实验 2　倒排索引

1. 实验要求

实验任务

请实现课堂上介绍的“带词频属性的文档倒排算法”。

在统计词语的倒排索引时，除了要输出带词频属性的倒排索引，还请计算每个词语的“平均出现次数”（定义见下）并输出。

“平均出现次数”在这里定义为：

平均出现次数 = 词语在全部文档中出现的频数总和 / 包含该词语的文档数

假如文档集中有四本小说：A、B、C、D。词语“江湖”在文档 A 中出现了 100 次，在文档 B 中出现了 200 次，在文档 C 中出现了 300 次，在文档 D 中没有出现。则词语“江湖”在该文档集中的“平均出现次数”为（100 + 200 + 300) / 3 = 200。

注意 这两个计算任务请在同一个 MapReduce Job 中完成。

输出格式

对于每个词语，输出一个键值对，该键值对的格式如下：

[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

输出中的小说名需要去掉“.txt.segmented”的文件名后缀。

下图展示了输出文件的一个片段（图中内容仅为格式示例）：

```
江湖  98.98, 金庸02雪山飞狐:43; 金庸04天龙八部:55; 金庸07鹿鼎记:123; ...
解药  42, 金庸12倚天屠龙记:41; 金庸15越女剑:45; ...
```

选做内容

该部分内容不做要求，供感兴趣的、学有余力的同学尝试练习。

1.使用另外一个 MapReduce Job 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。TF 定义为某个词语在某个作家的所有作品中的出现次数之和。IDF 定义为：

$$\mathrm{IDF}\left(\text{词语}\right) = \log\left(\frac{\text{语料库文档总数}}{\text{包含该词的文档数} + 1}\right)$$

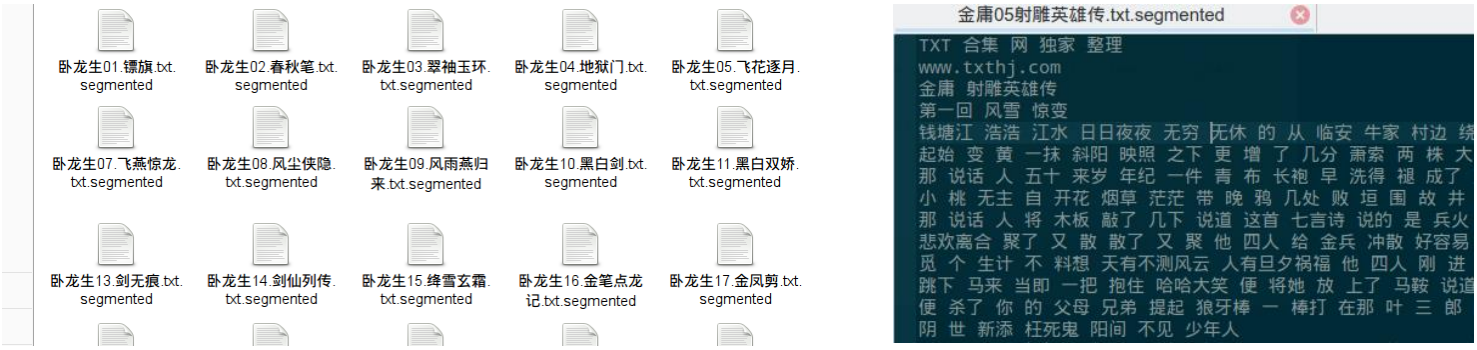
输出格式：作家名字，词语，该词语的 TF-IDF。

2. 实验数据

本次实验提供了金庸、梁羽生等五位小说家的作品全集。每部小说对应一个文本文件。

文本文件均使用 UTF-8 字符编码，并且已分词，两个汉语单词之间使用空格分隔。

输入数据的情况如下图所示：



输入数据文件示例

单机测试样例：提供金庸小说全集作为单机测试样例，请在“实验要求”文件夹下载。

该数据集主要供本地调试使用。

全部数据集：全部数据集位于集群的 HDFS 存储上（ http://114.212.190.91:50070/ ），数据集位置为：/data/wuxia_novels

