大数据作业六

191098273 徐冰清

作业要求

在作业5的数据集基础上完成莎士比亚文集单词的倒排索引，输出按字典序对单词进行排序，单词的索引按照单词在该文档中出现的次数从大到小排序。单词忽略大小写，忽略标点符号（punctuation.txt），忽略停词（stop-word-list.txt），忽略数字，单词长度>=3。输出格式如下：

单词1: 文档i#频次a, 文档j#频次b...

单词2: 文档m#频次x, 文档n#频次y...

...

思路分析

根据老师分享在群里的倒排索引代码 `InvertedIndexer.java`，加以适当改写，目的有三：

一是对单词做出处理

二是在输出时按照单词在该文档中出现的次数从大到小排序

三是使输出满足一定的格式规范

整体架构

Map

1. InvertedIndexMapper

输入	**key: 文件当前行偏移位置，value: 文件当前行内容**
输出	key: word#filename，value: 1

使用默认的 TextInputFormat 类对输入文件进行处理，得到文本中每行的偏移量及其内容。获取当前处理文件名filename，对value值进行切分得到多个word值，将每个word与filename拼接到一起作为输出key，其计数值为1，即value为1。

2. SumCombiner

输入	**key: word#filename，value: [1, 1, 1, ...]**
输出	key: word#filename，value: 同一key下的累加和

将Mapper输出的中间结果相同key部分的value累加，经过 map 方法处理后， Combine 过程将 key 值相同的 value 值累加，得到一个单词在文档中的词频。

Reduce

1. NewPartitioner

输入	key: word#filename， value: 累加和
输出	key: word#filename， value: 累加和

2. InvertedIndexReducer

输入	key: word#filename， vaule: [累加和1, 累加和2, ...]
输出	key: word,filename:词频;filename:词频;...

利用Reduce节点输入的key值都是有序的，将key拆分，对于同一word，每次都保存其filename和词频，并统计其总出现次数和总出现文档数；当同一word处理完后，filename及其词频作为value输出。

单词处理

将所有字母转为小写

```
String line = value.toString().toLowerCase();
```

忽略标点符号

```
Path punctuationsPath = new Path(patternsURIs[1].getPath());
String punctuationsFileName = punctuationsPath.getName().toString();
parseSkipPunctuations(punctuationsFileName);

fis = new BufferedReader(new FileReader(fileName));
String pattern = null;
while ((pattern = fis.readLine()) != null) {
  patternsToSkip.add(pattern);
}

for (String pattern : punctuations) {
        line = line.replaceAll(pattern, " ");
}
```

忽略长度小于3的单词

```
StringTokenizer itr = new StringTokenizer(line);
while (itr.hasMoreTokens()) {
  String one_word = itr.nextToken();
  if(one_word.length()<3) {
    continue;
  }
}
```

忽略停词文件里的词

```
  Path patternsPath = new Path(patternsURIs[0].getPath());
  String patternsFileName = patternsPath.getName().toString();
  parseSkipFile(patternsFileName);

  fis = new BufferedReader(new FileReader(fileName));
  String pattern = null;
  while ((pattern = fis.readLine()) != null) {
    patternsToSkip.add(pattern);
  }

  if(patternsToSkip.contains(one_word)){
    continue;
  }
```

次数排序

对 `InvertedIndexReducer` 中的 `reduce` 函数进行更改。因为默认对文件名排序，所以在设置word2的时候，将数字放在前面。

```
word2.set("<" + sum + "#" + temp + ">");
postingList.sort(Collections.reverseOrder());
```

