

# 文档倒排索引

## 一、实验设计

参考教学 ppt 及书上示例代码，文档倒排索引主要设计以下函数：

### Mapper 函数：

由于此次倒排索引需要携带词频属性，因此在从文件中获取 token 之后，需要添加 filename 来共同作为键值对中的 key。Token 与 filename 之间以#相连，组成的 key 为 text 类型。在 map 阶段，每次生成的 key 其 value 值均为 1，类型为 IntWritable。

```
private static final IntWritable one = new IntWritable(1);
private Text word = new Text();

StringTokenizer tokenizer = new StringTokenizer(value.toString().toLowerCase());
while (tokenizer.hasMoreTokens()) {
    word.set(tokenizer.nextToken() + "#" + fileName);
    context.write(word, one);
}
```

### Combiner 函数：

Mapper 输出的中间结果中会包含大量相同主键的键值对，combiner 将 mapper 输出的中间结果中的词频进行累加，来减少向 reduce 节点中传输的数据量。

```
int sum = 0;
for (IntWritable val : values)
    sum += val.get();
result.set(sum);
context.write(key, result);
```

### Partitioner 函数：

由于我们的 key 是 token#filename 的形式，但我们是想对 token 进行词频统计。为了保证具有相同 token 的键值对可以分配到相同的 reduce 节点，因此需要暂时将 key（token#filename）进行拆分。

```
public int getPartition(Text key, IntWritable value, int numReduceTasks) {
    String term = key.toString().split("#")[0];
    return super.getPartition(new Text(term), value, numReduceTasks);
}
```

### Reducer 函数：

Reducer 从 partitioner 拿到键值对后，需要对相同 token 在不同 file 中的计数进行叠加。因此此时我们需要把原本的 key（word#filename）进行拆分，k-v 由 word#filename - sum 转化为 word - filename: sum。在这部分我们使用 lastword 来记录上一个处理的 token，curword 记录当前处理的 token，维护一个 string 类型的队列 postingList 来保存相同 token 的文件:词频信息(即保存内容为 filename: sum 的字符串)，当某一 token 处

理完成时, list 的大小即为包含此 token 的文件数, 然后将此 list 清空, 然后用来继续保存下一个 token 的信息。

```
String[] keyPair = key.toString().split("#");
curWord.set(keyPair[0]);
String fileName = keyPair[1];
int sum = 0;
for (IntWritable val : values)
    sum += val.get();
if (!lastWord.equals(curWord) && !postingList.isEmpty())//当前 token 与上一个不同, 即上一个 token 已处理完毕
    commitResult(context); //自定义函数, 详见下
postingList.add(fileName + ": " + sum);
totalCount += sum;
lastWord.set(keyPair[0]);

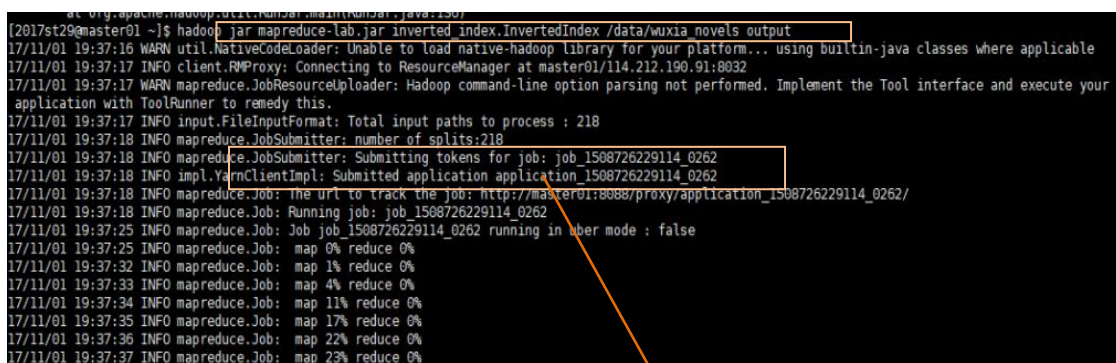
private void commitResult(Context context) //自定义函数
    throws IOException, InterruptedException {
    StringBuilder builder = new StringBuilder();
    builder.append(totalCount / (double) postingList.size()); //平均词频
    builder.append(", ");
    for (String str : postingList) {
        builder.append(str); //写入 filename: sum
        builder.append("; ");
    }
    context.write(lastWord, new Text(builder.toString()));
    totalCount = 0; //清空, 开始计算下一 token
    postingList.clear();
}
```

