

大数据实验2—倒排索引 实验报告

陈越琦

(121160005 Yueqichen.0x0@gmail.com)

刘威

(131220085 liuwei13cs@smail.nju.edu.cn)

杨杰才

(131220115 mark_grove@qq.com)

周子博

(121250229 441842096@qq.com)

摘要: 本次实验我们小组通过课堂上介绍的“带词频属性的文档倒排算法”，统计了词语的倒排索引并输出，按照要求输出了词语的平均出现次数，使用给定的小说数据集在集群上玩出了调试与测试。在此基础上，完成了两个选做任务：词语的全局排序与计算出每位作家每个词语 * 的TF-IDF并输出。实验分别在本地进行了测试并提交到集群运行获得运行结果。

关键词: Hadoop、倒排索引、全局排序、TF-IDF

§1. 引言

倒排索引是文档检索系统中最常用的数据结构，被广泛的应用于全文搜索引擎。它主要用来存储某个单词（或词组），在一个文档或一组文档中的存储位置的映射，即提供了一种根据内容来查找文档的方式，由于不是根据文档来确定文档所包含的内容，而是进行了相反的操作，因而被称为倒排索引。在本次实验中，我们实现了带词频属性的文档倒排算法。除了实验任务外，我们还设计了两个MapReduce Job来进行选作内容的设计，分别进行以下两个工作：(1)对每个词语的平均出现次数进行全局排序，输出排序后的结果；(2)为每位作家、计算每个词语的TF-IDF。

§2. 实验环境与概述

本次实验的本地开发与测试环境如下：

*陈越琦: 121160005 Yueqichen.0x0@gmail.com 刘威: 131220085 liuwei13cs@smail.nju.edu.cn
杨杰才: 121160005 mark_grove@qq.com 周子博: 121250229 441842096@qq.com

软件	版本
OS	Ubuntu-15.04
JDK	JDK 7u80
Hadoop	0.20.205.0 & 2.7.1

本次实验，我们完成了三个MapReduce任务的编写：(1)带词频属性的文档倒排算法。 (2)对每个词语的平均出现次数进行全局排序。 (3)为每位作家、计算每个词语的TF-IDF。同时，我们也初步熟悉了集群的使用与在集群上执行任务的方法。

下面几节我将详述实验的设计、测试与运行的结果。

§3. 实验具体设计

3.1 带词频属性的文档倒排算法

3.1.1 总体设计

为完成InvertedIndex任务我们共设计了五个类： InvertedIndex, InvertedIndexMapper, InvertedIndexReducer, NewPartitioner, SumCombiner.

InvertedIndex是主类，负责启动配置作业，提交作业与获得完成结果的过程。

NewPartitioner类是自定义的partitioner。由于对键值对进行shuffle处理传送给合适的Reducer时，将按照Map所得的新的键(词语, 文档id)进行排序和选择Reducer，因而同一个term的键值对可能被分到不同Reducer，因而需要对partitioner进行自定义，使同一个term的键值对分到相同的Reducer，避免实验结果产生问题。

InvertedIndexMapper, InvertedIndexReducer, SumCombiner分别为Map, Reduce与Combine类。

下面我将叙述Map与Reduce的设计思路：

1. Map过程： Map过程首先分析输入的(key,value)对，即文档id与具体文档，使用java的字符串处理函数得到索引中需要的信息：词语，文档id。得到key： (词语#文档id)(使用#便于后面的切割) 与 value (即词语出现次数)： 1。
2. Combine过程： 经过map方法处理后， Combine过程将key值相同的value值累加，得到一个词语在文档中的词频。
3. Reduce过程： 经过上述的两个过程后， Reduce过程将相同的词语的value值与词频属性组合成倒排引索文件的格式。

Map,Reduce各阶段的K,V类型见下表：

		Map	Combiner	Reduce
输入	key	—	词语+文档id	combiner的结果key
	value	—	词频	单个文档中出现的词频汇总
输出	key	词语+文档id	词语+文档id	词语
	value	词频 (1)	单个文档中出现的词频汇总	文档id+词频

3.1.2 Map与Reduce的伪代码

Map:

```

Class Mapper
procedure Map(doc_id j, doc d)
F ← new AssociativeArray
for all term m t in doc d do
    F{t} ← F{t} + 1
for all term t ∈ F do
    Emit(pair<t,j>, frequency F{t})

```

Reduce:

```

class Reducer
method Setup // 初始化
    tprev ← Ø;
    P ← new PostingsList
method Reduce(tuple <t,n>, tf [f])
    if t ≠ tprev ^ tprev ≠ Ø then
        Emit(tprev, P)
        P.Reset()
    P.Add(<n,f>)
    tprev ← t
method Close
    Emit(t, P)

```

3.2 对词语出现次数进行全局排序

3.2.1 任务设计思路

对词语出现次数进行全局排序是基于上面带词频属性的倒排索引的输出结果来操作的，将上面的输出按照平均出现次数由小到大进行排列。为完成这个任务，我们设计了ResultSort与ResultSortMapper两个类。该任务中，Mapper使用ResultSortMapper类，Partitioner与Reducer使用默认类。

