倒排索引实验报告

一、小组信息：

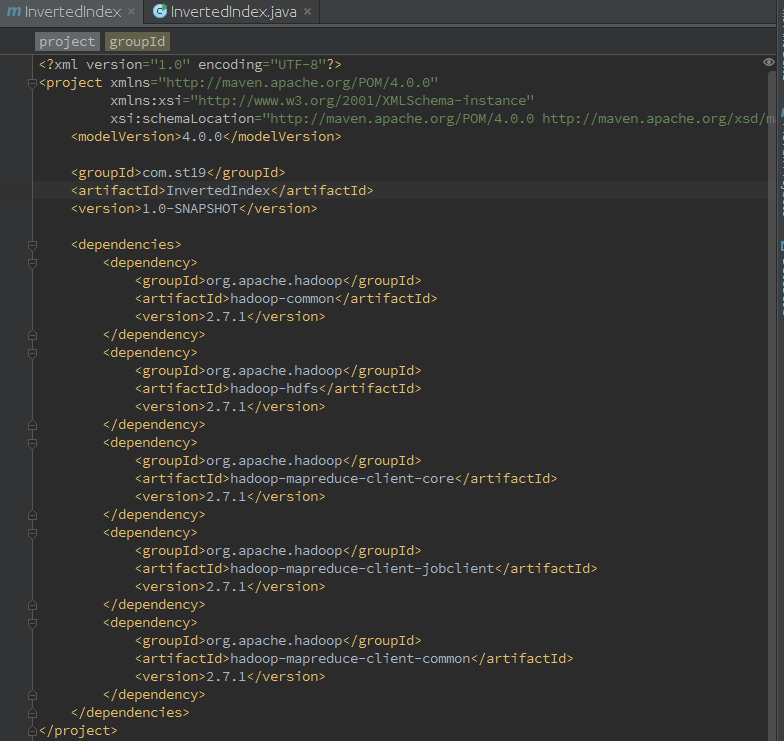
2016st19：张瑞MF1533072、董茂健MG1533007、杨杰MG1533075、金熠波131220168

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

本项目采用Maven构建：



四、Map、Reduce设计思路及实现

Map设计思路：

输入：<key(Object), value(Text), context(Context)>

key值无含义，value为文本信息，context包含文件

处理：遇到一个词组即生成一组键值对

输出：<某词组+所在文件名(Text), 1(IntWritable)>

Reduce设计思路：

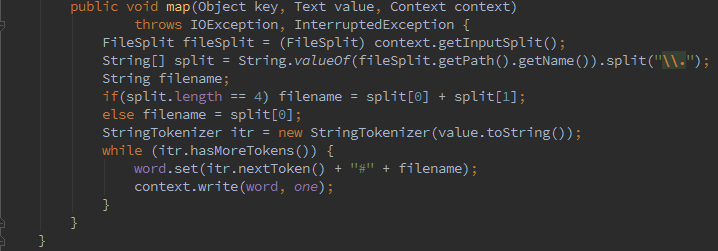
输入：<某单词+所在文件名(Text), 在该文件的出现次数(IntWritable)>

处理：将同一词组的所有词频信息整合，算出平均出现次数

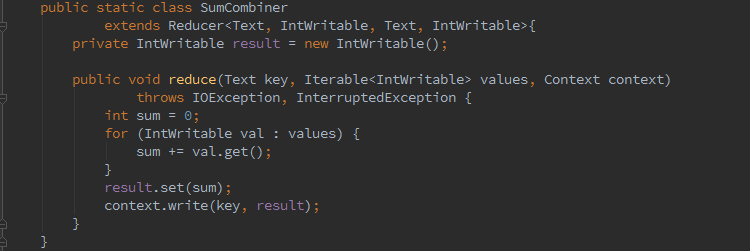
输出：<某词组(Text), 平均出现次数+词频信息(Text)>

具体实现细节：

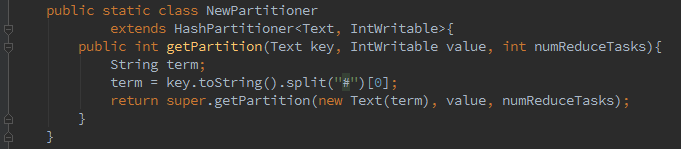
Map：遇到一个词组就生成一组键值对，在key中包含文件名信息，这里对文件名做了处理，去掉后缀.txt.segmented，生成的键值对key中包含词组和文件名信息。这里由于有不同命名格式的文件，所以加判断。



Combiner：将一个节点所处理的所有map中词组key一致的键值对做整合。由于Hadoop会对map键值对做本地排序，所以对于同一个key，Iterable的values就是所有相同key的value列表，只需统计value值的个数即可。生成新的一个键值对，value为key相同的键值对的个数。

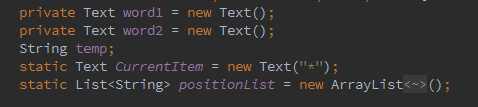


Partition：由于需要将同一个词组的所有键值对交由一个reducer处理，所以重写partitioner，让key中词组一样的键值对分配到一个reducer。