## 二、 运行截图

集群上执行 InvertedIndex 程序指令:

hadoop jar mapreduce-lab.jar inverted\_index.InvertedIndex /data/wuxia\_novels output

其输出内容前半部分:



```
at org.apache.hadoop.util.NativeCodeLoader.<init> (NativeCodeLoader.java:139)
[2017st29@master01 ~]$ hadoop jar mapreduce-lab.jar inverted_index.InvertedIndex /data/wuxia_novels output
17/11/01 19:37:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/11/01 19:37:17 INFO client.RMProxy: Connecting to ResourceManager at master01/114.212.190.91:8032
17/11/01 19:37:17 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your
application with ToolRunner to remedy this.
17/11/01 19:37:17 INFO input.FileInputFormat: Total input paths to process : 218
17/11/01 19:37:18 INFO mapreduce.JobSubmitter: number of splits:218
17/11/01 19:37:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1508726229114_0262
17/11/01 19:37:18 INFO impl.YarnClientImpl: Submitted application application_1508726229114_0262
17/11/01 19:37:18 INFO mapreduce.Job: The url to track the job: http://master01:8088/proxy/application_1508726229114_0262/
17/11/01 19:37:18 INFO mapreduce.Job: Running job: job_1508726229114_0262
17/11/01 19:37:25 INFO mapreduce.Job: Job job_1508726229114_0262 running in uber mode : false
17/11/01 19:37:25 INFO mapreduce.Job: map 0% reduce 0%
17/11/01 19:37:32 INFO mapreduce.Job: map 1% reduce 0%
17/11/01 19:37:33 INFO mapreduce.Job: map 4% reduce 0%
17/11/01 19:37:34 INFO mapreduce.Job: map 11% reduce 0%
17/11/01 19:37:35 INFO mapreduce.Job: map 17% reduce 0%
17/11/01 19:37:36 INFO mapreduce.Job: map 22% reduce 0%
17/11/01 19:37:37 INFO mapreduce.Job: map 23% reduce 0%
```

```
mapreduce.JobSubmitter: Submitting tokens for job: job_1508726229114_0262
impl.YarnClientImpl: Submitted application application_1508726229114_0262
```

集群中输出文件的位置在\$HOME/lab2 中，  
Hdfs 下的路径即在 inverted-index-output 文件夹下

```
[2017st29@master01 lab2]$ ll
total 246168
-rw-r--r-- 1 2017st29 hadoop_user 118238028 Nov 1 19:56 count-sort-result.txt
-rw-r--r-- 1 2017st29 hadoop_user 8587 Oct 28 22:24 fileList.txt
-rw-r--r-- 1 2017st29 hadoop_user 118238023 Nov 1 20:06 inverted-index-result.txt
-rw-r--r-- 1 2017st29 hadoop_user 19039 Nov 1 19:19 mapreduce-lab.jar
-rw-r--r-- 1 2017st29 hadoop_user 15298113 Nov 1 20:11 tf-idf-result.txt
[2017st29@master01 lab2]$ hdfs dfs -ls
17/11/01 20:24:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Found 4 items
drwxr-xr-x - 2017st29 hadoop_user 0 2017-11-01 19:58 count-sort-output
-rw-r--r-- 3 2017st29 hadoop_user 8587 2017-10-28 22:24 fileList.txt
drwxr-xr-x - 2017st29 hadoop_user 0 2017-11-01 20:08 inverted-index-output
drwxr-xr-x - 2017st29 hadoop_user 0 2017-11-01 20:10 tf-idf-output
[2017st29@master01 lab2]$
```

```
[2017st29@master01 ~]$ hdfs dfs -ls inverted-index-output
17/11/01 19:52:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Found 2 items
-rw-r--r-- 3 2017st29 hadoop_user 0 2017-11-01 19:38 inverted-index-output/_SUCCESS
-rw-r--r-- 3 2017st29 hadoop_user 118238028 2017-11-01 19:38 inverted-index-output/part-r-000000
```