本任务的Map与Reduce设计如下：

1. Map过程：从输入数据中使用字符串相关处理函数拆分出每个词语的平均出现次数作为key，原本的记录作为value。
2. Reduce过程：将reduce task数目设为1，此时只有一个Reducer，也就是只有一个Partitioner，那么所有Map的输出都会经过一个Partitioner到一个Reducer里，在一个Reducer里会自动根据key的值即平均出现次数来排序，从而实现排序的目标，此时不会对每个键值对做任何具体的操作。

此时Map,Reduce各阶段的K,V类型见下表：

		Map	Reduce
输入	key	——	平均出现次数
	value	每一条倒排索引	对应的倒排索引
输出	key	平均出现次数	平均出现次数
	value	对应的倒排索引	对应的倒排索引

3.2.2 Map的伪代码

Map的伪代码如下：

```

Class Mapper
procedure Map(doc_id j, doc d)
F ← new AssociativeArray
for all term m t in doc d do
    F{t} ← F{t} + 1
for all term t ∈ F do
    Emit(pair<t,j>, frequency F{t})

```

3.3 为每位作家、计算每个词语的TF-IDF

3.3.1 任务设计思路

TBD.

3.3.2 Map与Reduce伪代码

TBD.

§4. 实验测试与运行结果

4.1 JAR包执行方式说明

1. 首先使用scp指令将jar包传送到服务器。
2. 使用ssh登录到服务器
3. 使用hadoop jar InvertedIndex.jar /data/wuxia_novels InvertedIndexoutput指令执行InvertedIndex.jar
3. 使用hadoop jar ResultSort.jar InvertedIndexoutput指令以上面的执行结果为输入数据来执行ResultSort.jar

4.2 实验运行结果展示

我们的程序均首先在本地做了简单的测试后提交到集群进行运行。

在集群上运行的结果如下：

- 带词频属性的文档倒排算法InvertedIndex任务的运行结果：该任务的输出结果在HDFS上的存放路径为：hdfs://master01:9000/user/2016st21/Lab2/InvertedIndexoutput .
该任务在集群上对全部数据集运行的结果的部分截图如下（输出格式为[词语] TAB 平均出现次数， 小说1:词频； 小说2:词频； 小说3:词频； ...； 小说N:词频）：

```

0      1.6875, 卧龙生07飞燕惊龙:1; 卧龙生37天涯情侣:1; 卧龙生42新仙鹤神针:1; 卧龙生45燕子传奇:1; 李凉07赌棍小狂侠:5; 李凉15江湖一担皮:1; 李凉21六宝江湖行:1; 李凉23妙贼丁小勾:3; 李凉27奇神扬小邪:1; 李凉38笑哭江湖:4; 梁羽生01白发魔女传:1; 梁羽生11广陵剑:1; 梁羽生12瀚海雄风:1; 梁羽生25牧野流星:3; 梁羽生34武当一剑:1; 金庸07鹿鼎记:1;
007    1.0, 李凉12活宝小淘气:1;
01     1.0, 卧龙生45燕子传奇:1; 李凉23妙贼丁小勾:1;
01章   1.0, 古龙60神君别传:1;
02章   1.0, 古龙60神君别传:1;
03章   1.0, 古龙60神君别传:1;
04     1.0, 李凉26奇神扬小邪续集:1;
04章   1.0, 古龙60神君别传:1;
05     1.0, 李凉23妙贼丁小勾:1;
05章   1.0, 古龙60神君别传:1;
06     1.0, 李凉23妙贼丁小勾:1;
06章   1.0, 古龙60神君别传:1;
07章   1.0, 古龙60神君别传:1;
08张   1.0, 卧龙生47一代天骄:1;
08章   1.0, 古龙60神君别传:1;
09章   1.0, 古龙60神君别传:1;
0—     1.5, 卧龙生46摇花放鹰传:1; 李凉34天下第一当:2;
0年     2.5, 梁羽生01白发魔女传:4; 梁羽生33随笔集：三剑楼随笔:1;
@      1.0, 古龙60神君别传:1;

1      3.3114754098360657, 卧龙生03翠袖玉环:1; 卧龙生04地狱门:1; 卧龙生16金笔点龙记:2; 卧龙生24七绝剑:2; 卧龙生32天鹅谱:1; 卧龙生33天剑绝刀:1; 卧龙生42新仙鹤神针:3; 卧龙生47一代天骄:1; 卧龙生50玉钗盟:1; 古龙06彩环曲:4; 古龙18大旗英雄传:1; 古龙19大人物:2; 古龙24护花铃:4; 古龙26浣花洗剑录:1; 古龙30剑客行:1; 古龙39陆小凤02绣花大盗:1; 古龙48飘香剑雨:1; 古龙57七种武器07拳头:1; 古龙64湘妃剑:1; 李凉02霸枪艳血:2; 李凉06超级邪侠:77; 李凉09红顶记:1; 李凉10滑头傻小子:1; 李凉13江湖急救站:2; 李凉14江湖双响炮:2; 李凉15江湖一担皮:8; 李凉16江湖一品郎:1; 李凉20狂侠南宫鹰:2; 李凉21六宝江湖行:6; 李凉22矛盾大师:4; 李凉24妙贼丁小勾续集:1; 李凉25魔手邪怪:1; 李凉26奇神扬小邪续集:3; 李凉29神偷小千:1; 李凉30淘气世家:4; 李凉33天齐大帝:2; 李凉34天下第一当:1; 李凉40新蜀山剑侠传续:1; 梁羽生02冰川天女传:3; 梁羽生03水河洗剑录:3; 梁羽生05草莽龙蛇传:5; 梁羽生06大唐游侠传:1; 梁羽生07弹指惊雷:2; 梁羽生09风雷震九州:1; 梁羽生12瀚海雄风:1; 梁羽生14幻剑灵旗:3; 梁羽生15慧剑心魔:2; 梁羽生16剑网尘丝:1; 梁羽生18绝塞传烽录:1; 梁羽生19狂侠天娇魔女:8; 梁羽生20联剑风云录:1; 梁羽生22龙凤宝钗缘:2; 梁羽生24鸣镝风云录:6; 梁羽生31随笔集：笔不花:1; 梁羽生32随笔集：笔花六照:1; 梁羽生34武当一剑:1; 梁羽生35武林天骄:1; 梁羽生38云海玉弓缘:1; 金庸04天龙八部:2; 金庸05射雕英雄传:1; 金庸07鹿鼎记:4;
1.     1.0, 梁羽生20联剑风云录:1;
10     2.0, 李凉06超级邪侠:1; 李凉26奇神扬小邪续集:2; 梁羽生31随笔集：笔不花:4;
梁羽生38云海玉弓缘:1;
1000万钱   1.0, 梁羽生31随笔集：笔不花:1;
100间   1.0, 梁羽生31随笔集：笔不花:1;
102个   1.0, 梁羽生31随笔集：笔不花:1;
107    1.0, 梁羽生31随笔集：笔不花:1;