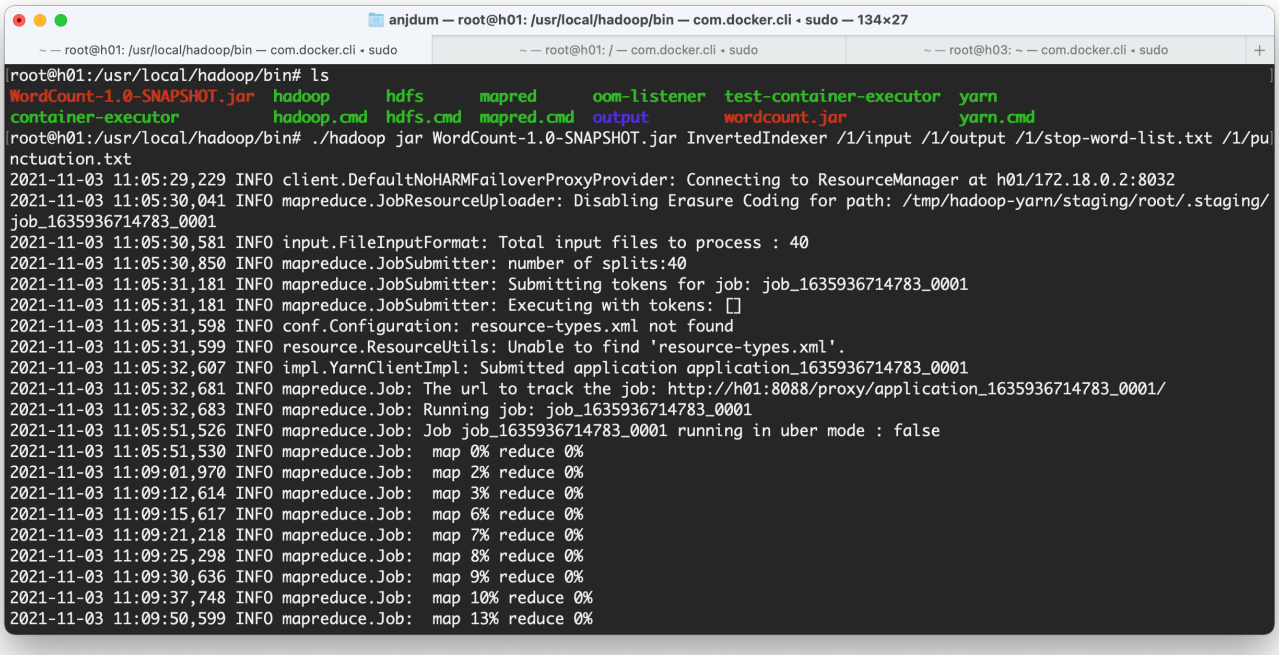
输出格式

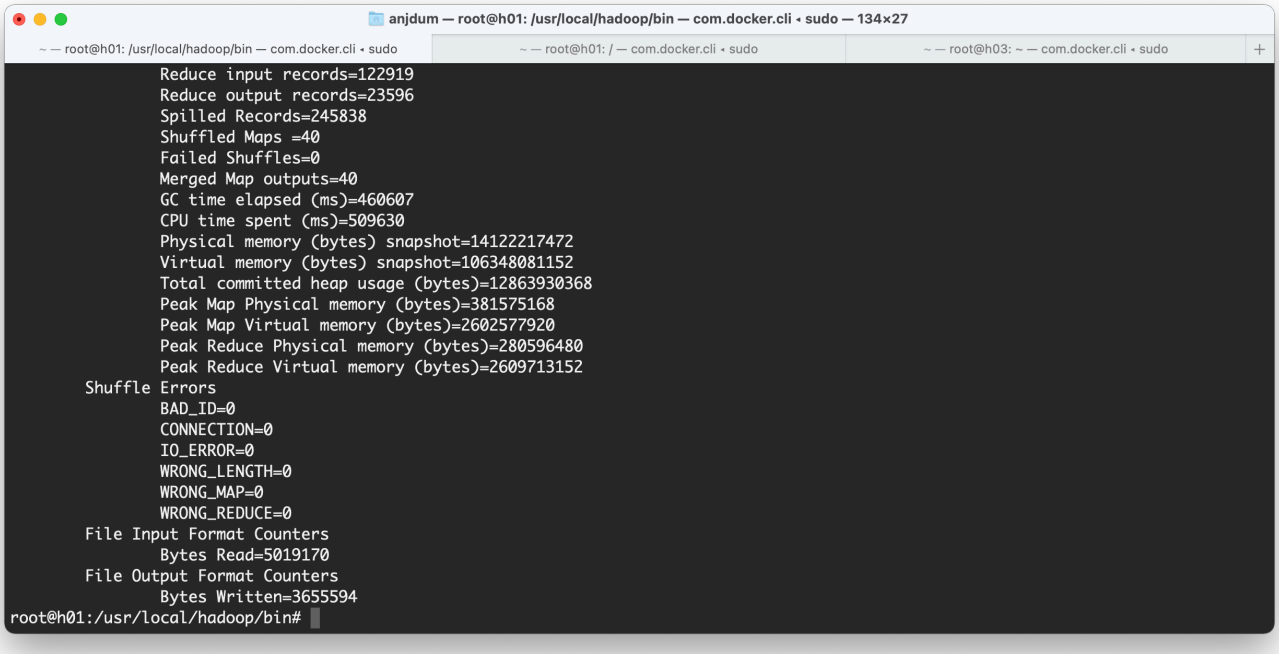
首先以数字+文件名的形式存入，再通过substring取出正确的顺序。

```
  word2.set("<" + sum + "#" + temp + ">");
  String currentItem = CurrentItem + ":";
  Text myItem = new Text(currentItem);
  for (String p : postingList) {
    String q = p.substring(p.indexOf("#")+1, p.indexOf(">")) + "#" + p.substring(p.indexOf("
  <")+1, p.indexOf("#"));
    out.append(q);
    out.append(",");
     count =
            count
                + Long.parseLong(p.substring(p.indexOf("<") + 1,
                     p.indexOf("#")));
  }
```

