# 电 子 科 技 大 学

# 实　验　报　告

**学生姓名：Lolipop　　学号：2018091202000　　指导教师：xx**

**实验地点：信软学院楼西 400　　实验时间：2020.11.18**

## 一、实验名称：Hadoop 下单词反向索引程序

## 二、实验学时：4 学时

## 三、实验目的：

1. 熟悉反向索引；
2. 加强 MapReduce 编程能力。

## 四、实验原理：

反向索引（Inverted index），也常被称为倒排索引，是一种索引方法，被用来存储在全文搜索下某个单词在一个文档或者一组文档中的存储位置的映射。它是文档检索系统中最常用的数据结构。

有两种不同的反向索引形式：

1. 一条记录的水平反向索引（或者反向档案索引）包含每个引用单词的文档的列表。
2. 一个单词的水平反向索引（或者完全反向索引）又包含每个单词在一个文档中的位置。

后者的形式提供了更多的兼容性（比如短语搜索），但是需要更多的时间和空间来创建。

举例：

以英文为例，下面是要被索引的文本：

```
T0 = "it is what it is"
T1 = "what is it"
T2 = "it is a banana"
```

我们就能得到下面的反向文件索引：

```
"a":          {2}
"banana": {2}
"is":         {0, 1, 2}
"it":         {0, 1, 2}
"what":    {0, 1}
```

检索的条件"what", "is" 和 "it" 将对应这个集合：

{0,1}∩{0,1,2}∩{0,1,2}={0,1}。

## 五、实验内容：

运用已学习的知识，在 Eclipse 上编写实现一个适用于 Hadoop 中进行反向索引的程序并执行，输出结果。

## 六、实验器材（设备、元器件）：

1. Ubuntu 20.04
2. jdk 1.8
3. hadoop 2.10.1
4. Eclipse Luna 4.4

## 七、实验步骤：

1. 编写一个反向索引程序。
2. 使用如下文本测试 InvertedIndex 程序。

```
T0 = "a singular fatality has ruled the destiny of nearly all the most famous of Leonardo da Vinci's works"
```

T1 = "Leonardo's literary labours in various departments both of Art and of Science were those essentially of an enquirer"

T2 = "no reader could find his way through such a labyrinth and Leonardo himself could not have done it"

T3= "thus the Manuscripts that remain represent a period of about thirty years"

T4= "within this space of time his handwriting altered so little that it is impossible to judge from it of the date of any particular text"

T5= “the beginning of Leonardo's literary labours dates from about his thirty-seventh year, and he seems to have carried them on without any serious interruption till his death”

T6= “the assistance these afford for a chronological arrangement of the Manuscripts is generally self evident”

T7=”the truth is that the labours of three centuries have hardly sufficed for the elucidation of some of the problems which occupied his mighty mind”

T8=”I have given as complete a review of these writings as seemed necessary in the Bibliographical notes”

T9=”I may venture to state that I have devoted especial care and thought to the due execution of this responsible task”

## 八、实验结果与分析（含重要数据结果分析或核心代码流程分析）

1. 编写一个反向索引程序

   a) 创建新的 Map/Reduce 项目，命名为 InvertedIndex。

   b) 根据实验指导书提供的代码，分别编写 TokenInputFormat.java，ValuePair.java 和 InvertedIndex.java 函数。