```

- 对词语的平均出现次数进行排序的ResultSort任务的运行结果：该任务的输出结果在HDFS上的存放路径为：hdfs://master01:9000/user/2016st21/Lab2/InvertedIndexoutput

该任务在集群上对全部数据集运行的结果的部分截图如下（输出格式为平均出现次数 TAB 原纪录）：

```

x - □ 终端 文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
1.0 因材 1.0, 梁羽生05草莽龙蛇传:1;
1.0 因时制宜 1.0, 卧龙生12剑气洞彻九重天:1; 卧龙生15绎雪玄霜:1; 卧龙>
生36天香飘:1; 卧龙生49幽灵四艳:1; 卧龙生50玉钗盟:1; 梁羽生32随笔集 : 笔花六照:1;
金庸13碧血剑:1;
1.0 因日 1.0, 古龙29剑毒梅香:1;
1.0 因人制宜 1.0, 卧龙生30素手劫:1;
1.0 因循 1.0, 卧龙生09风雨燕归来:1; 卧龙生22女捕头:1; 卧龙生26情剑无刃:1;
卧龙生30素手劫:1; 梁羽生29塞外奇侠传:1; 金庸07鹿鼎记:1;
1.0 因式 1.0, 梁羽生23龙虎斗京华:1;
1.0 因地制宜 1.0, 卧龙生50玉钗盟:1; 李凉07赌棍小狂侠:1; 梁羽生32随笔>
集 : 笔花六照:1;
1.0 因势利导 1.0, 卧龙生07飞燕惊龙:1; 卧龙生12剑气洞彻九重天:1; 卧龙>
生51玉手点将录:1; 梁羽生38云海玉弓缘:1;
1.0 因利乘便 1.0, 梁羽生14幻剑灵旗:1; 梁羽生22龙凤宝钗缘:1; 梁羽生37>
游剑江湖:1;
1.0 因公殉职 1.0, 李凉15江湖一担皮:1; 李凉28忍者龟:1; 金庸07鹿鼎记:1;

1.0 08张 1.0, 卧龙生47一代天骄:1;
1.0 08章 1.0, 古龙60神君别传:1;
1.0 09章 1.0, 古龙60神君别传:1;
1.0 突升 1.0, 卧龙生19惊鸿一剑震江湖:1;
1.0 突刺 1.0, 梁羽生03冰河洗剑录:1; 金庸09书剑恩仇录:1;
1.0 突击队 1.0, 李凉04本尊分身:1;
~/桌面/part-r-00000" 134881L, 112096626C 1,1 顶端

```

```

x - □ 终端 文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
1.0 突出表现 1.0, 梁羽生21梁羽生传奇:1;
1.0 1. 1.0, 梁羽生20联剑风云录:1;
1.0 突减 1.0, 卧龙生15绎雪玄霜:1; 李凉22矛盾天师:1;
1.0 1000万钱 1.0, 梁羽生31随笔集 : 笔不花:1;
1.0 100间 1.0, 梁羽生31随笔集 : 笔不花:1;
1.0 穿鼻 1.0, 李凉20狂侠南宫鹰:1;
1.0 102个 1.0, 梁羽生31随笔集 : 笔不花:1;
1.0 穿透力 1.0, 卧龙生17金凤剪:1; 卧龙生22女捕头:1; 李凉41杨小邪发威:1;
1.0 107 1.0, 梁羽生31随笔集 : 笔不花:1;
1.0 穿过来 1.0, 古龙67英雄无泪:1; 李凉17惊神关小刀:1; 梁羽生16剑网尘丝:1;
1.0 10岁 1.0, 梁羽生31随笔集 : 笔不花:1;
1.0 穿身而过 1.0, 卧龙生01镖旗:1; 卧龙生53岳小钗:1; 李凉11会醉才会赢:>
1; 金庸05射雕英雄传:1; 金庸10神雕侠侣:1;
1.0 10章 1.0, 卧龙生37天涯情侣:1; 古龙60神君别传:1;
1.0 穿衣镜 1.0, 古龙16楚留香08午夜兰花:1;
1.0 11 1.0, 李凉26奇神扬小邪续集:1; 梁羽生31随笔集 : 笔不花:1;
1.0 1100 1.0, 梁羽生31随笔集 : 笔不花:1;
1.0 穿街过巷 1.0, 卧龙生17金凤剪:1; 古龙62天涯·明月·刀:1; 金庸05射雕>
英雄传:1; 金庸07鹿鼎记:1;
1.0 11岁 1.0, 梁羽生31随笔集 : 笔不花:1;
1.0 11章 1.0, 古龙60神君别传:1;
1.0 12 1.0, 李凉26奇神扬小邪续集:1; 梁羽生31随笔集 : 笔不花:1; 梁羽生38>
云海玉弓缘:1;
37,1 0%

