选做实验

一、小组信息：

2016st19：张瑞MF1533072、董茂健MG1533007、杨杰MG1533075、金熠波131220168

二、实验环境

Ubuntu 12.04.2、Java 1.8.0\_102、Maven 3.3.9、IntelliJ 15.06

三、项目构建

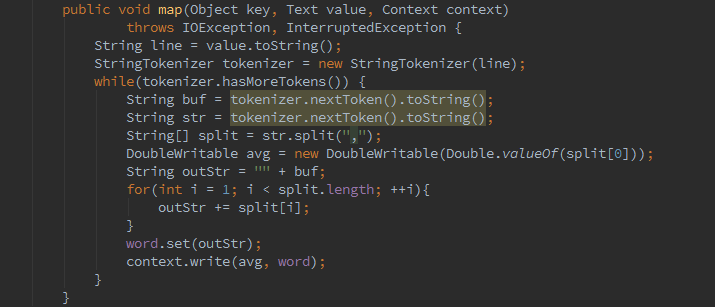
项目均采用Maven构建。由于是在倒排索引的基础上完成的选作实验，所以只列出不同的实现部分。

1. 具体实现

1. 平均词频的全排序：

这里采用的是，在原有生成词频文件的基础上只进行Map。由于Map结束后会自动进行排序，所以能够达到效果。

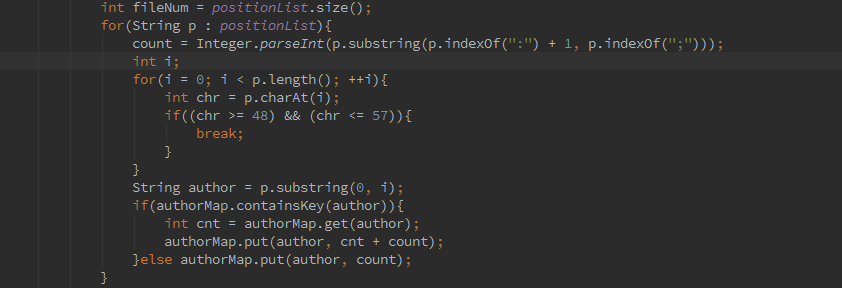
将词频作为Key（DoubleWritable）即可。



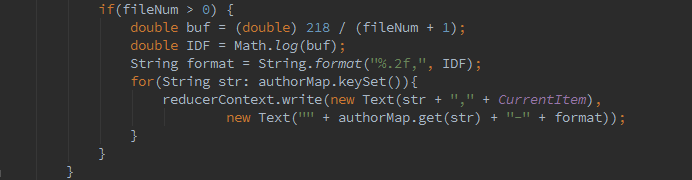
1. TF-IDF

这里采取在词频的MapReduce任务上进行修改。由于一个单词后的倒排索引包含所有作者，所以只需把其中的作者部分提取出来即可。由于需要统计同一个作者的所有作品，所以这里采用java.util.map集合。对同一个单词，每次判断是否已经处理过同一个作者的其他作品。如是，则更新map中的文档数；否则向map插入一条该作者的信息。随后将所有作者和这条单词结合输出。