   c) 编写后的项目如图 1 所示。

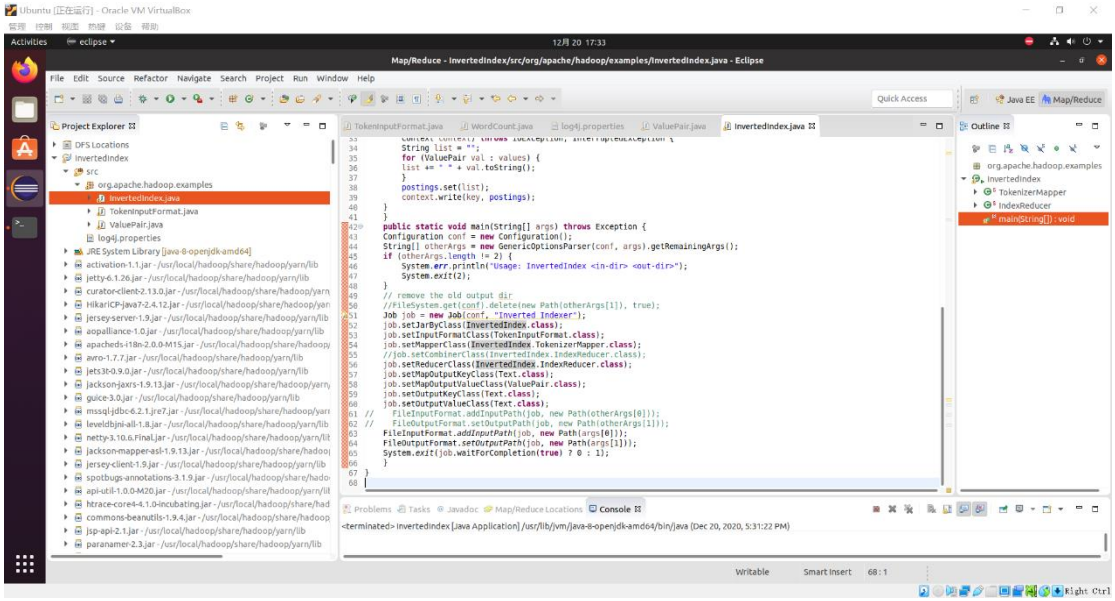


图 1 编写反向索引程序

2. 使用文本测试 InvertedIndex 程序

a) 在/usr/local/hadoop/input 目录下依次创建 file0.txt 至 file9.txt 文件，输入实验要求的测试文本。如图 2-1 所示。

```
hadooper@lolipop-VirtualBox:/usr/local/hadoop/input$ sudo vim file0.txt
hadooper@lolipop-VirtualBox:/usr/local/hadoop/input$ sudo vim file1.txt
hadooper@lolipop-VirtualBox:/usr/local/hadoop/input$ sudo vim file2.txt
hadooper@lolipop-VirtualBox:/usr/local/hadoop/input$ sudo vim file3.txt
hadooper@lolipop-VirtualBox:/usr/local/hadoop/input$ sudo vim file4.txt
hadooper@lolipop-VirtualBox:/usr/local/hadoop/input$ sudo vim file5.txt
hadooper@lolipop-VirtualBox:/usr/local/hadoop/input$ sudo vim file6.txt
hadooper@lolipop-VirtualBox:/usr/local/hadoop/input$ sudo vim file7.txt
hadooper@lolipop-VirtualBox:/usr/local/hadoop/input$ sudo vim file8.txt
hadooper@lolipop-VirtualBox:/usr/local/hadoop/input$ sudo vim file9.txt
hadooper@lolipop-VirtualBox:/usr/local/hadoop/input$ ls
file0.txt  file2.txt  file4.txt  file6.txt  file8.txt
file1.txt  file3.txt  file5.txt  file7.txt  file9.txt
hadooper@lolipop-VirtualBox:/usr/local/hadoop/input$ cat ./*
a singular fatality has ruled the destiny of nearly all the most famous of Leona
rdo da Vinci's works
Leonardo's literary labours in various departments both of Art and of Science we
re those essentially of an enquirer
no reader could find his way through such a labyrinth and Leonardo himself could
 not have done it
thus the Manuscripts that remain represent a period of about thirty years
within this space of time his handwriting altered so little that it is impossibl
e to judge from it of the date of any particular text
the beginning of Leonardo's literary labours dates from about his thirty-seventh
 year, and he seems to have carried them on without any serious interruption til
l his death
the assistance these afford for a chronological arrangement of the Manuscripts i
s generally self evident
the truth is that the labours of three centuries have hardly sufficed for the el
ucidation of some of the problems which occupied his mighty mind
I have given as complete a review of these writings as seemed necessary in the B
ibliographical notes
I may venture to state that I have devoted especial care and thought to the due
execution of this responsible task
hadooper@lolipop-VirtualBox:/usr/local/hadoop/input$
```

图 2-1 编写测试文件

b) 删除分布式文件系统中之前用作测试的 README.txt 文件，并将本地系统目录中/usr/local/hadoop/input 中的所有文件通过 put 命令复制到分布式文件系统中的/user/hadooper/input 目录下。如图 2-2 所示。

```
hadooper@lolipop-VirtualBox:/usr/local/hadoop$ bin/hdfs dfs -put input/* /user/h
adooper/input
hadooper@lolipop-VirtualBox:/usr/local/hadoop$ bin/hdfs dfs -ls /user/hadooper/i
nput
Found 10 items
-rw-r--r--   1 hadooper supergroup        101 2020-12-20 20:42 /user/hadooper/in
put/file0.txt
-rw-r--r--   1 hadooper supergroup        116 2020-12-20 20:42 /user/hadooper/in
put/file1.txt
-rw-r--r--   1 hadooper supergroup         98 2020-12-20 20:42 /user/hadooper/in
put/file2.txt
-rw-r--r--   1 hadooper supergroup         74 2020-12-20 20:42 /user/hadooper/in
put/file3.txt
-rw-r--r--   1 hadooper supergroup        134 2020-12-20 20:42 /user/hadooper/in
put/file4.txt
-rw-r--r--   1 hadooper supergroup        172 2020-12-20 20:42 /user/hadooper/in
put/file5.txt
-rw-r--r--   1 hadooper supergroup        105 2020-12-20 20:42 /user/hadooper/in
put/file6.txt
-rw-r--r--   1 hadooper supergroup        145 2020-12-20 20:42 /user/hadooper/in
put/file7.txt
-rw-r--r--   1 hadooper supergroup        101 2020-12-20 20:42 /user/hadooper/in
put/file8.txt
-rw-r--r--   1 hadooper supergroup        115 2020-12-20 20:42 /user/hadooper/in
put/file9.txt
hadooper@lolipop-VirtualBox:/usr/local/hadoop$
```