```

- “江湖”、“风雪”两个词语的输出结果如下：

江湖 116.06481481481481, 卧龙生 01 横旗 :275; 卧龙生 02 春秋笔 :329; 卧龙生 03 翠袖玉环 :402; 卧龙生 04 地狱门 :105; 卧龙生 05 飞花逐月 :298; 卧龙生 06 飞铃 :244; 卧龙生 07 飞燕惊龙 :269; 卧龙生 08 风尘侠隐 :228; 卧龙生 09 风雨燕归来 :198; 卧龙生 10 黑白剑 :223; 卧龙生 11 黑白双娇 :117; 卧龙生 12 剑气洞彻九重天 :299; 卧龙生 13 剑无痕 :261; 卧龙生 14 剑仙列传 :25; 卧龙生 15 绯雪玄霜 :274; 卧龙生 16 金点龙记 :318; 卧龙生 17 金凤剪 :326; 卧龙生 18 金剑雕翎 :397; 卧龙生 19 惊鸿一剑震江湖 :346; 卧龙生 20 梦幻之刀 :106; 卧龙生 21 妙绝天香 :28; 卧龙生 22 女捕头 :317; 卧龙生 23 飘花令 :263; 卧龙生 24 七绝剑 :130; 卧龙生 25 七绝剑 II 还情剑 :144; 卧龙生 26 情剑无刃 :7; 卧龙生 27 琼楼十二曲 :181; 卧龙生 28 神州豪侠 :261; 卧龙生 29 双凤旗 :261; 卧龙生 30 素手劫 :374; 卧龙生 31 桃花血令 :131; 卧龙生 32 天鹤谱 :106; 卧龙生 33 天剑绝刀 :321; 卧龙生 34 天龙甲 :293; 卧龙生 35 天马霜衣 :278; 卧龙生 36 天香飘 :246; 卧龙生 37 天涯情侣 :113; 卧龙生 38 铁剑玉佩 :71; 卧龙生 39 铁苗神剑 :358; 卧龙生 40 无名箫 :136; 卧龙生 41 无形剑 :256; 卧龙生 42 新仙鹤神针 :188; 卧龙生 43 血剑丹心 :350; 卧龙生 44 烟镇江湖 :238; 卧龙生 45 蔚子传奇 :132; 卧龙生 46 摆花放鹰传 :446; 卧龙生 47 一代天骄 :353; 卧龙生 48 银月飞霜 :132; 卧龙生 49 幽灵四艳 :221; 卧龙生 50 五铁盟 :320; 卧龙生 51 五手点将录 :150; 卧龙生 52 袁紫烟 :30; 卧龙生 53 岳小钗 :308; 卧龙生 54 指剑为媒 :57; 古龙 01 白玉老虎 :111; 古龙 02 白玉雕龙 :24; 古龙 03 碧血洗银枪 :81; 古龙 04 边城刀声 :50; 古龙 05 边城浪子 :62; 古龙 06 环彩曲 :76; 古龙 07 残金缺玉 :82; 古龙 08 苍穹神剑 :131; 古龙 09 楚留香 01 血海飘香 :47; 古龙 10 楚留香 02 大沙漠 :23; 古龙 11 楚留香 03 画眉鸟 :96; 古龙 12 楚留香 04 借尸还魂 (鬼恋传奇) :41; 古龙 13 楚留香 05 蝙蝠传奇 :80; 古龙 14 楚留香 06 桃花传奇 :8; 古龙 15 楚留香 07 新月传奇 :20; > 古龙 16 楚留香 08 午夜兰花 :68; 古龙 17 大地飞鹰 :56; 古龙 18 大旗英雄传 :151; 古龙 19 大人物 :37; 古龙 20 多情剑客无情剑 :150; 古龙 21 飞刀, 又见飞刀 :30; 古龙 22 风铃中的刀声 :76; 古龙 23 孤星传 :175; 古龙 24 护花铃 :193; 古龙 25 欢乐英雄 :64; 古龙 26 洗花洗剑录 :285; 古龙 27 血鹦鹉 :11; 古龙 28 刀剑·花·烟雨·江南 :18; 古龙 29 剑毒梅香 :110; 古龙 30 剑客行 :187; 古龙 31 剑气书香 :27; 古龙 32 剑玄录 :160; 古龙 33 九月鹰飞 :50; 古龙 34 绝不低头 :3; 古龙 35 绝代双骄 :249; 古龙 36 猎鹰·赌局 :107; 古龙 37 流星·蝴蝶·剑 :34; 古龙 38 陆小凤 01 陆小凤前传 (金鹏王朝) :15; 古龙 39 陆小凤 02 绣花大盗 :23; 古龙 40 陆小凤 03 决战前后 :26; 古龙 41 