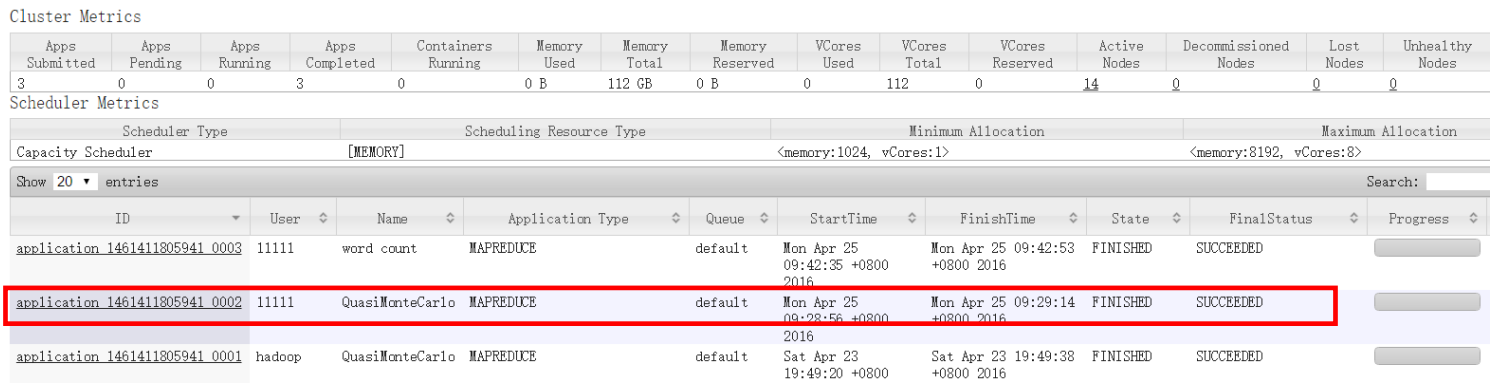
注意 最终每个小组的程序必须在课程指定集群上运行，而且输入数据集是全部数据集。结果输出到集群的 HDFS 上。

3. 实验报告要求

在最后提交的压缩包中，除了包含源代码、JAR 包、JAR 包执行方式说明，还需要包含一个实验报告。

实验报告中请包含：

1. Map 和 Reduce 的设计思路（含 Map、Reduce 阶段的 K、V 类型）。
2. MapReduce 中 Map 和 Reduce 的伪代码（或者带注释的实际代码，如果使用实际代码，请做好排版）。
3. 输出结果文件的部分截图。输出结果文件在 HDFS 上的路径（某些情况下助教会检查 HDFS 上的输出文件）。
4. “江湖”、“风雪”两个单词的输出结果。
5. 在以后的实验报告中，如果需要在集群上执行 MapReduce Job，请在实验报告中附带上相关的 MapReduce Job 的执行报告，以作为评分依据。如果没有执行报告，在评分时将会认为该 MapReduce Job 没有在集群上执行，会影响实验得分。在实验报告中添加 MapReduce Job 执行截图。截图内容如下所示：
   a) 在集群 All Application（ http://114.212.190.91:8088/ ） WebUI 页面中