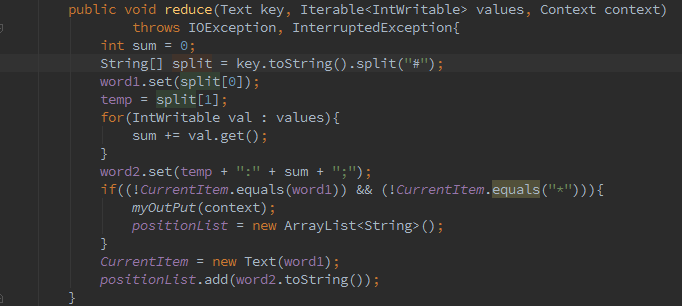


Reduce：这里的Reduce稍有些复杂。首先需要将同一个词组的不同键值对挑选出来。因为key不仅包含词组信息，更包含文件名信息，所以在做处理时先要剔除文件名。其次，相同的词组只生成一个键值对，所以需要找出所有相同词组的所有键值对，因此需要有一个数据列表用于存储当前处理的相同词组的词频信息（需要根据文件名和在给文件中出现的次数手动生成）。当发现当前处理的词组不是之前列表中处理的词组时（一个新的词组产生意味着之前的词组已经处理完毕），即对之前的词组生成新的键值对输出。由于最后一个词组到文件结束都没有被处理（等待下一个不一样的词组），所以需要额外一次输出处理。

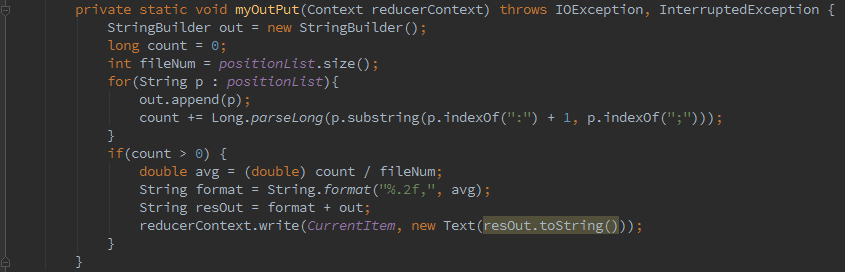
1. 用于记录处理当前词组和词频缓存列表的数据结构：



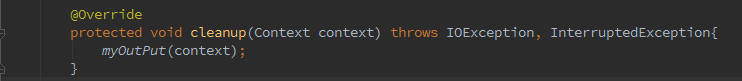
1. reduce函数：不断检查当前到来的词组，如该词组已经出现（和CurrentItem一致）则将其词频信息加在列表尾，否则将上一个词组输出，CurrentItem为当前词组。CurrentItem初始值为\*。