以“江湖”、“风雪”两个单词为例，它们的输出结果为（部分）：

```
[2017st29@master01 lab2]$ grep "\{江湖\|风雪\}" inverted-index-result.txt
江湖 116.06481481481481, 卧龙生01.镖旗: 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;
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.峰雪玄霜: 274; 卧龙生
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
卧龙生34.天龙甲: 293; 卧龙生35.天与福衣: 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;
子传奇: 132; 卧龙生46.梅花放鹰传: 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
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.彩环曲: 76; 古龙07.殃金缺玉: 82; 古龙08.苍穹神
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
193; 古龙25.欢乐英雄: 64; 古龙26.浣花洗剑录: 285; 古龙27.血玲珑: 11; 古龙28.剑·花·烟雨·江南: 18; 古龙29.剑毒梅香:
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
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
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69.圆月弯刀: 155; 古龙70.月异星邪: 102; 李凉01.暗器高手: 59; 李凉02.霸枪艳血: 34; 李凉03.百败小赢家: 79; 李凉04.本
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
李凉23.妙贼丁小勾: 16; 李凉24.妙贼丁小勾续集: 13; 李凉25.魔手邪怪: 6; 李凉26.奇神杨小邪续集: 98; 李凉27.奇神杨小邪:
; 李凉35.武林雄游记: 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: 2
生05.草莽龙蛇传: 143; 梁羽生06.大唐游侠传: 97; 梁羽生07.弹指惊鸿: 77; 梁羽生08.飞凤潜龙: 4; 梁羽生09.风雷震九洲: 13
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: 12
梁羽生27.萍踪侠影录: 40; 梁羽生28.七剑下天山: 77; 梁羽生29.塞外奇侠传: 2; 梁羽生30.散花女侠: 48; 梁羽生31.随笔集: 第
81; 金庸01.飞狐外传: 67; 金庸02.雪山飞狐: 29; 金庸03.连城诀: 35; 金庸04.天龙八部: 126; 金庸05.射雕英雄传: 57; 金庸06.白马
唐14.碧血令: 35;
风雪 4.5333333333333333, 卧龙生01.镖旗: 3; 卧龙生07.飞燕惊龙: 1; 卧龙生08.风尘侠隐: 1; 卧龙生09.风雨燕归来: 1; 卧
7.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.铁臂神剑: 7; 卧龙生42.新仙
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
配剑: 5; 古龙67.英雄无泪: 8; 李凉01.暗器高手: 1; 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
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
; 梁羽生21.梁羽生传奇: 1; 梁羽生22.龙凤宝钗缘: 5; 梁羽生25.牧野流星: 5; 梁羽生26.女帝奇侠传: 3; 梁羽生27.萍踪侠影录
白马啸西风: 11; 金庸07.鹿鼎记: 1; 金庸10.神雕侠侣: 2; 金庸12.倚天屠龙记: 1;
```

```
江湖 116.06481481481481, 卧龙生01.镖旗: 275; 卧龙生02.春秋笔: 329;
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
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
```

```
风雪 4.5333333333333333, 卧龙生01.镖旗: 3; 卧龙生07.飞燕惊龙: 1; 卧龙生08.风尘侠隐: 1; 卧龙生09.风雨燕归来: 1; 卧
7.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.铁臂神剑: 7; 卧龙生42.新仙
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
```

