# 实验2 倒排索引

## 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：

假如文档集中有四个文档：A、B、C、D。词语“同伴”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。

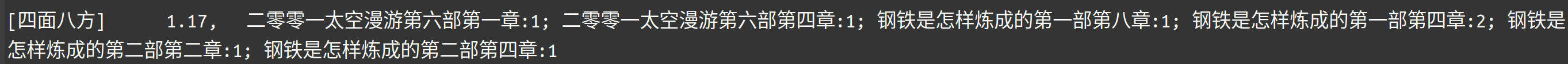
**注意 这两个计算任务请在同一个 MapReduce Job 中完成。**

## 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

下图展示了输出文件的一个片段（图中内容仅为格式示例）



### 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. 为每个作品计算每个词语的TF-IDF。TF定义为某个词语在该作品中的出现次数之和。IDF定义为：

**输出格式：作品名称，词语，该词语的 TF-IDF。由于不同作品的编排结构不一样，可以按照每个文件名的第一个“第”字进行字符串分解，从而获取作品名。**

## 实验数据

本次实验提供了五本外文名著的译文，每部小说的每个章节对应一个文本文件。

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《茶花女》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，HDFS 存储位置为：hdfs://master001:9000/data/classic\_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。**

## 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. “唯一”、“穿着”两个单词的输出结果。

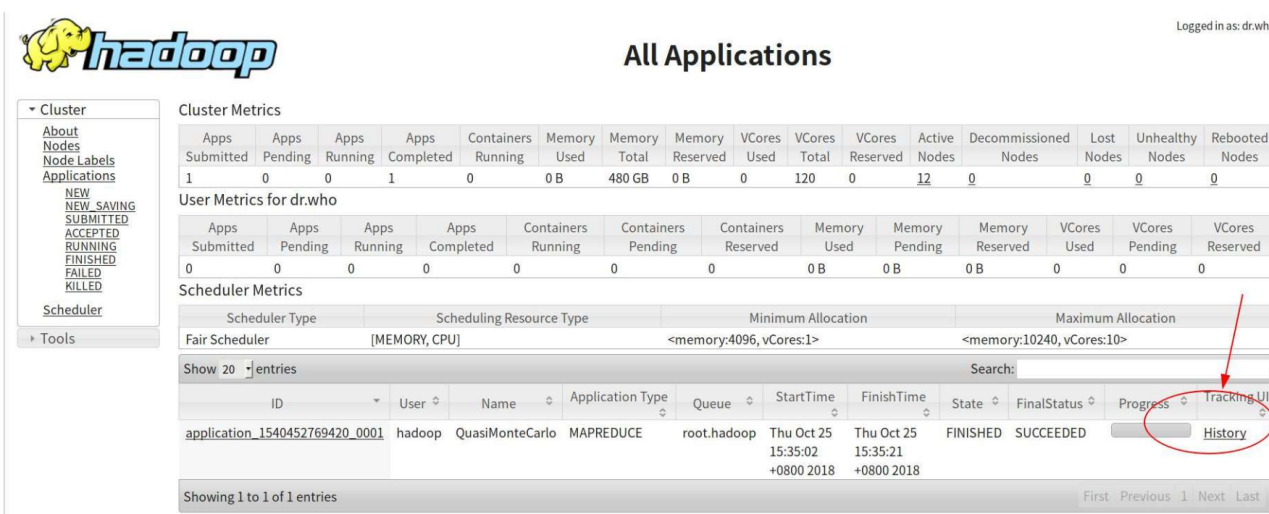
5. 请在报告中包含在集群上执行作业后，Yarn Resource Manager 的 WebUI 执行报告内容。请完整包括执行报告内容，否则影响分数。每个 MapReduce Job 对应一个报告）。执行报告内容示例见下文。

## WebUI执行结果

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。

校园网访问实验平台 114.212.190.95:8082。

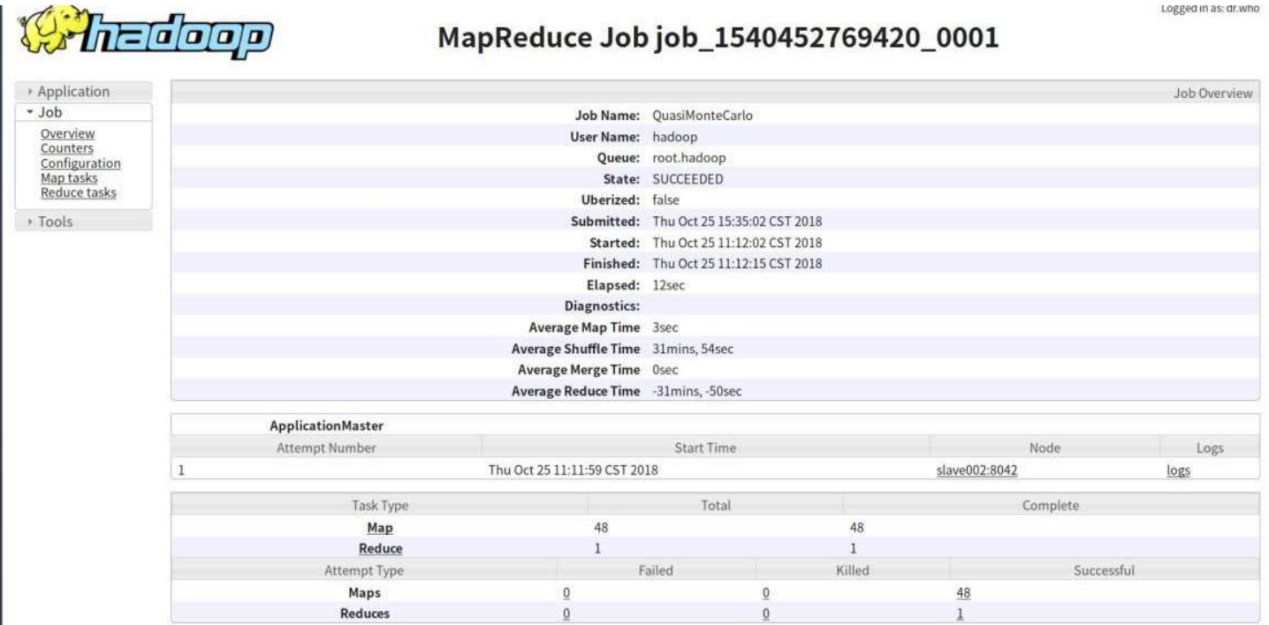
输入小组账户和密码，点击左侧栏“大数据并行计算平台”，再点击“MapReduce并行计算”可以进入集群监控页面（见下图）。



**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**