计算机科学与技术学院 大数据管理与分析 课程实验报告

|  |  |  |
| --- | --- | --- |
| 实验题目： 文档倒排索引算法实现 | | 学号：201705130113 |
| 日期：2020.4.22 | 班级：2017级计算机3班 | 姓名：黄瑞哲 |
| Email：jeshrz063@gmail.com | | |
| 实验目的：   1. 编程实现倒排索引 2. 熟练掌握MapReduce程序在集群上提交与执行的过程。 3. 加深对MapReduce编程框架的理解 | | |
| 实验软件和硬件环境：  操作系统：macOS Catalina 10.15.4  Hadoop版本：2.8.5  JDK版本：1.8.0\_241  Java IDE: IntelliJ IDEA | | |
| 实验原理和方法：   1. Map过程   使用默认的TextInputFormat类对文件输入进行处理，对于Map函数输入的每一个<key, value>，key是偏移量，value则是读进来的每一行文本。然后对value进行split操作，只保留单词与数字（统称为单词）。然后对于每一个单词到停用词表里去查找，如果停用词表里没有这个词，那么把 文件名#单词作为key，词频1作为value输出。   1. Combiner过程   Combiner其实是本地的一次Reduce过程。为了在最终的reduce函数中能够将一个单词全部聚集在一起，可以在map函数出口的时候进行一次预处理，就是我们这个Combiner完成的任务。对于map函数的输出<key,{value}>，首先遍历values可以统计出这个文件中某一个单词出现的次数，然后把key按照#分割，前者是文件名，后者是单词。最后把单独把单词作为key，文件名#出现次数作为value输出。   1. Reduce过程   每一个reduce函数会从所有的combiner输出中把同一单词的value聚集起来，得到的<key, {value}>。这里key就是单词，values里的每一个元素便是这个单词在哪个文件里出现过，以及出现了多少次。遍历一下values便可以知道这个单词在所有文档里出现的次数。最后按照输出要求按字典序从小到大排个序就好了。 | | |
| 实验步骤：（不要求罗列完整源代码）   1. 停用词表以及文件名的获取   在启动Map函数之前，系统会调用一次setup函数，可以在这个函数中获得我们想要的信息。  停用词表：首先在配置MapReduce的时候可以先把停用词文件传递给系统    然后就可以根据context在setup函数里拿到这个文件路径，读取这个文件放到一个集合里即可。    文件名也可以从context里获取，不过直接获取的是输入文件的完整路径，在这个实验里只需要知道输入文件的文件名就好，那么对原始输出的路径按/分割，取分割后的最后一个元素即可。     1. Map函数   首先按照非单词将输入的value分隔开，对于每一个非空单词先全部转换为小写，然后到停用词表里查找一下，如果没有查找到就将文件名#单词作为key，词频1作为value输出即可。     1. Combiner函数   Combiner的输入values是map函数输出的一次本地聚集，遍历一下就可以知道这个词出现的次数。再把key按照#分割一下，前者是文件名，后者是单词，把单词作为key，文件名#次数作为value输出。     1. Reducer函数   这里主要是对values的处理，首先遍历一下获得这个单词在所有文档里出现的次数，然后把每个value按照字典序从小到大排个序输出，最后把<total, sum>这个字段输出即可。 | | |
| 结论分析与体会：  MapReduce的部分输出如下      然后与数据集的标准答案对比一下    可以看到与标准答案没有差异，输出是正确的。 | | |
| 就实验过程中遇到和出现的问题，你是如何解决和处理的，自拟1－3道问答题：  对单词进行分割的时候，单纯用空格分割会出现很多标点符号，正确的分割方法是什么？  答：查了知乎，可以发现Java的split函数传递的参数可以试试正则表达式，查到\W表示非单词字符，把它作为split参数传递就是就好了，要注意把\转移一下，也就是说实际写的是\\W | | |