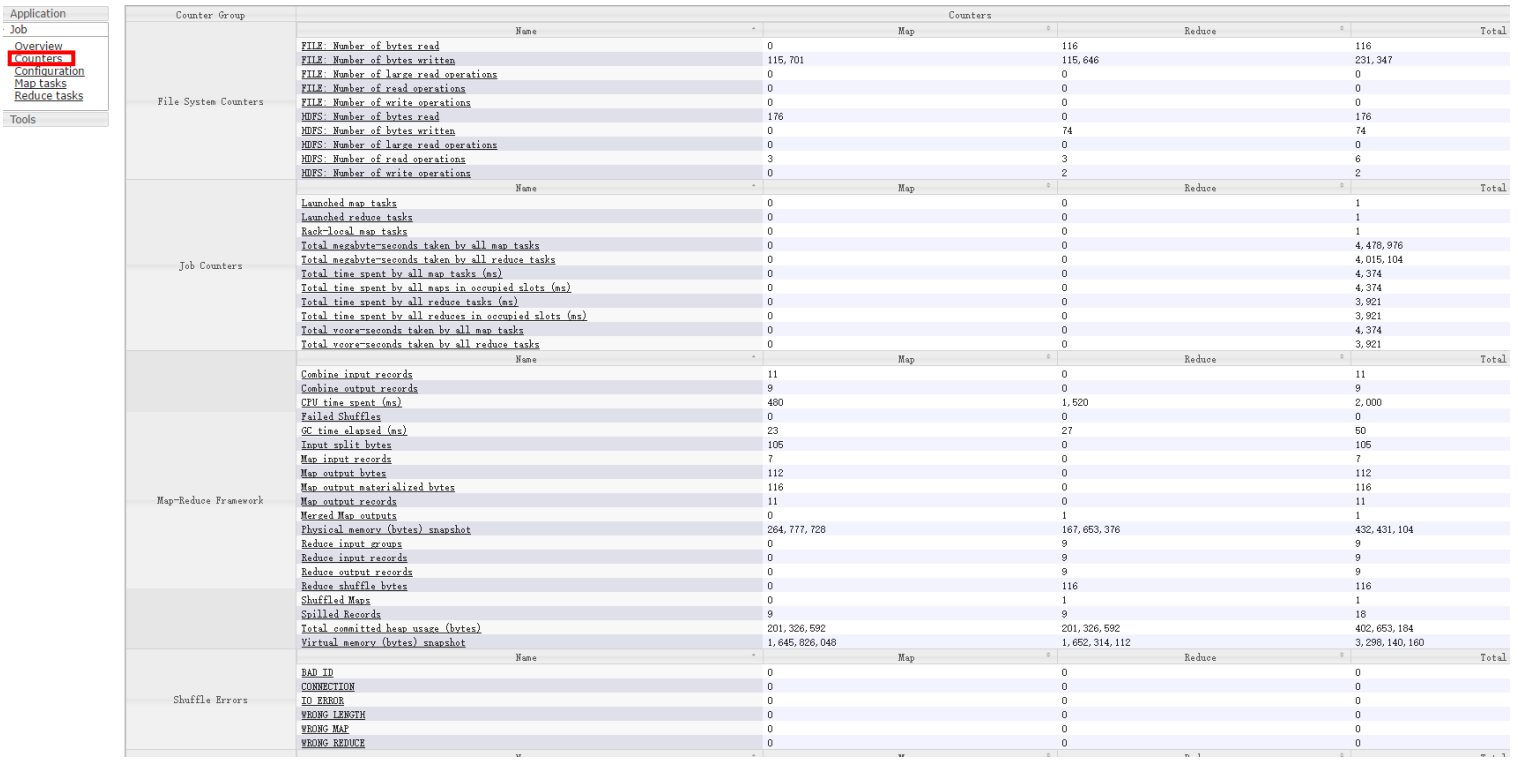


查看 Job 的执行状态。

b) 在 WebUI 页面（ http://114.212.190.91:19888/jobhistory ）

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed
2016.04.25 09:42:35 CST	2016.04.25 09:42:40 CST	2016.04.25 09:42:53 CST	job_1461411805941_0003	word count	11111	default	SUCCEEDED	1	1	1	1

根据 Job ID 链接进入 Job 详细页面，如下所示。



Counter Group	Counters			
	Name	Map	Reduce	Total
File System Counters	FILE: Number of bytes read	0	116	116
	FILE: Number of bytes written	115,701	115,646	231,347
	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	176	0	176
	HDFS: Number of bytes written	0	74	74
	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
	Name	Map	Reduce	Total
Job Counters	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	4,478,976
	Total megabyte-seconds taken by all reduce tasks	0	0	4,015,104
	Total time spent by all map tasks (ms)	0	0	4,374
	Total time spent by all maps in occupied slots (ms)	0	0	4,374
	Total time spent by all reduce tasks (ms)	0	0	3,921
	Total time spent by all reduces in occupied slots (ms)	0	0	3,921
	Total vcore-seconds taken by all map tasks	0	0	4,374
	Total vcore-seconds taken by all reduce tasks	0	0	3,921
	Name	Map	Reduce	Total
Map-Reduce Framework	Combine input records	11	0	11
	Combine output records	9	0	9
	CPU time spent (ms)	480	1,520	2,000
	Failed Shuffles	0	0	0
	GC time elapsed (ms)	23	27	50
	Input split bytes	105	0	105
	Map input records	7	0	7
	Map output bytes	112	0	112
	Map output materialized bytes	116	0	116
	Map output records	11	0	11
	Merged Map outputs	0	1	1
	Physical memory (bytes) snapshot	264,777,728	167,653,376	432,431,104
	Reduce input groups	0	9	9
	Reduce input records	0	9	9
	Reduce output records	0	9	9
	Reduce shuffle bytes	0	116	116
	Shuffled Maps	0	1	1
	Spilled Records	9	9	18
	Total committed heap usage (bytes)	201,326,592	201,326,592	402,653,184
	Virtual memory (bytes) snapshot	1,645,826,048	1,652,314,112	3,298,140,160
	Name	Map	Reduce	Total
Shuffle Errors	BAD_ID	0	0	0
	CONNECTION	0	0	0
	IO_ERROR	0	0	0
	WRONG_LENGTH	0	0	0
	WRONG_MAP	0	0	0
	WRONG_REDUCE	0	0	0