Lab4 - MapReduce

- 学号：516030910141
- 姓名：谢添翼
- 邮箱：330281987@qq.com

Part 1

Part 1主要需要完成doMap()和doReduce()两个函数

doMap逻辑主要如下:

1. 读取给定的file内容
2. 将给定的file文件名和读取的内容作为参数，调用用户定义的map()函数，获得key-value列表
3. 根据key的值将不同的key-value划分到不同的中间文件中，写入的格式采用json格式

其中注意中间文件的总个数是nMap*nReduce，一个mapper会将键值对放入其map id对应的nReduce个中间文件个数中。划分的依据则是将key值hash后，再模nReduce

json格式则使用JSON.toJSONString()函数即可实现

doReduce逻辑主要如下：

1. 读取reducer id对应的nMap个中间文件，并将所有具有相同key的key-value对整理至同一个key下的value数组
2. 对整理完后的每一个key分别调用用户定义的reduce()函数，生成新的key-value对
3. 将这些新的key-value对按key排序后，再按json格式写入指定的输出文件中

其中对于sort要求，采用的是用TreeMap数据结构来存储经reduce()处理过的新key-value对

Part 2

part 2要求完成wordcount的map()和reduce()函数

map()函数需要读取内容中的每一个word作为key，对应value则均为1，表示出现1次。最后包含所有的key-value对的list即可。其中寻找word采用正则表达式 "[a-zA-Z0-9]+"，利用Pattern和Matcher不断地去匹配即可

reduce()函数仅需将key中string数组的每一个元素转为int，再求和转为string返回即可

```
jos@cosmic:/mnt/hgfs/share$ sort -n -k2 mrtmp.wcseq | tail -10
that: 7871
it: 7987
in: 8415
was: 8578
a: 13382
of: 13536
I: 14296
to: 16079
and: 23612
the: 29748
```

Part 3&4

part 3和4都是要求对scheduler()函数进行修改

part 3要求scheduler将nTasks个并发分配给不同的worker，并且对于一个worker需要等待其做完一个任务后才能给其分发下一个任务。part 4则要求处理worker failure的问题，如果一个任务通过RPC发给worker但TimeOut了，则须将该任务分配给另一个worker

对于part 3中的并发分发任务：

- 应当对每一个任务都单独开一个线程并分别使用RPC发送请求，并且需等待所有任务结束完毕后，scheduler函数才能返回。
- 对于worker一次只能执行一个任务的要求，则采用registerChan的阻塞队列即可。每一个分发线程，先调用register.read()，若队列为空则会阻塞直至队列中有空闲的worker；每一个worker执行任务结束后，再调用register.writer()将该worker重新排入队列中。
- 对于等待所有任务执行完才返回的要求，采用CountDownLatch的方法：函数开头定义CountDownLatch latch = new CountDownLatch(nTasks)，每一个线程中等待worker成功执行完任务后调用latch.countDown()。这样函数最后的latch.await()便会等待所有latch结束后才返回

对于part 4中的worker failure问题，仅需将rpc call的exception catch，再通过循环重新从队列中选取新的worker分配任务即可。当call成功返回时，将worker写回等待队列再跳出循环即可

Part 5

part 5要求完成Inverted index generation的map()和reduce()函数

map()函数与wordcount类似，仅需将每一个key对应的value改为传入的file name

reduce()函数对value数组的操作改为，保留数组中互不重复的元素，重复元素去除。并将处理过的元素个数和元素内容按格式要求写入输出字符串中即可

```
jos@cosmic:/mnt/hgfs/share$ sort -k1,1 mrtmp.iiseq | sort -snk2,2 | grep -v '16' | tail -10
yesterday: 8 pg-being_ernest.txt,pg-dorian_gray.txt,pg-frankenstein.txt,pg-grimm.txt,pg-huckleberry_finn.txt,pg-metamorphosis.txt,pg-sherlock_holmes.txt,pg-tom_sawyer.txt
yet: 8 pg-being_ernest.txt,pg-dorian_gray.txt,pg-frankenstein.txt,pg-grimm.txt,pg-huckleberry_finn.txt,pg-metamorphosis.txt,pg-sherlock_holmes.txt,pg-tom_sawyer.txt
you: 8 pg-being_ernest.txt,pg-dorian_gray.txt,pg-frankenstein.txt,pg-grimm.txt,pg-huckleberry_finn.txt,pg-metamorphosis.txt,pg-sherlock_holmes.txt,pg-tom_sawyer.txt
You: 8 pg-being_ernest.txt,pg-dorian_gray.txt,pg-frankenstein.txt,pg-grimm.txt,pg-huckleberry_finn.txt,pg-metamorphosis.txt,pg-sherlock_holmes.txt,pg-tom_sawyer.txt
YOU: 8 pg-being_ernest.txt,pg-dorian_gray.txt,pg-frankenstein.txt,pg-grimm.txt,pg-huckleberry_finn.txt,pg-metamorphosis.txt,pg-sherlock_holmes.txt,pg-tom_sawyer.txt
young: 8 pg-being_ernest.txt,pg-dorian_gray.txt,pg-frankenstein.txt,pg-grimm.txt,pg-huckleberry_finn.txt,pg-metamorphosis.txt,pg-sherlock_holmes.txt,pg-tom_sawyer.txt
your: 8 pg-being_ernest.txt,pg-dorian_gray.txt,pg-frankenstein.txt,pg-grimm.txt,pg-huckleberry_finn.txt,pg-metamorphosis.txt,pg-sherlock_holmes.txt,pg-tom_sawyer.txt
Your: 8 pg-being_ernest.txt,pg-dorian_gray.txt,pg-frankenstein.txt,pg-grimm.txt,pg-huckleberry_finn.txt,pg-metamorphosis.txt,pg-sherlock_holmes.txt,pg-tom_sawyer.txt
yourself: 8 pg-being_ernest.txt,pg-dorian_gray.txt,pg-frankenstein.txt,pg-grimm.txt,pg-huckleberry_finn.txt,pg-metamorphosis.txt,pg-sherlock_holmes.txt,pg-tom_sawyer.txt
zip: 8 pg-being_ernest.txt,pg-dorian_gray.txt,pg-frankenstein.txt,pg-grimm.txt,pg-huckleberry_finn.txt,pg-metamorphosis.txt,pg-sherlock_holmes.txt,pg-tom_sawyer.txt
```