mapper.py

```
import sys
for line in sys.stdin:
        words=line.strip().split()
        for word in words:
                print(f'{word}\t1')
```

reducer.py

```
import sys

current_word = None
current_cnt = 0
word = None

for line in sys.stdin:
    line = line.strip()
    word, cnt = line.split('\t', 1)
    try:
        cnt = int(cnt)
    except ValueError:
        continue

    if current_word == word:
        current_cnt += cnt
    else:
        if current_word:
            print(f'{current_word}\t{current_cnt}')
        current_word = word
        current_cnt = cnt

# Ensure the last word is printed
if current_word == word:
    print(f'{current_word}\t{current_cnt}')
```

wordcount.txt

```
pig deer river
beer pig dear
beer bear river
```

code in hdfs
#put mapper.py,reducer.py,wordcount.txt in your folder

```
hadoop@ubuntu22:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
```

```
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ubuntu22]
Starting resourcemanager
Starting nodemanagers
hadoop@ubuntu22:~$ jps
11697 ResourceManager
12214 Jps
11097 NameNode
11484 SecondaryNameNode
11837 NodeManager
11277 DataNode
hdfs dfs -mkdir /160122771091
hadoop@ubuntu22:~$ hdfs dfs -put wordcount.txt  /160122771091
hadoop@ubuntu22:~$ hdfs dfs -put mapper.py  /160122771091
hadoop@ubuntu22:~$ hdfs dfs -put reducer.py  /160122771091
hadoop@ubuntu22:~$ hdfs dfs -ls /160122771091
Found 6 items
-rw-r--r--   1 hadoop supergroup        104 2025-01-28 14:45 /160122771091/mapper.py
drwxr-xr-x   - hadoop supergroup          0 2025-01-28 14:52 /160122771091/output.txt
drwxr-xr-x   - hadoop supergroup          0 2025-01-28 14:58 /160122771091/output1.txt
drwxr-xr-x   - hadoop supergroup          0 2025-01-28 15:03 /160122771091/output2.txt
-rw-r--r--   1 hadoop supergroup        392 2025-01-28 14:46 /160122771091/reducer.py
-rw-r--r--   1 hadoop supergroup         47 2025-01-28 14:38 /160122771091/wordcount.txt
hadoop@ubuntu22:~$ hadoop jar /home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-
streaming-3.3.6.jar -file mapper.py -mapper "python3 mapper.py" -file reducer.py -reducer
"python3 reducer.py" -input /160122771091/wordcount.txt -output /160122771091/output.txt
2025-01-28 15:10:37,691 WARN streaming.StreamJob: -file option is deprecated, please use
generic option -files instead.
packageJobJar: [mapper.py, reducer.py] [] /tmp/streamjob5561908958838107554.jar tmpDir=null
2025-01-28 15:10:38,343 INFO impl.MetricsConfig: Loaded properties from hadoop-
metrics2.properties
2025-01-28 15:10:38,476 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10
second(s).
2025-01-28 15:10:38,476 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-01-28 15:10:38,485 WARN impl.MetricsSystemImpl: JobTracker metrics system already
initialized!