图 2-2 移动测试文件到伪分布式文件服务器中

c) 执行 InvertedIndex,java 文件，结果保存在分布式文件系统中的 /user/hadooper/output 目录下。使用 cat 命令读取分布式文件系统中 /user/hadooper/output 目录下的内容。如图 2-3 所示。

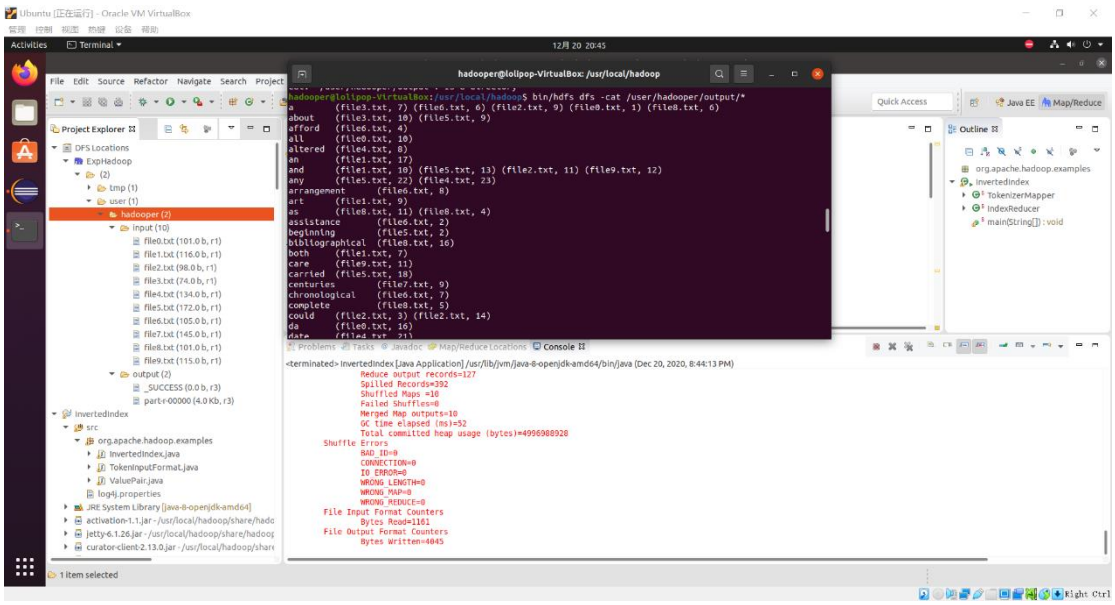


图 2-3 执行反向索引程序结果

读取目录中记录的内容如下所示。

a (file3.txt, 7) (file6.txt, 6) (file2.txt, 9) (file0.txt, 1) (file8.txt, 6)

about (file3.txt, 10) (file5.txt, 9)

afford (file6.txt, 4)

all (file0.txt, 10)

altered (file4.txt, 8)

an (file1.txt, 17)

and (file1.txt, 10) (file5.txt, 13) (file2.txt, 11) (file9.txt, 12)

any (file5.txt, 22) (file4.txt, 23)

arrangement (file6.txt, 8)

art (file1.txt, 9)

as (file8.txt, 11) (file8.txt, 4)

assistance (file6.txt, 2)

beginning (file5.txt, 2)

bibliographical (file8.txt, 16)

both (file1.txt, 7)

care (file9.txt, 11)

carried (file5.txt, 18)

centuries (file7.txt, 9)

chronological (file6.txt, 7)

complete (file8.txt, 5)

could (file2.txt, 3) (file2.txt, 14)

da (file0.txt, 16)

date (file4.txt, 21)

dates (file5.txt, 7)

death (file5.txt, 27)

departments (file1.txt, 6)

destiny (file0.txt, 7)

devoted (file9.txt, 9)

done (file2.txt, 17)

due (file9.txt, 16)

elucidation (file7.txt, 15)

enquirer (file1.txt, 18)

