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。

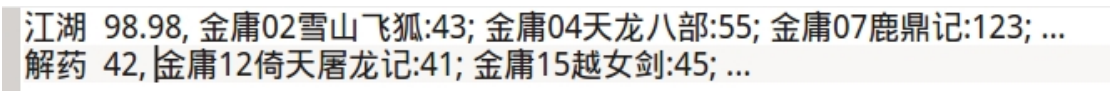
输出格式

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）：



选做内容

该部分内容不做要求，供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. 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，HDFS 存储位置为：

hdfs://master01:54310/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。

在实验报告中，需要展示输出文件的部分内容和在集群上执行时 JobTracker WebUI（http://114.212.190.91:50030/jobtracker.jsp）的截屏。