2025-01-28 15:10:38,676 INFO mapred.FileInputFormat: Total input files to process : 1
2025-01-28 15:10:38,748 INFO mapreduce.JobSubmitter: number of splits:1
2025-01-28 15:10:38,835 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local1225768448_0001
2025-01-28 15:10:38,835 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-01-28 15:10:39,078 INFO mapred.LocalDistributedCacheManager: Localized
file:/home/hadoop/mapper.py as file:/tmp/hadoop-
hadoop/mapred/local/job_local1225768448_0001_337c50e5-f0f5-4e7d-9ba7-
a29c1c46c71e/mapper.py
2025-01-28 15:10:39,090 INFO mapred.LocalDistributedCacheManager: Localized
file:/home/hadoop/reducer.py as file:/tmp/hadoop-
hadoop/mapred/local/job_local1225768448_0001_2689bd17-21f3-4172-bb79-
03453e30527b/reducer.py
2025-01-28 15:10:39,253 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-01-28 15:10:39,254 INFO mapred.LocalJobRunner: OutputCommitter set in config null
```

```
2025-01-28 15:10:39,254 INFO mapreduce.Job: Running job: job_local1225768448_0001
2025-01-28 15:10:39,255 INFO mapred.LocalJobRunner: OutputCommitter is
org.apache.hadoop.mapred.FileOutputCommitter
2025-01-28 15:10:39,257 INFO output.FileOutputCommitter: File Output Committer Algorithm
version is 2
2025-01-28 15:10:39,257 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup
_temporary folders under output directory:false, ignore cleanup failures: false
2025-01-28 15:10:39,313 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-01-28 15:10:39,315 INFO mapred.LocalJobRunner: Starting task:
attempt_local1225768448_0001_m_000000_0
2025-01-28 15:10:39,331 INFO output.FileOutputCommitter: File Output Committer Algorithm
version is 2
2025-01-28 15:10:39,331 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup
_temporary folders under output directory:false, ignore cleanup failures: false
2025-01-28 15:10:39,343 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-01-28 15:10:39,349 INFO mapred.MapTask: Processing split:
hdfs://localhost:9000/160122771091/wordcount.txt:0+47
2025-01-28 15:10:39,367 INFO mapred.MapTask: numReduceTasks: 1
2025-01-28 15:10:39,455 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-01-28 15:10:39,455 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-01-28 15:10:39,455 INFO mapred.MapTask: soft limit at 83886080
2025-01-28 15:10:39,455 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-01-28 15:10:39,455 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-01-28 15:10:39,457 INFO mapred.MapTask: Map output collector class =
org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-01-28 15:10:39,461 INFO streaming.PipeMapRed: PipeMapRed exec [/usr/bin/python3,
mapper.py]
2025-01-28 15:10:39,463 INFO Configuration.deprecation: mapred.work.output.dir is deprecated.
Instead, use mapreduce.task.output.dir
2025-01-28 15:10:39,464 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead,
use mapreduce.cluster.local.dir
2025-01-28 15:10:39,464 INFO Configuration.deprecation: map.input.file is deprecated. Instead,
use mapreduce.map.input.file
2025-01-28 15:10:39,464 INFO Configuration.deprecation: map.input.length is deprecated. Instead,
use mapreduce.map.input.length
2025-01-28 15:10:39,464 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead,
use mapreduce.job.id
2025-01-28 15:10:39,465 INFO Configuration.deprecation: mapred.task.partition is deprecated.
Instead, use mapreduce.task.partition
2025-01-28 15:10:39,469 INFO Configuration.deprecation: map.input.start is deprecated. Instead,
use mapreduce.map.input.start
2025-01-28 15:10:39,469 INFO Configuration.deprecation: mapred.task.is.map is deprecated.
Instead, use mapreduce.task.ismap
2025-01-28 15:10:39,469 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead,
use mapreduce.task.attempt.id
2025-01-28 15:10:39,469 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use
mapreduce.task.id
2025-01-28 15:10:39,469 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead,
use mapreduce.job.skiprecords
2025-01-28 15:10:39,470 INFO Configuration.deprecation: user.name is deprecated. Instead, use
mapreduce.job.user.name
2025-01-28 15:10:39,712 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
```

```
2025-01-28 15:10:39,714 INFO streaming.PipeMapRed: Records R/W=3/1
2025-01-28 15:10:39,715 INFO streaming.PipeMapRed: MRErrorThread done
2025-01-28 15:10:39,715 INFO streaming.PipeMapRed: mapRedFinished
2025-01-28 15:10:39,724 INFO mapred.LocalJobRunner:
2025-01-28 15:10:39,724 INFO mapred.MapTask: Starting flush of map output
2025-01-28 15:10:39,724 INFO mapred.MapTask: Spilling map output
2025-01-28 15:10:39,724 INFO mapred.MapTask: bufstart = 0; bufend = 63; bufvoid = 104857600
2025-01-28 15:10:39,724 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend =
26214364(104857456); length = 33/6553600
2025-01-28 15:10:39,735 INFO mapred.MapTask: Finished spill 0
2025-01-28 15:10:39,768 INFO mapred.Task: Task:attempt_local1225768448_0001_m_000000_0
is done. And is in the process of committing
2025-01-28 15:10:39,780 INFO mapred.LocalJobRunner: Records R/W=3/1
2025-01-28 15:10:39,780 INFO mapred.Task: Task 'attempt_local1225768448_0001_m_000000_0'
done.