陆小凤 04 银钩赌坊 :10; 古龙 42 陆小凤 05 幽灵山庄 :40; 古龙 43 陆小凤 06 凤舞九天 :42; 古龙 44 陆小凤 07 剑神一笑 :47; 古龙 45 八卦剑风流 :201; 古龙 46 那一剑的风情 :32; 古龙 47 怒剑狂花 :102; 古龙 48 飘香剑雨 :67; 古龙 49 七杀手 :20; 古龙 50 七星龙王 :55; 古龙 51 七种武器 01 长生剑 :23; 古龙 52 七种武器 02 碧玉刀 :22; 古龙 53 七种武器 03 孔雀翎 :10; 古龙 54 七种武器 04 多情环 :8; 古龙 55 七种武器 05 霸王枪 :54; 古龙 56 七种武器 06 离别钩 :39; 古龙 57 七种武器 07 拳头 :16; 古龙 58 情人箭 :197; 古龙 59 三少爷的剑 :86; 古龙 60 神君别传 :12; 古龙 61 失魂引 :96; 古龙 62 天涯·明月·刀 :69; 古龙 63 武林外史 :179; 古龙 64 淑妃剑 :169; 古龙 65 萧十一郎 :56; 古龙 66 火并萧十一郎 :89; 古龙 67 英雄无泪 :54; 古龙 68 游侠录 :103; 古龙 69 圆月弯刀 :155; 古龙 70 异星邪 :102; 李凉 01 暗器高手 :53; 李凉 02 霸枪艳血 :34; 李凉 03 百败小赢家 :79; 李凉 04 本尊分身 :97; 李凉 05 超霸的男人 :7; 李凉 06 超级邪侠 :209; 李凉 07 赌棍小狂侠 :31; 李凉 08 公孙小刀 :69; 李凉 09 红顶记 :48; 李凉 10 滑头傻小子 :43; 李凉 11 会醉才会赢 :189; 李凉 12 活宝小淘气 :121; 李凉 13 江湖急救站 :118; 李凉 14 江湖双响炮 :291; 李凉 15 江湖一担皮 :197; 李凉 16 江湖一品郎 :186; 李凉 17 惊神关小刀 :38; 李凉 18 酒赌小浪子 :17; 李凉 19 酒狂任小赌 :167; 李凉 20 狂侠南宫鹰 :15; 李凉 21 六宝江湖行 :72; 李凉 22 矛盾天师 :59; 李凉 23 炙赋丁小勾 :16; 李凉 24 炙赋丁小勾续集 :13; 李凉 25 魔手邪怪 :6; 李凉 26 奇神扬小邪续集 :98; 李凉 27 奇神杨小邪 :102; 李凉 28 忍者龟 :77; 李凉 29 神偷小千 :17; 李凉 30 淘气世家 :81; 李凉 31 才天混混 :89; 李凉 32 才天混混外集 :27; 李凉 33 天齐大帝 :62; 李凉 34 天下第一当 :52; 李凉 35 武林暗游 :180; 李凉 36 小鬼大赢家 :29; 李凉 37 小鱼吃大鱼 :42; 李凉 38 笑笑江湖 :69; 李凉 39 新蜀山剑侠传 :20; 李凉 40 新蜀山剑侠传续 :1; 李凉 41 杨小邪 :发威 :34; 梁羽生 01 白发魔女传 :63; 梁羽生 02 冰川天女传 :5; 梁羽生 03 水河洗剑录 :65; 梁羽生 04 水魄寒光剑 :4; 梁羽生 05 草莽龙蛇传 :143; 梁羽生 06 大唐游侠传 :97; 梁羽生 07 弹指惊雷 :77; 梁羽生 08 飞凤潜龙 :4; 梁羽生 09 风雷震九洲 :135; 梁羽生 10 风云雷电 :139; 梁羽生 11 广陵剑 :149; 梁羽生 12 潘海雄风 :69; 梁羽生 13 还剑奇情录 :8; 梁羽生 14 幻剑灵旗 :27; 梁羽生 15 慧剑心魔 :102; 梁羽生 16 剑网尘丝 :127; 梁羽生 17 江湖三女侠 :252; 梁羽生 18 绝塞传烽录 :40; 梁羽生 19 狂侠天娇魔女 :180; 梁羽生 20 联剑风云录 :126; 梁羽生 21 梁羽生传奇 :74; 梁羽生 22 龙凤宝钗缘 :109; 梁羽生 23 金庸 01 飞传 :3; 金庸 03 连城诀 :1; 金庸 04 天龙八部 :1; 金庸 05 射雕英雄传 :7; 金庸 06 白马啸西风 :11; 金庸 07 鹿鼎记 :1; 金庸 10 神雕侠侣 :2; 金庸 12 倚记 :1;