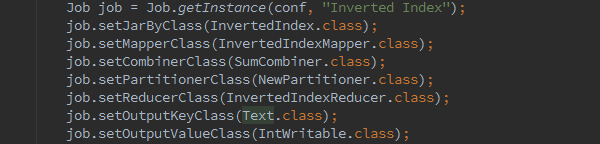
1. 自定义的reduce键值对生成函数



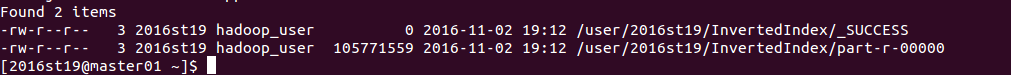
1. 最后一个词组的特殊处理，reduce结束自动调用

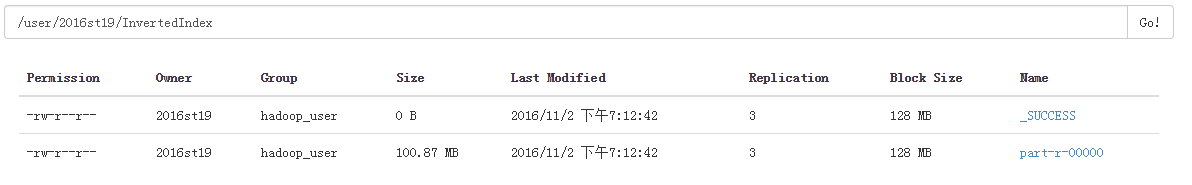


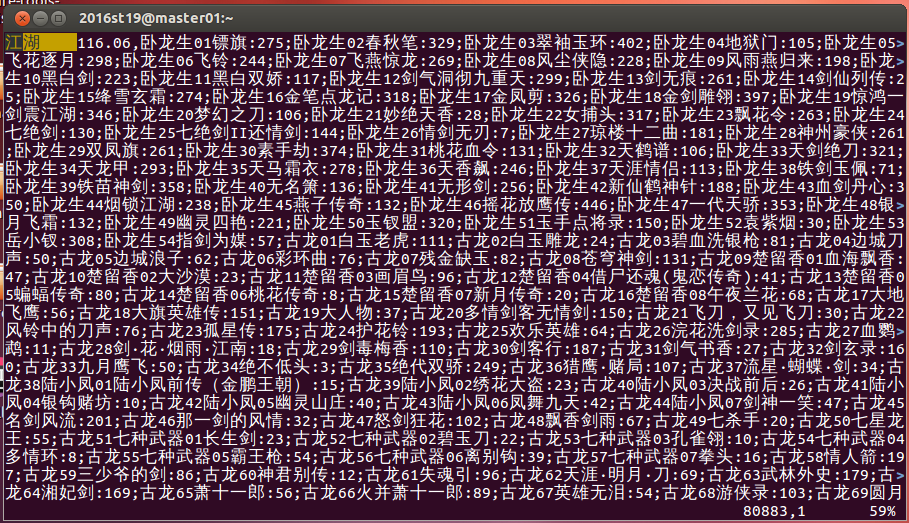
1. Main函数及配置信息

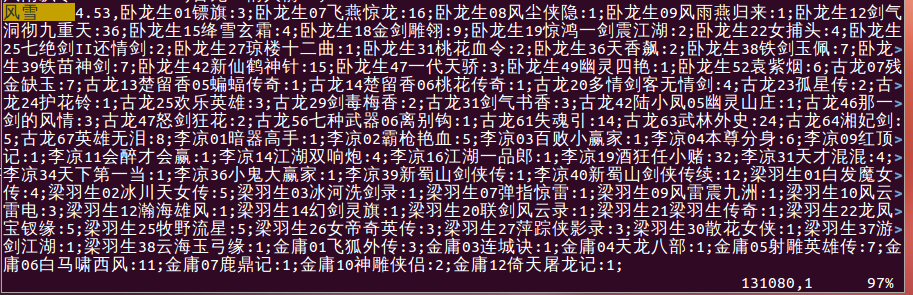


五、输出文件及结果

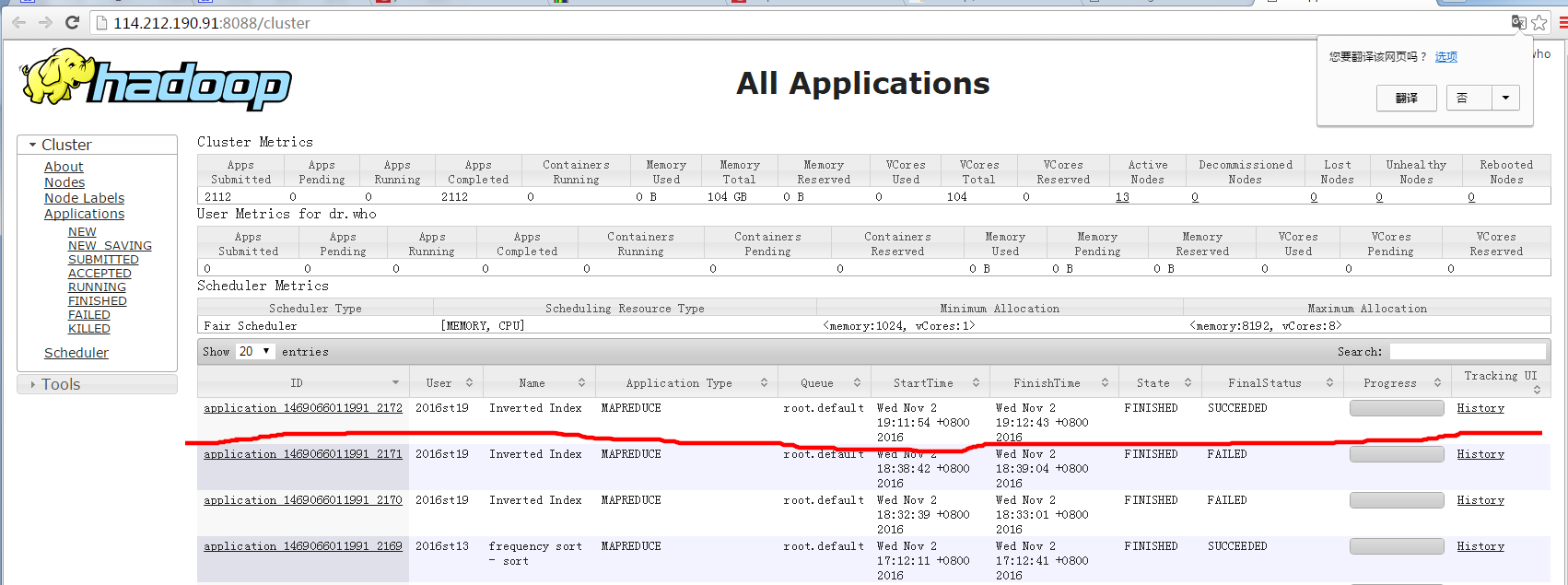








六、集群Web信息





七、附件信息

注InvertedIndex.jar使用命令：

1. 上传jar包：



1. 执行：



附件包含本实验报告、选做实验报告、InvertedIndex.jar和对应工程（本实验）

InvertedGlobal.jar和对应工程、InvertedTFIDF.jar和对应工程（选做实验）

所有工程项目均无out生成文件夹（已删，太大）