**课程名称：**新技术相关课程2

**实验名称：**倒排索引

# 一、实验目的

熟悉并掌握hadooop中MapReduce的用法；

理解并掌握倒排索引的设计与实现。

# 二、实验内容

**1、倒排索引：**

由已给出的“文档集合”建立其“倒排列表”:

1. 、遍历文档中的每一个单词；
2. 、将文档中一个单词出现的次数进行统计；
3. 、将文档集合中得出来的所有单词存储到一个词典（txt）中，并记录这个 单词在某个文件出现的次数。
4. 、为避免词典中单词过多导致的搜索效率慢问题，采用hash算法将不同 单词存入不同文件中。算法内容暂定为根据单词首字母将单词存入 a~z.txt的文件中，其他情况存入other.txt。
5. **MapReduce:**

将单词计频工作拆分，利用分布式hadoop的MapReduce分配给多人共同完 成。

1. 、Mapper拆分单词计频工作，它接收<文件名，单词>形式的输入，同时 形成一个<单词，文件路径+“1”>的输出。
2. 、Combiner接收以上<单词，文件路径+“1”>的输入，将value值的文件 路径与“1”分开，然后将所有“1”相加，得到词频，输出<单词，文件 路径===>n>。
3. 、Reducer接收以上Combiner的输出，将相同的单词合并在一起，且不同 文件间用“；”隔开。

# 三、实验环境

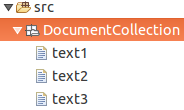
Linux;

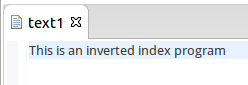
hadoop-0.20.203.0rc1.tarl;

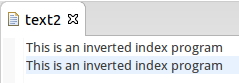
Eclipse;

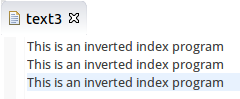
# 四、实验结果

**1、文档集合：**

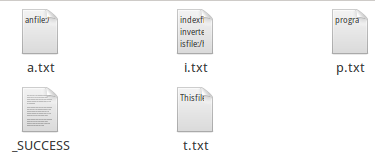








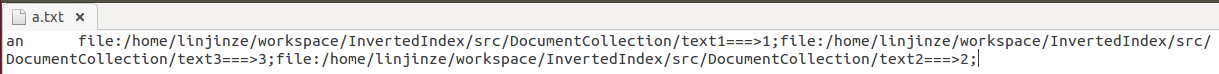
# 2、单词索引列表：

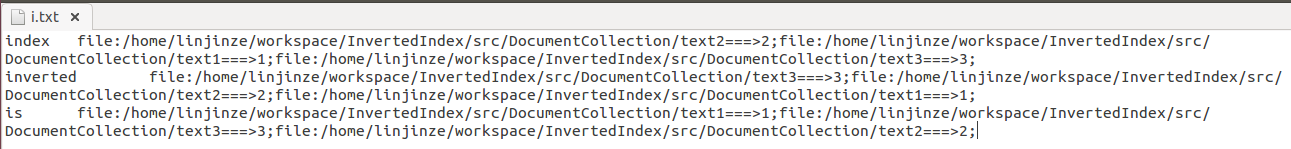


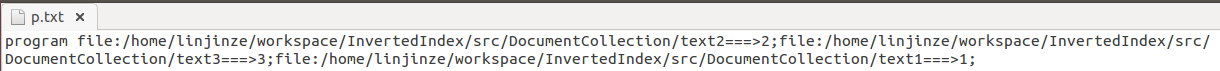
存储示例

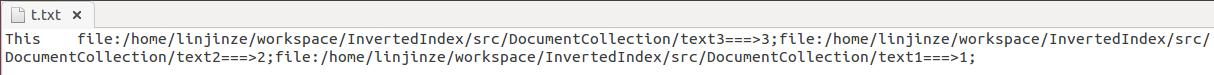


所有存储文件：









# 五、附录

1、方案设计：

A)、Mapper（extends Mapper）拆分单词计频工作，它接收<文件名，单词>形式 的输入，同时形成一个<单词，文件路径+“1”>的输出。

B)、Combiner（extends Reducer）接收以上<单词，文件路径+“1”>的输入，将 value值的文件路径与“1”分开，然后将所有“1”相加，得到词频， 输出 <单词，文件路径===>n>。

C)、Reducer （extends Reducer）接收以上Combiner的输出，将相同的单词合并 在一起，且不同文件归入此单词用“；”隔开。

D)、OutputFileName（extends MultipleOutputFormat）设置多文件输出模式，单 词的输出文件名由单词根据某hash算法得出，此处hash算法简单设为根据 单词首字母决定输出文件名a~z.txt，其他情况为other.txt。

1. 调试心得：
2. 、对于需要传输的数据，类型应该定义为hadoop提供的数据类型，以便用这 些类型定义的数据可以被序列化进行网络传输和文件存 储，以及进行大小 比较。如：IntWritable：整型数； Text：使用UTF8格式存储的文本。
3. 、combiner统计单词频率时，由于传输的value值包含了文件URL和数值 1 因此要先将二者分开，以便形成新的value（形式： URL===>n）时能输入 正确是URL。
4. 、多文件输出：Hadoop默认的输出是TextOutputFormat，输出文件名不可定制。 hadoop 0.19.X中有一个 org.apache.hadoop.mapred.lib.MultipleOutputFormat， 可以输出多份文件且可以自定义文件名，但是从hadoop 0.20.x中 MultipleOutputFormat所在包的所有类被标记为“已过时”，它将 MultipleOutputFormat与MultipleOutputs合并为MultipleOutputs，却并未 出 与之兼容的方法。因此本项目到网上找了一个别人写的新 MultipleOutputFormat类（附带LineRecordWrite类r：RecordWriter的一个实 现，用于把<Key, Value>转化为一行文本。在Hadoop中，这个类作为 TextOutputFormat的一个子类存在，protected访问权限，因此普通程序无法 访问。这里仅仅是把LineRecordWriter从TextOutputFormat抽取出来，作为 一个独立的公共类使用。）使用此MultipleOutputFormat类实现了新版本的 多文件输出。
5. 、Job初始化：

①、对Job进行合理的命名有助于更快地找到Job，以便在JobTracker和 Tasktracker的页面中对其进行监视 ；

②、setMapperClass（即我们所写的Mapper），setMapOutputKeyClass， setMapOutputValueClass（Map输出数据 类型）；

③、setReducerClass（即我们所写的Reducer）,setOutputKeyClass， setOutputValueClass（Reducer输出数据类型）,setOutputFormatClass （设置多文件输出）；

④、 FileInputFormat、FileOutputFormat设置输入、输出文件路径。