80883, 4-3 59%

“江湖”的输出结果

风雪 4.5333333333333333, 卧龙生 01 横旗 :3; 卧龙生 07 飞燕惊龙 :16; 卧龙生 08 风尘侠隐 :1; 卧龙生 09 风雨燕归来 :1; 卧龙生 12 剑气洞彻九重天 :36; 卧龙生 15 绯雪玄霜 :4; 卧龙生 18 金剑雕翎 :9; 卧龙生 19 惊鸿一剑震江湖 :2; 卧龙生 22 女捕头 :4; 卧龙生 25 七绝剑 II 还情剑 :2; 卧龙生 27 琼楼十二曲 :1; 卧龙生 31 桃花血令 :2; 卧龙生 36 天香飘 :2; 卧龙生 38 铁剑玉佩 :7; 卧龙生 39 铁苗神剑 :7; 卧龙生 42 新仙鹤神针 :15; 卧龙生 47 一代天骄 :1; 卧龙生 49 幽灵四艳 :1; 卧龙生 52 袁紫烟 :6; 古龙 07 残金缺玉 :7; 古龙 13 楚留香 05 蝙蝠传奇 :1; 古龙 14 楚留香 06 桃花传奇 :1; 古龙 20 多情剑客无情剑 :4; 古龙 23 孤星传 :2; 古龙 24 护花铃 :1; 古龙 25 欢乐英雄 :3; 古龙 29 剑毒梅香 :2; 古龙 31 剑气书香 :3; 古龙 42 陆小凤 05 幽灵山庄 :1; 古龙 46 那一剑的风情 :3; 古龙 47 怒剑狂花 :2; 古龙 56 七种武器 06 离别钩 :1; 古龙 61 失魂引 :14; 古龙 63 武林外史 :24; 古龙 64 淑妃剑 :5; 古龙 67 英雄无泪 :8; 李凉 02 霸枪艳血 :5; 李凉 03 百败小赢家 :1; 李凉 04 本尊分身 :6; 李凉 09 红顶记 :1; 李凉 11 会醉才会赢 :1; 李凉 14 江湖双响炮 :4; 李凉 15 江湖一品郎 :1; 李凉 19 酒狂任小赌 :32; 李凉 31 才天混混 :4; 李凉 34 天下第一当 :1; 李凉 36 小鬼大赢家 :1; 李凉 39 新蜀山剑侠传 :1; 李凉 40 新蜀山剑侠传续 :12; 梁羽生 01 白发魔女传 :4; 梁羽生 02 冰川天女传 :5; 梁羽生 03 水河洗剑录 :1; 梁羽生 07 弹指惊雷 :1; 梁羽生 09 风雷震九洲 :1; 梁羽生 10 风云雷电 :3; 梁羽生 12 潘海雄风 :1; 梁羽生 13 还剑奇情录 :8; 梁羽生 14 幻剑灵旗 :1; 梁羽生 20 联剑风云录 :1; 梁羽生 21 梁羽生传奇 :1; 梁羽生 22 龙凤宝钗缘 :5; 梁羽生 23 金庸 01 飞传 :3; 金庸 03 连城诀 :1; 金庸 04 天龙八部 :1; 金庸 05 射雕英雄传 :7; 金庸 06 白马啸西风 :11; 金庸 07 鹿鼎记 :1; 金庸 10 神雕侠侣 :2; 金庸 12 倚记 :1;

“风雪”的输出结果

本实验在集群上执行MapReduce Job后获得的执行报告如下：

- 在集群All Application (<http://114.212.190.91:8088/>) 的WebUI页面中查看Job的执行状态的截图如下：

All Applications

Logged in as: dr.who

Cluster Metrics																
Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Vcores Used	Vcores Total	Vcores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	
72	0	0	72	0	0 B	112 GB	0 B	0	112	0	14	0	0	0	0	0
Scheduler Metrics																
Scheduler Type: Capacity Scheduler Scheduling Resource Type: [MEMORY] Minimum Allocation: <memory:1024, vCores:1> Maximum Allocation: <memory:8192, vCores:8>																
Show 20 entries Search:																
ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI						
application_1461411805941_0067	2016st21	InvertedIndex Table	MAPREDUCE	default	Tue Apr 26 19:47:45 2016	Tue Apr 26 19:48:33 2016	FINISHED	SUCCEEDED								History
application_1461411805941_0064	2016st21	InvertedIndex Table	MAPREDUCE	default	Tue Apr 26 19:31:50 2016	Tue Apr 26 19:32:40 2016	FINISHED	SUCCEEDED								History
application_1461411805941_0063	2016st21	InvertedIndex Table	MAPREDUCE	default	Tue Apr 26 19:27:53 2016	Tue Apr 26 19:28:57 2016	FINISHED	SUCCEEDED								History
application_1461411805941_0060	2016st21	InvertedIndex Table	MAPREDUCE	default	Tue Apr 26 19:17:36 2016	Tue Apr 26 19:18:26 2016	FINISHED	SUCCEEDED								History
application_1461411805941_0058	2016st21	InvertedIndex Table	MAPREDUCE	default	Tue Apr 26 19:12:35 2016	Tue Apr 26 19:13:17 2016	FINISHED	SUCCEEDED								History
application_1461411805941_0042	2016st21	InvertedIndex Table	MAPREDUCE	default	Tue Apr 26 15:30:46 2016	Tue Apr 26 15:31:32 2016	FINISHED	SUCCEEDED								History
application_1461411805941_0041	2016st21	InvertedIndex Table	MAPREDUCE	default	Tue Apr 26 15:16:18 2016	Tue Apr 26 15:17:08 2016	FINISHED	SUCCEEDED								History
application_1461411805941_0025	2016st21	QuasiMonteCarlo	MAPREDUCE	default	Mon Apr 25 23:58:50 2016	Mon Apr 25 23:59:07 2016	FINISHED	SUCCEEDED								History
application_1461411805941_0070	2016st21	Result Sort	MAPREDUCE	default	Tue Apr 26 19:51:37 2016	Tue Apr 26 19:52:01 2016	FINISHED	SUCCEEDED								History
application_1461411805941_0053	2016st21	InvertedIndex Table	MAPREDUCE	default	Tue Apr 26 18:23:09 2016	Tue Apr 26 18:23:43 2016	FINISHED	FAILED								History

