大数据综合处理实验

*Lab3 文档倒排索引及词频统计*

**30号小组**

小组成员：

|  |  |  |
| --- | --- | --- |
| 学号 | 姓名 | 邮箱 |
| 161220018 | 陈昕元 | 161220018@smail.nju.edu.cn |
| 161220017 | 陈翔 |  |
| 161220019 | 陈亚栋 |  |

# 1. 实验背景：问题和数据集介绍

# 2. 实验环境

* + 操作系统版本：Ubuntu 18.04
  + Java版本：1.7.0
  + Hadoop版本：2.7.1

# 3. 使用算法及实现

### （1）分布式倒排索引算法介绍

### （2）Map过程实现

### （3）Combiner过程实现

### （4）Partition过程实现

### （5）Reduce过程实现

### （6）项目改进及扩展功能实现

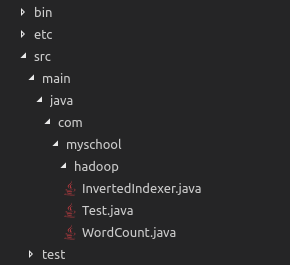
# 3. 实验过程介绍

### （1）Maven的下载与使用

首先从Maven的官网下载最新的maven压缩包，版本号为3.6.1，解压到特定文件夹。这里最好更换掉maven的默认镜像，否则下载插件会很慢。接着用指令

mvn archetype:generate -DgroupId=com.myschool.hadoop -DartifactId= hadoop -Dpackage=com.myschool.hadoop -Dversion=1.0-SNAPSHOT

生成我们的java框架，如下：



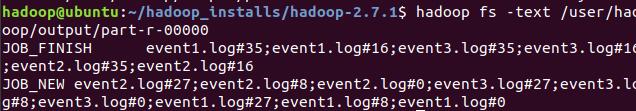
注意，其中的所有java文件都是之后手动添加的。

接着编辑其中的InvertedIndexer文件，上文中已经给出框架，编辑完成后运行如下指令：

mvn package

我们就可以得到一个target文件夹，里面有生成的jar文件，同时还有所有class文件，后面我们会用到其中的InvertedIndexer类。

最后我们在本地使用规模较小的样例测试，其中每个文档都完全相同，只有单词JOB\_NEW和JOB\_FINISH。运行的具体指令与实验2类似，下面为结果截图：



# 4. 实验总结

# 5. 代码使用方法