实验二 文档倒排索引算法实现

一、实验目的

倒排索引（Inverted Index）被用来存储在全文搜索下某个单词在一个文档或者一组文档中的存储位置的映射，是目前几乎所有支持全文索引的搜索引擎都需要依赖的一个数据结构。通过对倒排索引的编程实现，熟练掌握 MapReduce 程序在集群上的提交与执行过程，加深对 MapReduce 编程框架的理解。

二、实验平台

1）操作系统：Linux（实验室版本为 Ubuntu17.04，集群环境为 centos6.5）；

2）Hadoop 版本：2.9.0；

3）JDK 版本：1.8；

4）Java IDE：Eclipse 3.8。

三、实验内容

1）在本地编写程序和调试

在本地 eclipse 上编写带词频属性的对英文文档的文档倒排索引程序，要求程序能够实现对 stop-words(如 a,an,the,in,of 等词)的去除，能够统计单词在每篇文档中出现的频率。可自行准备文档数据和停词表，也可从下述集群的 hdfs 文件系统上下载，在伪分布式环境下完成程序的编写和调试。

2）在集群上提交作业并执行

集群的服务器地址为 10.102.0.197，用户名和密码为自己的学号，用户主目录为/home/用户名，hdfs 目录为/user/用户名。集群上的实验文档存放目录为 hdfs://master:9000/txt_input/. 英文停词表文件存放位置为 hdfs://master:9000/stop_words/stop_words_eng.txt ，具体步骤如下：

(1)使用 scp InvertedIndex.jar 用户名@10.102.0.197:/home/用户名 命令将本地程序提交到 Hadoop 集群,通过 ssh 用户名@10.102.0.197 命令远程登录到 Hadoop 集群进行操作；

(2)使用 hadoop jar InvertedIndex.jar /txt_input output 命令在集群上运行 Hadoop 作业，指定输出目录为自己 hdfs 目录下的 output；

(3)在浏览器中打开 http://10.102.0.197:50070,可以查看集群的基本信息以及 hdfs 目录；在浏览器中打开 http://10.102.0.197:8088，可以查看集群上作业的基本执行情况。

四、实验要求

实验结果的输出类似如下格式，单词以及单词对应列表里的文档名按照字母顺序排列：

tonight <MACBETH.txt,10>;<OTHELLO.txt,24>;<total,34>. took <MACBETH.txt,2>;<OTHELLO.txt,4>;<total,6>. tooth <MACBETH.txt,2>;<OTHELLO.txt,1>;<total,3>.

使用 hdfs dfs -get 命令将自己的运行结果与 hdfs 上/output 目录的标准结果 part-r-00000 保存到本地，使用 diff 命令判断差异，结果正确无误，方可提交。

五、实验报告

计算机科学与技术学院 大数据管理与分析 课程实报告

实验题目：		学号：201500000000
日期：2018.3.20	班级：2015 级 1 班/菁英班	姓名：张三
Email：zhangsan@qq.com		
实验目的：		
实验软件和硬件环境：		
实验原理和方法：		
实验步骤：（不要求罗列完整源代码）		
结论分析与体会：		

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：