运行截图

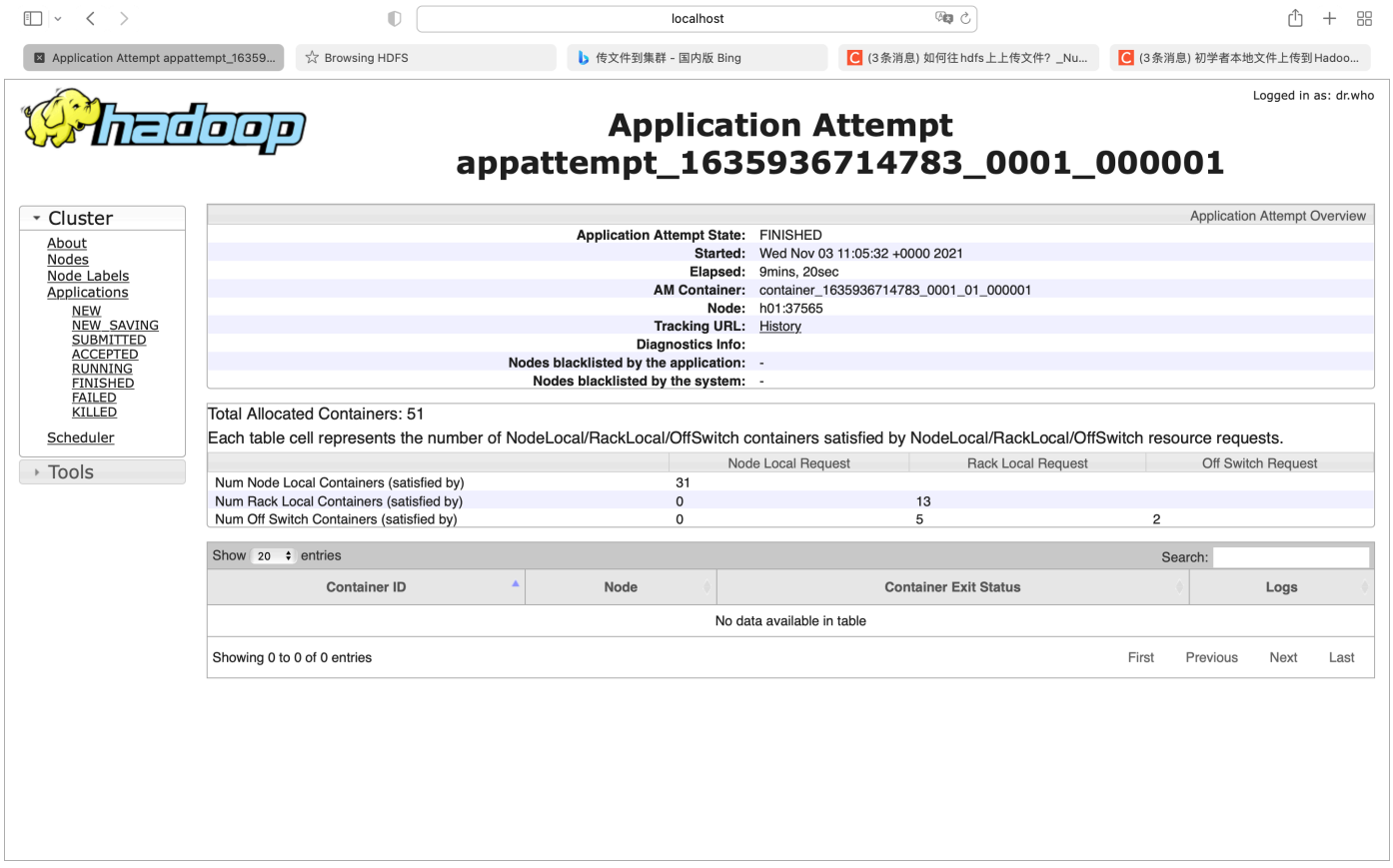
伪分布与集群


anjdum — root@h01: /usr/local/hadoop/bin — com.docker.cli • sudo — 134×27
~ — root@h01: /usr/local/hadoop/bin — com.docker.cli • sudo
~ — root@h01: / — com.docker.cli • sudo
~ — root@h03: ~ — com.docker.cli • sudo
root@h01:/usr/local/hadoop/bin# ls
WordCount-1.0-SNAPSHOT.jar hadoop hdfs mapred oom-listener test-container-executor yarn
container-executor hadoop.cmd hdfs.cmd mapred.cmd output wordcount.jar yarn.cmd
root@h01:/usr/local/hadoop/bin# ./hadoop jar WordCount-1.0-SNAPSHOT.jar InvertedIndexer /1/input /1/output /1/stop-word-list.txt /1/punctuation.txt
2021-11-03 11:05:29,229 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at h01/172.18.0.2:8032
2021-11-03 11:05:30,041 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1635936714783_0001
2021-11-03 11:05:30,581 INFO input.FileInputFormat: Total input files to process : 40
2021-11-03 11:05:30,850 INFO mapreduce.JobSubmitter: number of splits:40
2021-11-03 11:05:31,181 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635936714783_0001