提取和处理作者：



生成一条条键值对（Text, Text），第一个Text包含作者和单词信息，后一个为具体TF-IDF的值：

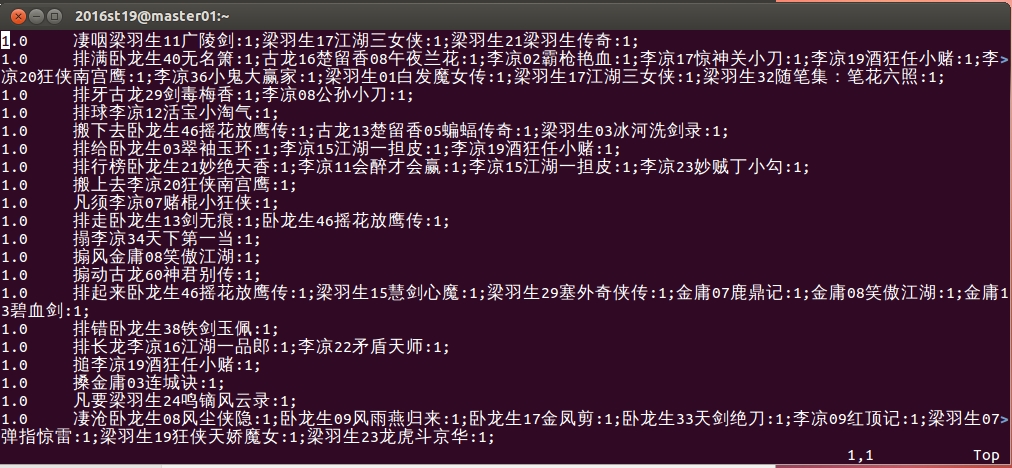


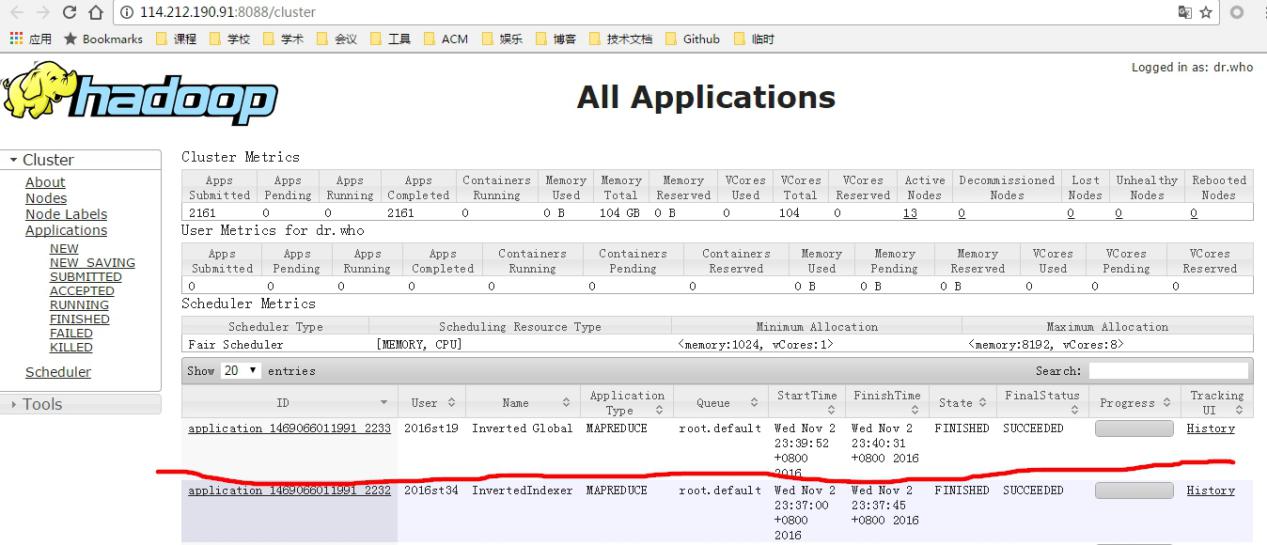
1. 输出效果
   1. 全局词频排序