Mapreduce Job 执行截图：

Counters for job_1508726229114_0262									
Counter Group	Name	Map	Reduce	Total					
File System Counters	File: Number of bytes read	0	21,137,473	21,137,473					
	File: Number of bytes written	47,481,960	21,253,635	68,735,595					
	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	268,310,944	0	268,310,944					
	HDFS: Number of bytes written	0	118,238,028	118,238,028					
	HDFS: Number of large read operations	0	0	0					
	HDFS: Number of read operations	654	3	657					
	HDFS: Number of write operations	0	2	2					
Job Counters	Data-local map tasks	0	0	194					
	Killed map tasks	0	0	1					
	Launched map tasks	0	0	219					
	Launched reduce tasks	0	0	1					
	Spilled map tasks	0	0	25					
	Total megabyte-seconds taken by all map tasks	0	0	9,713,262,592					
	Total megabyte-seconds taken by all reduce tasks	0	0	472,055,808					
	Total time spent by all map tasks (ms)	0	0	1,185,701					
	Total time spent by all maps in occupied slots (ms)	0	0	2,371,402					
	Total time spent by all reduce tasks (ms)	0	0	57,624					
Map-Reduce Framework	Total time spent by all reducers in occupied slots (ms)	0	0	115,248					
	Total vcore-seconds taken by all map tasks	0	0	1,185,701					
	Total vcore-seconds taken by all reduce tasks	0	0	57,624					
	Combine input records	45,567,096	0	45,567,096					
	Combine output records	3,976,539	0	3,976,539					
	CPU time spent (ms)	795,179	43,670	838,849					
	Failed shuffles	0	0	0					
	GC time elapsed (ms)	17,449	216	17,665					
	Input split bytes	31,041	0	31,041					
	Map input records	1,954,740	0	1,954,740					
Map-Reduce Framework	Map output bytes	1,571,080,862	0	1,571,080,862					
	Map output materialized bytes	22,145,068	0	22,145,068					
	Map output records	45,567,096	0	45,567,096					
	MergeMap outputs	0	218	218					
	Physical memory (bytes) snapshot	94,506,188,800	873,660,416	95,379,849,216					
	Reduce input groups	0	3,976,539	3,976,539					
	Reduce input records	0	3,976,539	3,976,539					
	Reduce output records	0	134,846	134,846					
	Reduce shuffle bytes	0	22,145,068	22,145,068					
	Shuffled Maps	0	218	218					
Map-Reduce Framework	Spilled Records	3,976,539	3,976,539	7,953,078					
	Total committed heap usage (bytes)	222,472,977,820	1,271,398,400	223,744,376,220					
	Virtual memory (bytes) snapshot	1,038,705,413,840	8,114,761,728	1,046,820,175,568					
	Write failures	0	0	0					
	Yarn: Map task CPU time spent (ms)	0	0	0					
	Yarn: Map task memory (MB)	0	0	0					
	Yarn: Map task virtual memory (MB)	0	0	0					
	Yarn: Reduce task CPU time spent (ms)	0	0	0					
	Yarn: Reduce task memory (MB)	0	0	0					
	Yarn: Reduce task virtual memory (MB)	0	0	0					

## Application application\_1508726229114\_0262

User: 2017st29	
Name: inverted index	
Application Type: MAPREDUCE	
Application Tags:	
YarnApplicationState: FINISHED	
FinalStatus Reported by AM: SUCCEEDED	
Started: Wed Nov 01 19:37:18 +0800 2017	
Elapsed: 1mins, 14sec	
Tracking URL: <a href="#">History</a>	
Diagnostics:	
Total Resource Preempted: <memory:0, vCores:0>	
Total Number of Non-AM Containers Preempted: 0	
Total Number of AM Containers Preempted: 0	
Resource Preempted from Current Attempt: <memory:0, vCores:0>	
Number of Non-AM Containers Preempted from Current Attempt: 0	
Aggregate Resource Allocation: 13061387 MB-seconds, 1495 vcore-seconds	
Started	
Wed Nov 1 19:37:18 +0800 2017	
Node	
<a href="http://slave014:8042">http://slave014:8042</a>	

选做题:

两道选做题均以上述倒排索引算法的输出作为输入文件，代码以一并打包入 `mapreduce-lab.jar` 中。其集群本地输出文件与 `hdfs` 输出路径与倒排算法相同，在上面的路径截图的虚线框中体现。

对每个词语的平均出现次数全局排序算法的执行命令为:

```
hadoop jar mapreduce-lab.jar inverted_index.CountSort inverted-index-output/part-r-00000
output
```

