构成了一个完整的新闻文本分类流程。

使用语料：某门户网站新闻列表

# 课程设计3—新闻自动分类


| | | |
|---|---|---|
| test | 2021/6/15 8:17 | 文件夹 |
| train | 2021/6/15 8:17 | 文件夹 |

| | | | | |
|---|---|---|---|---|
| | | 财经_798977.txt | 2021/6/15 8:20 | 文本文档 |
| 财经 | 2021/6/15 8:20 | 财经_798978.txt | 2021/6/15 8:20 | 文本文档 |
| 彩票 | 2021/6/15 8:28 | 财经_798979.txt | 2021/6/15 8:20 | 文本文档 |
| 房产 | 2021/6/15 8:28 | 财经_798980.txt | 2021/6/15 8:20 | 文本文档 |
| 股票 | 2021/6/15 8:21 | 财经_798981.txt | 2021/6/15 8:20 | 文本文档 |
| 家居 | 2021/6/15 8:29 | 财经_798982.txt | 2021/6/15 8:20 | 文本文档 |
| 教育 | 2021/6/15 8:27 | 财经_798983.txt | 2021/6/15 8:20 | 文本文档 |
| 科技 | 2021/6/15 8:22 | 财经_798984.txt | 2021/6/15 8:20 | 文本文档 |
| 社会 | 2021/6/15 8:23 | 财经_798985.txt | 2021/6/15 8:20 | 文本文档 |
| 时尚 | 2021/6/15 8:26 | 财经_798986.txt | 2021/6/15 8:20 | 文本文档 |
| 时政 | 2021/6/15 8:25 | 财经_798987.txt | 2021/6/15 8:20 | 文本文档 |
| 体育 | 2021/6/15 8:31 | 财经_798988.txt | 2021/6/15 8:20 | 文本文档 |
| 星座 | 2021/6/15 8:24 | 财经_798989.txt | 2021/6/15 8:20 | 文本文档 |
| 游戏 | 2021/6/15 8:24 | 财经_798990.txt | 2021/6/15 8:20 | 文本文档 |
| 娱乐 | 2021/6/15 8:30 | 财经_798991.txt | 2021/6/15 8:20 | 文本文档 |

新闻分类目录结构，新闻的类别在文件名中用 _ 隔开；

# 课程设计3—新闻自动分类

## 任务 1 文本特征选择

本任务的主要工作是对原始新闻中的文本进行特征选择，选择能够表征新闻特性的特征词，为后续的文本分类做准备。

输入输出

输入：

1. 新闻文本训练数据和测试数据；
2. 停用词表

输出：

新闻文本特征

# 课程设计3—新闻自动分类


图 2　特征词

# 课程设计3—新闻自动分类

## 任务 2 文本特征表示

基于任务 1 得到的特征词，为每条新闻文本计算特征表示。

输入输出

输入：1. 任务 1 的输出；2. 新闻文本数据
输出：每条新闻文本的特征向量

示例：

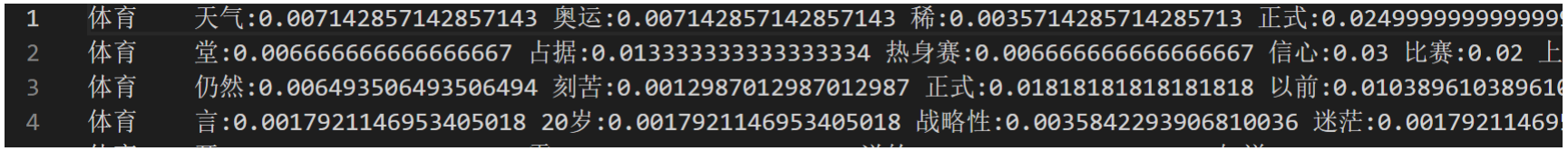



图 3 样本特征表示

# 课程设计3—新闻自动分类


## 任务 3 文本分类

得到了每个新闻的特征向量之后，就可以利用机器学习分类算法实现新闻文本的分类。具体采用何种分类算法，请同学们自行选择，也可以验证多种分类算法的优劣。

## 任务 4 样本预测

在任务 3 中得到了分类模型，接下来使用该模型对测试数据中的新闻文本进行分类，输出分类结果。

## 附加题（选做）

考虑另外使用 Apache Spark 实现新闻自动分类算法，并对新闻文本进行分类，输出分类结果。

# 课程设计3—新闻自动分类


**4. 提交材料**

请各位同学提交如下材料：

1. 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
2. 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
3. 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。

# MapReduce课程设计

- 最终课题完成与提交


## ■ 课程设计结果提交（以下内容打包提交）

- 课程设计报告，内容包括

1. 小组信息（人员，学号，联系信息，导师及研究领域）
2. 课题小组分工：需要明确说明各成员在整个课题中分工负责完成的内容
3. 课程设计题目
4. 摘要
5. 研究问题背景
6. 主要技术难点和拟解决的问题，尤其要解释说明哪些地方、为什么需要采用MapReduce
7. 主要解决方法和设计思路，尤其要解释说明如何采用MapReduce并行化算法解决问题
8. 详细设计说明，包括详细算法设计、程序框架、功能模块、主要类的设计说明，包括主要类、函数的输入输出参数、**尤其是map和reduce函数的输入输出键值对详细数据格式和含义**，主要功能和算法代码中加清晰的注释说明。对于引用的部分，需要给出参考文献。
9. **输入文件数据和详细输入数据格式**，输出结果文件数据片段和详细输出数据格式（必须清晰描述）
10. 程序运行实验结果说明和分析
11. 总结：特点总结，功能、性能、扩展性等方面存在的不足和可能的改进之处
12. 参考文献

- 带注释的源程序 (**必须提交源程序以备检查实现情况，无源程序的以未完成课程设计处理**）
- 输入数据文件和运行结果文件（**必须提交输入输出文件数据，数据量太大可取部分数据**）
- 执行程序


# 谢 谢！