especial (file9.txt, 10)

essentially (file1.txt, 15)

evident (file6.txt, 15)

execution (file9.txt, 17)

famous (file0.txt, 13)

fatality (file0.txt, 3)

find (file2.txt, 4)

for (file6.txt, 5) (file7.txt, 13)

from (file5.txt, 8) (file4.txt, 17)

generally (file6.txt, 13)

given (file8.txt, 3)

handwriting (file4.txt, 7)

hardly (file7.txt, 11)

has (file0.txt, 4)

have (file7.txt, 10) (file9.txt, 8) (file5.txt, 17) (file8.txt, 2) (file2.txt, 16)

he (file5.txt, 14)

himself (file2.txt, 13)

his (file2.txt, 5) (file4.txt, 6) (file7.txt, 23) (file5.txt, 26) (file5.txt, 10)

i (file9.txt, 1) (file9.txt, 7) (file8.txt, 1)

impossible (file4.txt, 14)

in (file8.txt, 14) (file1.txt, 4)

interruption (file5.txt, 24)

is (file4.txt, 13) (file6.txt, 12) (file7.txt, 3)

it (file2.txt, 18) (file4.txt, 18) (file4.txt, 12)

judge (file4.txt, 16)

labours (file7.txt, 6) (file1.txt, 3) (file5.txt, 6)

labyrinth (file2.txt, 10)

leonardo (file2.txt, 12) (file0.txt, 15)

leonardo's (file1.txt, 1) (file5.txt, 4)

literary (file5.txt, 5) (file1.txt, 2)

little (file4.txt, 10)

manuscripts (file3.txt, 3) (file6.txt, 11)

may (file9.txt, 2)

mighty (file7.txt, 24)

mind (file7.txt, 25)

most (file0.txt, 12)

nearly (file0.txt, 9)

necessary (file8.txt, 13)

no (file2.txt, 1)

not (file2.txt, 15)

notes (file8.txt, 17)

occupied (file7.txt, 22)

of (file7.txt, 7) (file7.txt, 18) (file7.txt, 16) (file8.txt, 8) (file9.txt, 18) (file0.txt, 8) (file0.txt, 14) (file4.txt, 4) (file4.txt, 22) (file4.txt, 19) (file1.txt, 11) (file1.txt, 8) (file1.txt, 16) (file6.txt, 9) (file3.txt, 9) (file5.txt, 3)

on (file5.txt, 20)

particular (file4.txt, 24)

period (file3.txt, 8)

problems (file7.txt, 20)

reader (file2.txt, 2)

remain (file3.txt, 5)

represent (file3.txt, 6)

responsible (file9.txt, 20)

review (file8.txt, 7)

ruled (file0.txt, 5)

science (file1.txt, 12)

seemed (file8.txt, 12)

seems (file5.txt, 15)

self (file6.txt, 14)

serious (file5.txt, 23)

singular (file0.txt, 2)

so (file4.txt, 9)

some (file7.txt, 17)

space (file4.txt, 3)

state (file9.txt, 5)

such (file2.txt, 8)

sufficed (file7.txt, 12)

task (file9.txt, 21)

text (file4.txt, 25)

that (file4.txt, 11) (file9.txt, 6) (file3.txt, 4) (file7.txt, 4)

the (file7.txt, 1) (file7.txt, 14) (file7.txt, 19) (file7.txt, 5) (file3.txt, 2) (file4.txt, 20) (file8.txt, 15) (file9.txt, 15) (file0.txt, 11) (file0.txt, 6) (file5.txt, 1) (file6.txt, 10) (file6.txt, 1)

them (file5.txt, 19)

these (file8.txt, 9) (file6.txt, 3)

thirty (file3.txt, 11)

thirty-seventh (file5.txt, 11)

this (file4.txt, 2) (file9.txt, 19)

those (file1.txt, 14)

thought (file9.txt, 13)

three (file7.txt, 8)

through (file2.txt, 7)

thus (file3.txt, 1)

till (file5.txt, 25)

time (file4.txt, 5)

to (file4.txt, 15) (file9.txt, 14) (file9.txt, 4) (file5.txt, 16)

truth (file7.txt, 2)

various (file1.txt, 5)

venture (file9.txt, 3)

vinci's (file0.txt, 17)

way (file2.txt, 6)

were (file1.txt, 13)

which (file7.txt, 21)

within (file4.txt, 1)

without (file5.txt, 21)

works (file0.txt, 18)

writings (file8.txt, 10)

year (file5.txt, 12)

years (file3.txt, 12)

## 九、总结及心得体会：

// removed

## 十、对本实验过程及方法、手段的改进建议：

// removed

**报告评分：**

**指导教师签字：**