

**课 程 实 验 报 告**

**课程名称： 大数据分析**

**专业班级：**

**学 号：**

**姓 名：**

**指导教师：**

**报告日期：**

**计算机科学与技术学院**

**目录**

[实验一 wordCount算法及其实现 1](#_Toc57053298)

[**1.1实验目的** 1](#_Toc57053299)

[**1.2 实验内容** 1](#_Toc57053300)

[**1.3 实验过程** 1](#_Toc57053301)

[1.3.1 编程思路 1](#_Toc57053302)

[1.3.2 遇到的问题及解决方式 1](#_Toc57053303)

[1.3.3 实验测试与结果分析 1](#_Toc57053304)

[**1.4 实验总结** 2](#_Toc57053305)

# 实验一 Map-reduce算法及其实现

## **1.1实验目的**

1、理解map-reduce算法思想与流程；

2、应用map-reduce思想解决问题；

3、（可选）掌握并应用combine与shuffle过程。

## **1.2 实验内容**

提供9个预处理过的文件夹（folder\_1-9）模拟9个分布式节点中的数据，每个源文件夹中包含大约6千个文件，每个文件标题为维基百科条目标题，内容为对应的网页内容。提供words.txt文件作为待统计的词汇。

要求应用map-reduce思想，模拟9个map节点与3个reduce节点实现对维基百科条目词汇的词频的统计。

map节点输出<((title1,key1),1),…,((titlem,keyn),1)>，其中key为文件title.txt中出现的且在words.txt中词。同时，要求最终的reduce节点输出出现次数最多的前1000个词汇，以及这些词汇的跳转关系。

输出对应的map文件和最终的reduce结果文件。要求使用多线程来模拟分布式节点。

学有余力的同学可以在map-reduce的基础上添加combine与shuffle过程，并可以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。

## **1.3 实验过程**

### 1.3.1 编程思路

### 1.3.2 遇到的问题及解决方式

### 1.3.3 实验测试与结果分析

## **1.4 实验总结**