2021-11-03 11:05:31,181 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-03 11:05:31,598 INFO conf.Configuration: resource-types.xml not found
2021-11-03 11:05:31,599 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-03 11:05:32,607 INFO impl.YarnClientImpl: Submitted application application_1635936714783_0001
2021-11-03 11:05:32,681 INFO mapreduce.Job: The url to track the job: http://h01:8088/proxy/application_1635936714783_0001/
2021-11-03 11:05:32,683 INFO mapreduce.Job: Running job: job_1635936714783_0001
2021-11-03 11:05:51,526 INFO mapreduce.Job: Job job_1635936714783_0001 running in uber mode : false
2021-11-03 11:05:51,530 INFO mapreduce.Job: map 0% reduce 0%
2021-11-03 11:09:01,970 INFO mapreduce.Job: map 2% reduce 0%
2021-11-03 11:09:12,614 INFO mapreduce.Job: map 3% reduce 0%
2021-11-03 11:09:15,617 INFO mapreduce.Job: map 6% reduce 0%
2021-11-03 11:09:21,218 INFO mapreduce.Job: map 7% reduce 0%
2021-11-03 11:09:25,298 INFO mapreduce.Job: map 8% reduce 0%
2021-11-03 11:09:30,636 INFO mapreduce.Job: map 9% reduce 0%
2021-11-03 11:09:37,748 INFO mapreduce.Job: map 10% reduce 0%
2021-11-03 11:09:50,599 INFO mapreduce.Job: map 13% reduce 0%