其执行截图如下:



```
[2017st29@master01 ~]$ hadoop jar mapreduce-lab.jar inverted_index.CountSort inverted-index-output/part-r-00000 output
17/11/01 19:53:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
17/11/01 19:53:08 INFO client.RMProxy: Connecting to ResourceManager at master01/114.212.190.91:8032
17/11/01 19:53:09 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface
17/11/01 19:53:09 INFO input.FileInputFormat: Total input paths to process : 1
17/11/01 19:53:10 INFO mapreduce.JobSubmitter: number of splits:1
17/11/01 19:53:10 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1508726229114_0264
17/11/01 19:53:10 INFO impl.YarnClientImpl: Submitted application application_1508726229114_0264
17/11/01 19:53:10 INFO mapreduce.Job: The url to track the job: http://master01:8088/proxy/application_1508726229114_0264/
17/11/01 19:53:10 INFO mapreduce.Job: Running job: job_1508726229114_0264
17/11/01 19:53:17 INFO mapreduce.Job: Job job_1508726229114_0264 running in uber mode : false
17/11/01 19:53:17 INFO mapreduce.Job: map 0% reduce 0%
17/11/01 19:53:27 INFO mapreduce.Job: map 36% reduce 0%
17/11/01 19:53:31 INFO mapreduce.Job: map 56% reduce 0%
```



Counters for job\_1508726229114\_0264

Logged in as: drwho

Category	Counter Group	Name	Map	Reduce	Total
File System Counters	File System Counters	FILE: Number of bytes read	14,484,301	14,398,210	28,882,511
		FILE: Number of bytes written	28,998,540	14,514,192	43,512,732
		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	118,238,162	0	118,238,162
		HDFS: Number of bytes written	0	118,238,028	118,238,028
		HDFS: Number of large read operations	0	0	0
		HDFS: Number of read operations	3	3	6
		HDFS: Number of write operations	0	2	2
Job Counters	Job Counters	Launched map tasks	0	0	1
		Launched reduce tasks	0	0	1
		Back-local map tasks	0	0	1
		Total megabyte-secs taken by all map tasks	0	0	196,722,688
		Total megabyte-secs taken by all reduce tasks	0	0	92,069,888
		Total time spent by all map tasks (ms)	0	0	24,014
		Total time spent by all maps in occupied slots (ms)	0	0	48,028
		Total time spent by all reduce tasks (ms)	0	0	11,239
		Total time spent by all reduces in occupied slots (ms)	0	0	22,478
		Total vcore-seconds taken by all map tasks	0	0	24,014
Map-Reduce Framework	Map-Reduce Framework	Combine input records	0	0	0
		Combine output records	0	0	0
		CPU time spent (ms)	19,840	11,360	31,200
		Failed Shuffles	0	0	0
		GC time elapsed (ms)	93	40	133
		Input split bytes	134	0	134
		Map input records	134,846	0	134,846
		Map output bytes	119,469,991	0	119,469,991
		Map output materialized bytes	14,398,202	0	14,398,202
		Map output records	134,846	0	134,846
Physical memory (bytes) snapshot	Physical memory (bytes) snapshot	Merged Map outputs	0	1	1
		Physical memory (bytes) snapshot	589,283,328	386,162,688	975,446,016

Application application\_1508726229114\_0264

User:	2017st29
Name:	count sort
Application Type:	MAPREDUCE
Application Tags:	
YarnApplicationState:	FINISHED
FinalStatus Reported by AM:	SUCCEEDED
Started:	Wed Nov 01 19:53:10 +0800 2017
Elapsed:	45sec
Tracking URL:	<a href="#">History</a>
Diagnostics:	
Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	531036 MB-seconds, 90 vcore-seconds
Started	Node
Wed Nov 1 19:53:10 +0800 2017	<a href="#">http://slave007:8042</a>

