**Mapreduce实现倒排索引**

赖皓，硕士1905班，M201973231

# 1 实验目的

1. 掌握倒排索引概念及实现流程
2. 学习掌握mapreduce基本概念
3. 使用mapreduce框架对提供的文件建立倒排索引项

# 2 实验内容

使用mapreduce实现倒排索引实验，输入目录为以下三个文件。

* + file1.txt //内容：mapreduce is simple
  + file2.txt //内容：mapreduce is powerful and simple
  + file3.txt //内容：mapreduce and mapreduce

期待倒排后输出文件内容为

* + and file3.txt:1;file2.txt:1;
  + is file2.txt:1;file1.txt:1;
  + mapreduce file1.txt:1;file2.txt:1;file3.txt:2;
  + powerful file2.txt:1;
  + simple file2.txt:1;file1.txt:1;

# 3 实验方法

## 3.1 Map的实现

Map接受的value为文件split的内容，输出的KV对为<word:filename，1>，key为单词和所属的文件名，类型为Text。value为数量1，类型为Intwritable。

用StringTokennizer对输入的文件按照空格分离出单词，每分离出一个单词，就和文件名组合成一个key值，并用1作为value值将KV对交给Combine处理。

## 3.2 Combine的实现

Combine接受的KV对为Map的输出，输出的KV对为<word,filename:count>。key为单词，类型为Text。value为文件名和单词在这个文件出现的次数，类型为Text。

在Combine中统计一个单词在一个文件中的出现次数，并将key值设置为单词后输出给reduce处理。

## 3.3 Reduce的实现

Reduce接受的KV对为Combine的输出，输出的KV对为<word,<filename1:count1>,<filename2:count2>,…,>。key为单词，类型为Text。value为文件名和单词在这个文件出现次数的汇总，类型为Text。

在Reduce中将接受的value结合为一个Text即可完成。

# 4 实验结果

成功通关网站编程题：



图 1 通关成功截图

# 5总结与收获

1. 学习并了解了hadoop/mapreduce编程框架，并用其实现了简单的程序。
2. 学习并了解了倒排索引相关知识，并结合mapreduce编程框架建立了倒排索引项。