- 在WebUI页面 (<http://114.212.190.91:19888/jobhistory>) 找到对应的job如下

Retired Jobs

Retired Jobs													Search:		
Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed				
2016.04.26 19:17:36 CST	2016.04.26 19:17:41 CST	2016.04.26 19:18:25 CST	job_1461411805941_0067	InvertedIndex Table	2016st21	default	SUCCEEDED	218	218	1	1				
2016.04.26 19:27:53 CST	2016.04.26 19:27:57 CST	2016.04.26 19:28:56 CST	job_1461411805941_0063	InvertedIndex Table	2016st21	default	SUCCEEDED	218	218	1	1				
2016.04.26 19:31:50 CST	2016.04.26 19:31:55 CST	2016.04.26 19:32:40 CST	job_1461411805941_0064	InvertedIndex Table	2016st21	default	SUCCEEDED	218	218	1	1				
2016.04.26 19:47:45 CST	2016.04.26 19:47:49 CST	2016.04.26 19:48:33 CST	job_1461411805941_0060	InvertedIndex Table	2016st21	default	SUCCEEDED	218	218	1	1				
2016.04.26 19:51:37 CST	2016.04.26 19:51:41 CST	2016.04.26 19:52:01 CST	job_1461411805941_0070	Result Sort	2016st21	default	SUCCEEDED	1	1	1	1				

- 根据Job ID链接进入Job详细页面，几个job的详细信息如下所示。

– InvertedIndexTable:

Logged in as: ur.wm



Counters for job_1461411805941_0067

Counter Group	Name	Map	Reduce	Total
File System Counters	FILE: Number of bytes read	0	148,774,946	148,774,946
	FILE: Number of bytes written	173,963,852	148,890,429	322,854,281
	FILE: Number of large read operations	0	0	0
	FILE: Number of read operations	0	0	0
	FILE: Number of write operations	0	0	0
	HDFS: Number of bytes read	268,312,252	0	268,312,252
	HDFS: Number of bytes written	0	110,648,830	110,648,830
	HDFS: Number of large read operations	0	0	0
	HDFS: Number of read operations	654	3	657
Job Counters	HDFS: Number of write operations	0	2	2
	Data-local map tasks	0	0	215
	Launched map tasks	0	0	218
	Launched reduce tasks	0	0	1
	Rack-local map tasks	0	0	3
	Total megabyte-seconds taken by all map tasks	0	0	1,653,874,688
	Total megabyte-seconds taken by all reduce tasks	0	0	31,096,832
	Total time spent by all map tasks (ms)	0	0	1,615,112
	Total time spent by all maps in occupied slots (ms)	0	0	1,615,112
Map-Reduce Framework	Total time spent by all reduce tasks (ms)	0	0	30,368
	Total time spent by all reduces in occupied slots (ms)	0	0	30,368
	Total vcore-seconds taken by all map tasks	0	0	1,615,112
	Total vcore-seconds taken by all reduce tasks	0	0	30,368
	Name	Map	Reduce	Total
	Combine input records	45,567,096	0	45,567,096
	Combine output records	3,977,262	0	3,977,262
	CPU time spent (ms)	991,980	32,260	1,024,220
	Failed Shuffles	0	0	0
Shuffle Errors	GC time elapsed (ms)	28,291	925	29,216
	Input split bytes	32,349	0	32,349
	Map input records	1,954,746	0	1,954,746
	Map output bytes	1,529,780,369	0	1,529,780,369
	Map output materialized bytes	148,776,242	0	148,776,242
	Map output records	45,567,096	0	45,567,096
	Merged Map outputs	0	218	218
	Physical memory (bytes) snapshot	58,413,002,752	291,491,840	58,704,494,592
	Reduce input groups	0	3,977,262	3,977,262
File Input Format Counters	Reduce input records	0	3,977,262	3,977,262
	Reduce output records	0	134,881	134,881
	Reduce shuffle bytes	0	148,776,242	148,776,242
	Shuffled Maps	0	218	218
	Spilled Records	3,977,262	3,977,262	7,954,524
	Total committed heap usage (bytes)	43,146,805,248	209,190,912	43,355,996,160
File Output Format Counters	Virtual memory (bytes) snapshot	357,850,693,632	1,654,018,048	359,504,711,680
	Name	Map	Reduce	Total
Bytes Read	Bytes Read	268,279,903	0	268,279,903
	Name	Map	Reduce	Total
Bytes Written	Bytes Written	0	110,648,830	110,648,830