对每位作家，计算每个词语的 TF-IDF 算法的执行语句为：  
hadoop jar mapreduce-lab.jar inverted\_index.Tfidf fileList.txt inverted-index-output/part-r-00000  
output  
其执行截图如下：

```
[2017st29@master01 ~]$ hadoop jar mapreduce-lab.jar inverted index.flidf fileList.txt inverted-index-output/part-r-00000 output
17/11/01 20:09:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/11/01 20:09:28 INFO client.RMProxy: Connecting to ResourceManager at master01/114.212.190.91:8032
17/11/01 20:09:29 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and redefine options to override Hadoop defaults. Command-line options won't be able to override configuration.
17/11/01 20:09:29 INFO input.FileInputFormat: Total input paths to process : 1
17/11/01 20:09:29 INFO mapreduce.JobSubmitter: number of splits:1
17/11/01 20:09:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1508726229114_0266
17/11/01 20:09:30 INFO impl.YarnClientImpl: Submitted application application_1508726229114_0266
17/11/01 20:09:30 INFO mapreduce.Job: The url to track the job: http://master01:8088/proxy/application_1508726229114_0266/
17/11/01 20:09:30 INFO mapreduce.Job: Running job: job_1508726229114_0266
17/11/01 20:09:36 INFO mapreduce.Job: Job job_1508726229114_0266 running in uber mode : false
17/11/01 20:09:36 INFO mapreduce.Job: map 0% reduce 0%
17/11/01 20:09:47 INFO mapreduce.Job: map 23% reduce 0%
17/11/01 20:09:50 INFO mapreduce.Job: map 39% reduce 0%
17/11/01 20:09:53 INFO mapreduce.Job: map 48% reduce 0%
```

hadoop		Counters for job_1508726229114_0266					
action	counter group	name	map	reduce	total		
File System Counters	File System Counters	FILE: Number of bytes read	8,225,265	8,234,772	16,450,037		
		FILE: Number of bytes written	16,567,032	8,341,720	24,908,752		
		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	118,238,157	8,587	118,246,744		
		HDFS: Number of bytes written	0	15,298,113	15,298,113		
		HDFS: Number of large read operations	0	0	0		
		HDFS: Number of read operations	3	4	7		
		HDFS: Number of write operations	0	2	2		
Job Counters	Job Counters	Launched map tasks	0	0	1		
		Launched reduce tasks	0	0	1		
		Back-local map tasks	0	0	1		
		Total mapreduce-seconds taken by all map tasks	0	0	411,230,208		
		Total mapreduce-seconds taken by all reduce tasks	0	0	208,175,104		
		Total time spent by all map tasks (ms)	0	0	50,199		
		Total time spent by all maps in occupied slots (ms)	0	0	100,398		
		Total time spent by all reduce tasks (ms)	0	0	25,412		
		Total time spent by all reducers in occupied slots (ms)	0	0	50,824		
		Total vcore-seconds taken by all map tasks	0	0	50,199		
Map-Reduce Framework	Map-Reduce Framework	Combine input records	0	0	0		
		Combine output records	0	0	0		
		CPU time spent (ms)	54,450	25,760	80,210		
		Failed shuffles	0	0	0		
		GC time elapsed (ms)	288	73	361		
		Input split bytes	134	0	134		
		Map input records	134,846	0	134,846		
		Map output bytes	83,174,486	0	83,174,486		
		Map output materialized bytes	8,224,764	0	8,224,764		
		Map output records	3,976,538	0	3,976,538		
Map-Reduce Framework	Map-Reduce Framework	Merged Map outputs	0	1	1		

## Application application\_1508726229114\_0266

User: 2017st29			
Name: tf-idf			
Application Type: MAPREDUCE			
Application Tags:			
YarnApplicationState: FINISHED			
FinalStatus Reported by AM: SUCCEEDED			
Started: Wed Nov 01 20:09:30 +0800 2017			
Elapsed: 1mins, 25sec			
Tracking URL: <a href="#">History</a>			
Diagnostics:			
Total Resource Preempted: <memory0, vCores0>			
Total Number of Non-AM Containers Preempted: 0			
Total Number of AM Containers Preempted: 0			
Resource Preempted from Current Attempt: <memory0, vCores0>			
Number of Non-AM Containers Preempted from Current Attempt: 0			
Aggregate Resource Allocation: 1026006 MB-seconds, 170 vcore-seconds			
Search:			
Attempt ID	Started	Node	Logs
66_000001	Wed Nov 1 20:09:30 +0800 2017	<a href="http://slave004:8042">http://slave004:8042</a>	<a href="#">Logs</a>