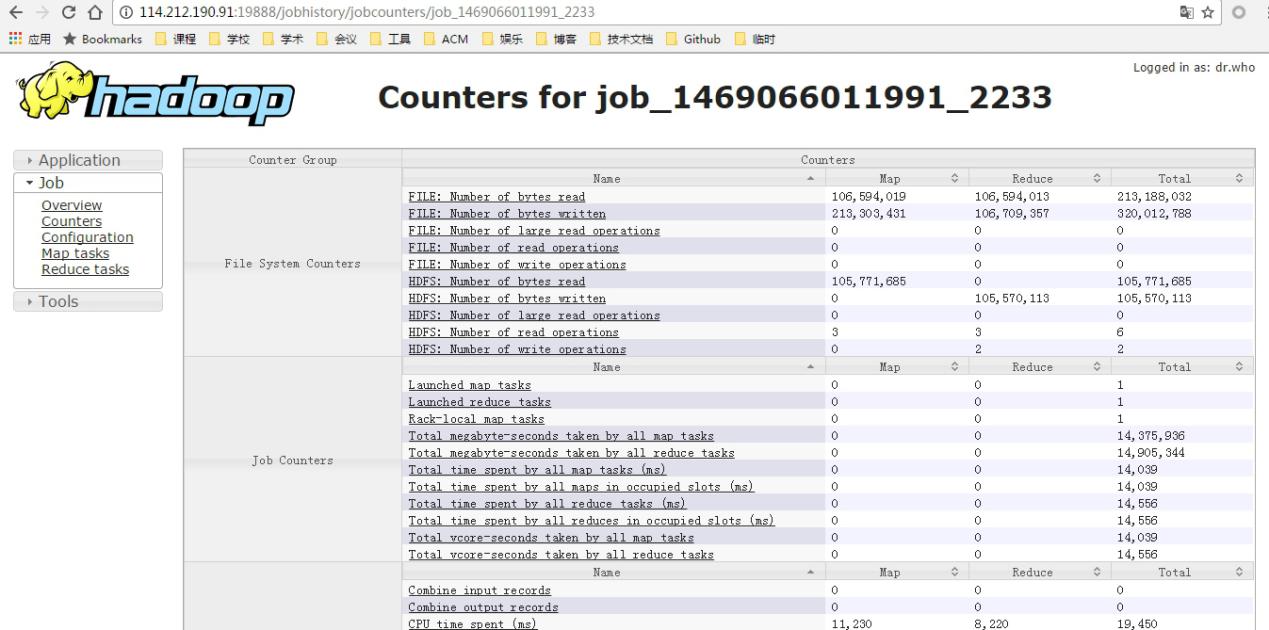
此处的输入为HDFS上之前实验的输出

生成的文件在/user/2016st19/InvertedGlobal中

screenshot_20161102_234439



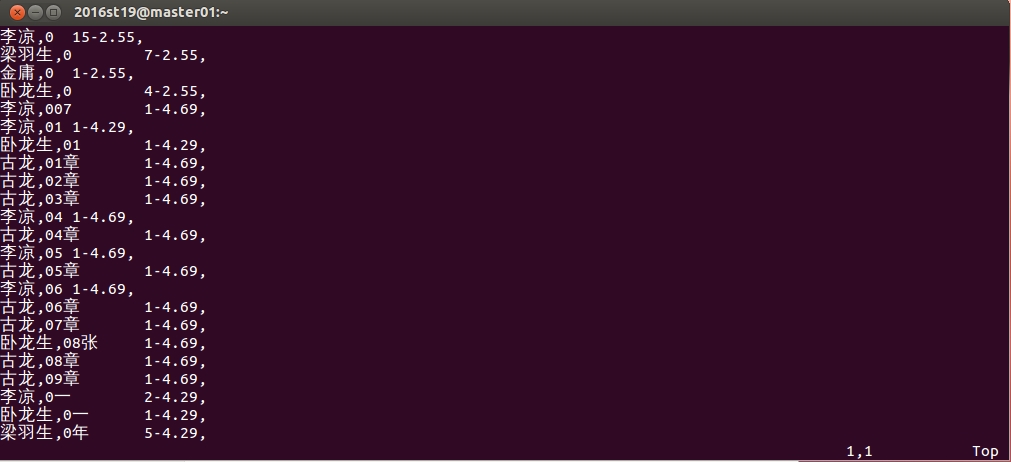


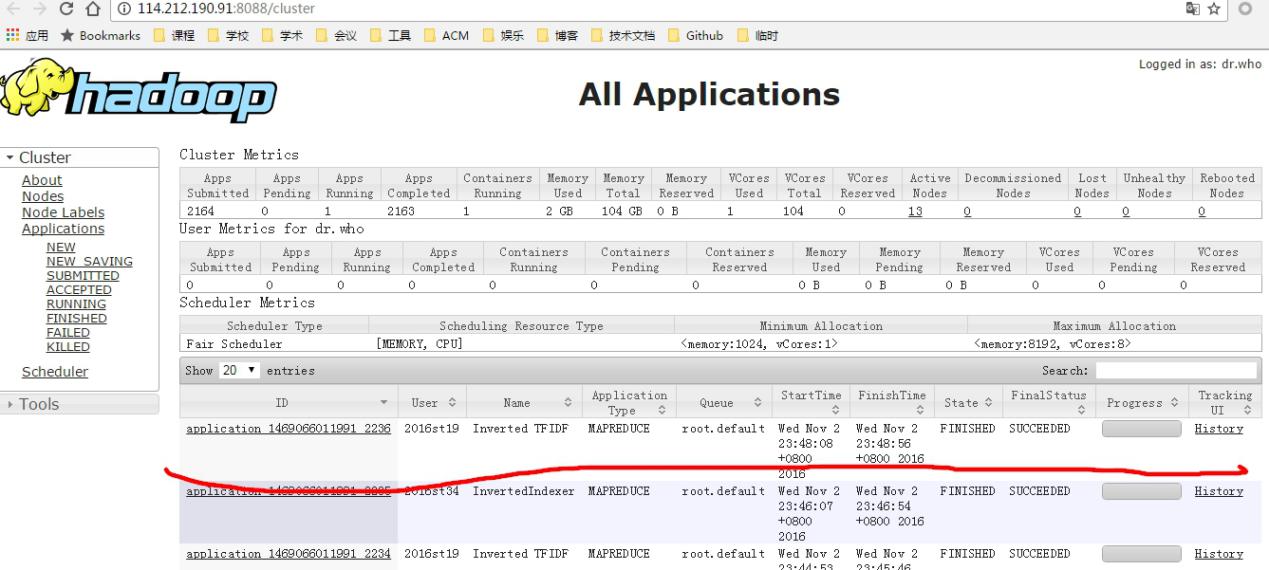


