

---

## 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：

$$\text{平均出现次数} = \frac{\text{词语在全部文档中出现的频数}}{\text{包含该词语的文档数}}$$

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。

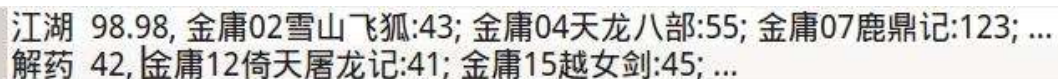
#### 输出格式

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：

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

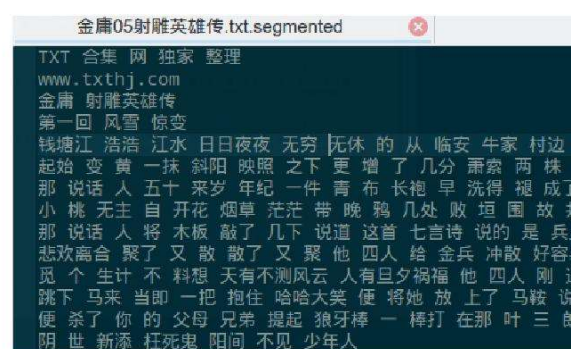
输出格式：作家名字，词语，该词语的 TF-IDF。

## 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：



**单机测试样例：**提供金庸小说全集作为单机测试样例，请在“cslabcms 教学支撑平台上”获取。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“实验集群” /data/exercise\_2目录下，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 Key、Value 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. 请在报告中包含在集群上执行作业后，Yarn Resource Manager 的 WebUI 执行报告内容。请完整包括执行报告内容，否则影响分数。每个 MapReduce Job 对应一个报告）。执行报告内容示例见下文。

## 4. WebUI 执行报告

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。

校园网访问实验平台 **114.212.190.95:8082**

输入小组账户和密码，点击左侧栏“大数据并行计算平台”，再点击“MapReduce 并行计算”可以进入集群监控页面（见下图）。

The screenshot displays the Hadoop YARN All Applications monitoring interface. The top left features the Hadoop logo and a navigation menu with options like 'Cluster', 'About Nodes', 'Node Labels', 'Applications', 'NEW', 'NEW SAVING', 'SUBMITTED', 'ACCEPTED', 'RUNNING', 'FINISHED', 'FAILED', 'KILLED', 'Scheduler', and 'Tools'. The main content area is titled 'All Applications' and shows 'Cluster Metrics' for 'dr.who'. The metrics table includes columns for Apps Submitted, Pending, Running, Completed, Containers Running, Memory Used, Total, Reserved, VCoers Used, Total, Reserved, Active Nodes, Decommissioned Nodes, Lost Nodes, Unhealthy Nodes, and Rebooted Nodes. Below this, 'User Metrics for dr.who' and 'Scheduler Metrics' are also displayed. A table of applications is shown, with columns for ID, User, Name, Application Type, Queue, StartTime, FinishTime, State, FinalStatus, Progress, and Tracking UI. The application 'application\_1540452769420\_0001' is highlighted, and its 'Tracking UI' link is circled in red. The bottom of the page shows 'Showing 1 to 1 of 1 entries' and navigation links 'First Previous 1 Next Last'.

图 1. 集群监控页面

在该页面上，每个 MapReduce Job 都有一项记录，在记录最右侧“Tracking UI”一栏可以访问到该 Job 的执行情况（见上图画圈的位置）。在执行情况页面（见下图）记录的有 Job 的执行时间、执行状态（是否 SUCCEEDED）等信息。

请在实验报告中附上 MapReduce Job 的执行情况页面截屏，以表明该 Job 是在集群上实际执行过的。



图 2. Job 执行情况页面