课程设计 3: 新闻分类

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和机器学习任务，包括数据预处理，机器学习建模、样本预测等。

图 1 提供了新闻数据集的基本目录结构，其中子目录的名称和文件名称的前缀就表示当前新闻文本的类别。提供的数据中包含训练集和测试集，需要使用训练集训练模型并使用测试集报告精度。

图 1: 数据集目录结构

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

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。不同类别的新闻往往反映不同的主题和特征。本课程设计的任务是通过 MapReduce 技术和机器学习算法实现新闻文本的自动分类。具体包含如下若干子任务，这些子任务组合起来就构成了一个完整的新闻文本分类流程。

任务 1：文本特征选择

本任务的主要工作是在分词后对原始新闻的文本中包含的词语进行特征选择，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

图 2 为从新闻文本中生成的特征词的示例。

任务 2：关键词提取

本任务要求在任务 1 的基础上，为每个新闻类别提取 20 个最能够代表该类别的关键词。

输入输出

输入： 1. 新闻文本训练数据；2. 任务 1 得到的输出；
输出： 每个新闻类别的关键词

任务 3：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。注意，本任务不是必须的，文本的特征表示形式视选取的分类算法而定，或许不需要将文本表示为特定向量形式。

图 2: 特征词

输入输出

输入： 1. 任务 1 中的输出；2. 新闻文本数据

输出： 每条新闻文本的特征向量

文本特征的一个示例如图 3 所示。

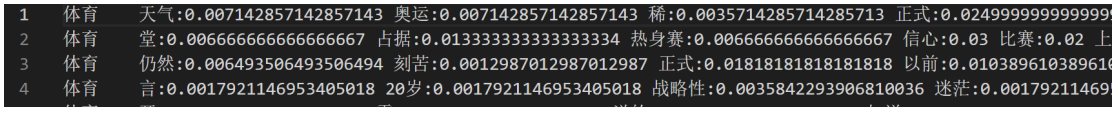

图 3: 文本特征

任务 4：文本分类

得到了每个新闻文本的特征向量之后，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。在得到的分类模型之后，接下来使用该模型对测试数据中的新闻文本进行预测，输出预测结果，并统计预测的正确率。

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。