实验一 wordCount 算法及其实现

实验目的

1、理解 map-reduce 算法思想与流程；

2、应用 map-reduce 思想解决 wordCount 问题；

3、掌握并应用 combine 与 shuffle 过程。

实验内容

提供预处理过的数据 wordCount-data.zip（里面包含预处理过的 source01-09 源文件）模拟 9 个分布式节点，每个源文件中包含一百万个由英文、数字和字符（不包括逗号）构成的单词，单词由逗号与换行符分割。

要求应用 map-reduce 思想，模拟 9 个 map 节点与 3 个 reduce 节点实现 wordCount 功能，完成以下任务：

1) 输出对应的 map 文件和最终的 reduce 结果文件。由于源文件较大，要求使用多线程来模拟分布式节点；
2) 在任务 1）的基础上，添加 combine 与 shuffle 过程，并计算线程运行时间来考察这些过程对算法整体的影响。

提示：实现 shuffle 过程时应保证每个 reduce 节点的工作量尽量相当，来减少整体运行时间。