

哈爾濱工業大學

实验报告

题 目：大数据高级数据结构设计与实践作业三

专 业 大数据科学与技术

学 号 1170300916

姓 名 彭钰驯

课 程 大数据高级数据结构设计与实践

日 期 2020-5-3

一、实验要求

任务一：

- 1) 修改样例程序 wordcount，通过 reduce 过程计算词频最高的 5 个单词。
- 2) 输出格式为： Top5 word is %s %s %s %s %s.

任务二：

- 1) 输入文件为 friends.txt
- 2) 数据的格式以 “:” 分割成两部分，前面是用户，后面是该用户的粉丝，以 A:B, C, D, F, E, O 为例，B, C, D, E, F, O 是用户 A 的粉丝

二、实验环境

系统环境：Windows10

IDE：Visual Studio

三、人员安排

一人完成

四、实验过程

4.1 词频 topk 设计思路

对于词频统计 topk 利用 mapreduce 算法进行计算，首先需要进行词频统计(wordcount)，然后将 wordcount 的统计结果输出到文件之中。其中每一行两个元素，分别是单词(word)以及出现的次数(num)。之后运行另一个 mapreduce 程序，读取 wordcount 生成的文件。接下来分为两个阶段：

map 阶段：利用 java 自带的二叉搜索树，在 map 的过程中，将数据构造成大小小于 K 的树。即每读入一个单词，将其与其出现次数插入到树中，在每次 map 后判断树的大小和 K 的大小，当树的数据量大于 K 时，取出最小的数。在 map 方法结束后会执行 cleanup 方法，该方法将 map 任务中的前 K 个数据传入 reduce 任务中。

reduce 阶段：在 reduce 阶段中，依次将 map 方法中传入的 K 个数据放入 java 自带的二叉搜索树中，并依靠平衡特性来维持数据的有序性。从而将 K 个数据利用二叉搜索树的 firstKey 方法按从大到小或者利用二叉搜索树的 lastKey 方法按从小到大的顺序排列。从而求出前 K 个数。

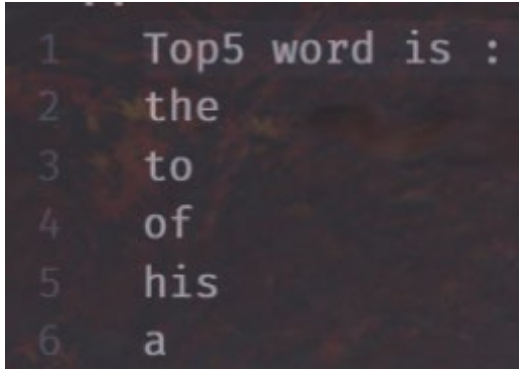
4.2 共同粉丝设计思路

同样需要两边 mapreduce，首先求得某一个人是哪些人的粉丝，比如 B 是 A,E,F,J 的粉丝。这是第一步需要求的结果。第二步进行两两配对，即 A，E 的共同粉丝有 B。A，F

的共同粉丝有 B。然后在 reduce 阶段进行合并。

五、实验分析

任务一：



```
1 Top5 word is :
2 the
3 to
4 of
5 his
6 a
```

任务二：

首先求出某一人是那些人的粉丝，然后将该文件作为计算共同粉丝的输入文件，得到所有人的共同粉丝。文件放在另一个文档（friend.txt）中。