– ResultSort



Counters for job_1461411805941_0070

Counter Group	Name	Map	Reduce	Total
File System Counters	FILE: Number of bytes read	112,297,666	112,297,660	224,595,326
	FILE: Number of bytes written	224,710,423	112,412,702	337,123,125
	FILE: Number of large read operations	0	0	0
	FILE: Number of read operations	0	0	0
	FILE: Number of write operations	0	0	0
	HDFS: Number of bytes read	110,648,967	0	110,648,967
	HDFS: Number of bytes written	0	112,096,626	112,096,626
	HDFS: Number of large read operations	0	0	0
	HDFS: Number of read operations	3	3	6
Job Counters	HDFS: Number of write operations	0	2	2
	Launched map tasks	0	0	1
	Launched reduce tasks	0	0	1
	Rack-local map tasks	0	0	1
	Total megabyte-seconds taken by all map tasks	0	0	8,190,976
	Total megabyte-seconds taken by all reduce tasks	0	0	6,990,848
	Total time spent by all map tasks (ms)	0	0	7,999
	Total time spent by all maps in occupied slots (ms)	0	0	7,999
	Total time spent by all reduce tasks (ms)	0	0	6,827
	Total time spent by all reduces in occupied slots (ms)	0	0	6,827
Map-Reduce Framework	Total vcore-seconds taken by all map tasks	0	0	7,999
	Total vcore-seconds taken by all reduce tasks	0	0	6,827
	Combine input records	0	0	0
	Combine output records	0	0	0
	CPU time spent (ms)	7,710	5,860	13,570
	Failed Shuffles	0	0	0
	GC time elapsed (ms)	129	85	214
	Input split bytes	137	0	137
	Map input records	134,881	0	134,881
	Map output bytes	111,877,679	0	111,877,679
	Map output materialized bytes	112,297,660	0	112,297,660
	Map output records	134,881	0	134,881
	Merged Map outputs	0	1	1
	Physical memory (bytes) snapshot	267,386,880	169,914,368	437,301,248
	Reduce input groups	0	19,569	19,569
	Reduce input records	0	134,881	134,881
	Reduce output records	0	134,881	134,881
	Reduce shuffle bytes	0	112,297,660	112,297,660
	Shuffled Maps	0	1	1
Shuffle Errors	Spilled Records	269,762	134,881	404,643
	Total committed heap usage (bytes)	199,229,440	190,316,544	389,545,984
	Virtual memory (bytes) snapshot	1,645,809,664	1,652,441,088	3,298,250,752
	BAD_ID	0	0	0
	CONNECTION	0	0	0
	IO_ERROR	0	0	0
File Input Format Counters	WRONG_LENGTH	0	0	0
	WRONG_MAP	0	0	0
File Output Format Counters	WRONG_REDUCE	0	0	0
	Bytes Read	110,648,830	0	110,648,830
	Bytes Written	0	112,096,626	112,096,626

§5. 实验中遇到的问题与解决方案

§6. 实验总结

本次实验是大数据处理综合实验的第二次实验，在这次实验中我们完成了三个MapReduce任务的编写：(1)带词频属性的文档倒排算法。(2)对每个词语的平均出现次数进行全局排序。(3)为每位作家、计算每个词语的TF-IDF。通过本次实验，我们小组的人都对MapReduce算法设计与编程有了一定的了解，同时进一步加深了对前面所学的Hadoop MapReduce基本框架的理解。

这次实验各个任务中最主要的问题就是Map与Reduce过程的设计与(key,value)键值对类型的确定。在正确确定了键值对的类型转换过程后，整个编程任务的完成就变得容易了。此外，在具体实现时，我们更进一步的了解了java字符串相关的操作函数。灵活使用这些函数，我们也得以方便地从输入文本与文件名中分割出想要的信息，例如作者的姓名与各个文件名。

本次实验是我们第一次深入的了解Hadoop MapReduce的算法设计与程序的编写和运行，整体完成较为顺利，也为后面的实验打下了较好的基础。

参 考 文 献

- [1] 作者. 文章题目[J]. 期刊名称, 年份, 卷号(期数): 起始页码.
- [2] 作者. 书名[M]. 出版地: 出版社, 年份.
- [3] 作者. 章节名[M]// 编者. 书名. 出版地: 出版社, 年份: 起始页码.
- [4] 作者. 文章题目[C]// 编者. 会议论文集名. 出版地: 出版社, 年份: 起始页码.
- [5] Reckdahl K. *Using Import graphics in L^AT_EX2e*[M]. 王磊, 译. [出版地不详]: [出版社不详], 2000.
- [6] 胡伟. L^AT_EX2 ε 完全学习手册[M]. 北京: 清华大学出版社, 2011.
- [7] 李平. L^AT_EX2 ε 及常用宏包使用指南[M], 北京: 清华大学出版社, 2004.
- [8] 桑大勇, 王瑛. 科技文献排版系统: L^AT_EX入门与提高[M]. 武汉: 武汉大学出版社, 2001.
- [9] 陈志杰, 赵书钦, 万福永. L^AT_EX入门与提高[M]. 北京: 高等教育出版社, 2002.
- [10] 邓建松, 彭冉冉, 陈长松. L^AT_EX2 ε 科技排版指南[M]. 北京: 科学出版社, 2001.