anjdum — root@h01: /usr/local/hadoop/bin — com.docker.cli • sudo — 134×27
~ — root@h01: /usr/local/hadoop/bin — com.docker.cli • sudo
~ — root@h01: / — com.docker.cli • sudo
~ — root@h03: ~ — com.docker.cli • sudo
Reduce input records=122919
Reduce output records=23596
Spilled Records=245838
Shuffled Maps =40
Failed Shuffles=0
Merged Map outputs=40
GC time elapsed (ms)=460607
CPU time spent (ms)=509630
Physical memory (bytes) snapshot=14122217472
Virtual memory (bytes) snapshot=106348081152
Total committed heap usage (bytes)=12863930368
Peak Map Physical memory (bytes)=381575168
Peak Map Virtual memory (bytes)=2602577920
Peak Reduce Physical memory (bytes)=280596480
Peak Reduce Virtual memory (bytes)=2609713152
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=5019170
File Output Format Counters
Bytes Written=3655594
root@h01:/usr/local/hadoop/bin#

```
libexec — -zsh — 127×24
/usr/local/cellar/hadoop/3.3.1/libexec — -zsh
[anjdum@AnJDumdeMacBook-Air libexec % bin/hdfs dfs -ls /InvertedIndex/output
2021-11-03 18:03:45,441 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jav
a classes where applicable
Found 2 items
-rw-r--r--   1 anjdum supergroup          0 2021-11-03 18:02 /InvertedIndex/output/_SUCCESS
-rw-r--r--   1 anjdum supergroup    3655592 2021-11-03 18:02 /InvertedIndex/output/part-r-00000
[anjdum@AnJDumdeMacBook-Air libexec % bin/hdfs dfs -cat /InvertedIndex/output/part-r-00000
2021-11-03 18:04:11,781 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jav
a classes where applicable
aaron:  shakespeare-titus-50.txt#98,
abaissiez:      shakespeare-life-54.txt#1,
abandon:        shakespeare-as-12.txt#4,shakespeare-twelfth-20.txt#1,shakespeare-troilus-22.txt#1,shakespeare-timon-49.txt#1,sh
akespeare-third-53.txt#1,shakespeare-taming-2.txt#1,shakespeare-othello-47.txt#1,
abandoned:      shakespeare-titus-50.txt#1,shakespeare-alls-11.txt#1,
abase:  shakespeare-tragedy-58.txt#1,
abash:  shakespeare-troilus-22.txt#1,
abate:  shakespeare-life-54.txt#5,shakespeare-venus-60.txt#1,shakespeare-tragedy-58.txt#1,shakespeare-titus-50.txt#1,shakespear
e-taming-2.txt#1,shakespeare-romeo-48.txt#1,shakespeare-midsummer-16.txt#1,shakespeare-merchant-5.txt#1,shakespeare-loves-8.txt
#1,shakespeare-hamlet-25.txt#1,shakespeare-cymbeline-17.txt#1,
abated:  shakespeare-second-52.txt#1,shakespeare-king-45.txt#1,shakespeare-coriolanus-24.txt#1,
abatement:      shakespeare-twelfth-20.txt#1,shakespeare-king-45.txt#1,shakespeare-cymbeline-17.txt#1,
abatements:     shakespeare-hamlet-25.txt#1,
abates:  shakespeare-tempest-4.txt#1,
abbess:  shakespeare-comedy-7.txt#8,
```

web



Logged in as: dr.who

Application application_1635936714783_0001

Application Overview	
User:	root
Name:	inverted index
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Wed Nov 03 11:05:32 +0000 2021
Launched:	Wed Nov 03 11:05:34 +0000 2021
Finished:	Wed Nov 03 11:14:53 +0000 2021
Elapsed:	9mins, 21sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Application Metrics	
Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	12431549 MB-seconds, 11542 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1635936714783_0001_000001	Wed Nov 3 19:05:32 +0800 2021	http://h01:8042	Logs	0	0

Showing 1 to 1 of 1 entries First Previous 1 Next Last