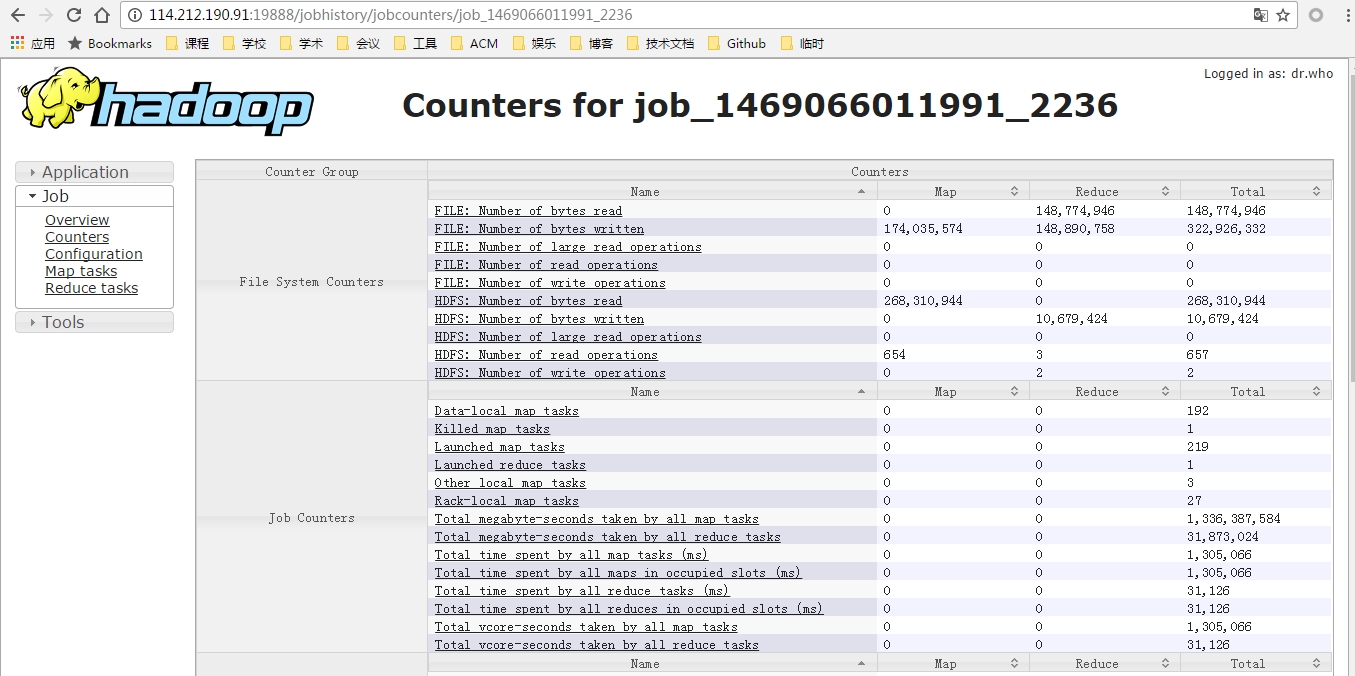
* 1. TF-IDF

生成文件在/user/2016st19/InvertedTFIDF下

screenshot_20161102_235011







1. 附件及说明

随压缩包附有实验报告、本选做实验补充报告、三个项目的工程文件即对应的jar包