2025-01-28 15:10:39,787 INFO mapred.Task: Final Counters for
attempt_local1225768448_0001_m_000000_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=1358
                FILE: Number of bytes written=650361
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=47
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=3
                Map output records=9
                Map output bytes=63
                Map output materialized bytes=87
                Input split bytes=100
                Combine input records=0
                Spilled Records=9
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=275251200
        File Input Format Counters
                Bytes Read=47
2025-01-28 15:10:39,787 INFO mapred.LocalJobRunner: Finishing task:
attempt_local1225768448_0001_m_000000_0
2025-01-28 15:10:39,788 INFO mapred.LocalJobRunner: map task executor complete.
2025-01-28 15:10:39,792 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2025-01-28 15:10:39,792 INFO mapred.LocalJobRunner: Starting task:
attempt_local1225768448_0001_r_000000_0
2025-01-28 15:10:39,811 INFO output.FileOutputCommitter: File Output Committer Algorithm
version is 2
```

```
2025-01-28 15:10:39,811 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup
_temporary folders under output directory:false, ignore cleanup failures: false
2025-01-28 15:10:39,812 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-01-28 15:10:39,818 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin:
org.apache.hadoop.mapreduce.task.reduce.Shuffle@5473bd50
2025-01-28 15:10:39,820 WARN impl.MetricsSystemImpl: JobTracker metrics system already
initialized!
2025-01-28 15:10:39,861 INFO reduce.MergeManagerImpl: MergerManager:
memoryLimit=1272394496, maxSingleShuffleLimit=318098624, mergeThreshold=839780416,
ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-01-28 15:10:39,864 INFO reduce.EventFetcher: attempt_local1225768448_0001_r_000000_0
Thread started: EventFetcher for fetching Map Completion Events
2025-01-28 15:10:39,904 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map
attempt_local1225768448_0001_m_000000_0 decomp: 83 len: 87 to MEMORY
2025-01-28 15:10:39,909 INFO reduce.InMemoryMapOutput: Read 83 bytes from map-output for
attempt_local1225768448_0001_m_000000_0
2025-01-28 15:10:39,911 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of
size: 83, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->83
2025-01-28 15:10:39,914 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2025-01-28 15:10:39,914 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-01-28 15:10:39,915 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory
map-outputs and 0 on-disk map-outputs
2025-01-28 15:10:39,928 INFO mapred.Merger: Merging 1 sorted segments
2025-01-28 15:10:39,929 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left
of total size: 76 bytes
2025-01-28 15:10:39,930 INFO reduce.MergeManagerImpl: Merged 1 segments, 83 bytes to disk
to satisfy reduce memory limit
2025-01-28 15:10:39,930 INFO reduce.MergeManagerImpl: Merging 1 files, 87 bytes from disk
2025-01-28 15:10:39,931 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from
memory into reduce
2025-01-28 15:10:39,931 INFO mapred.Merger: Merging 1 sorted segments
2025-01-28 15:10:39,931 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left
of total size: 76 bytes
2025-01-28 15:10:39,932 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-01-28 15:10:39,942 INFO streaming.PipeMapRed: PipeMapRed exec [/usr/bin/python3,
reducer.py]
2025-01-28 15:10:39,944 INFO Configuration.deprecation: mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2025-01-28 15:10:39,945 INFO Configuration.deprecation: mapred.map.tasks is deprecated.
Instead, use mapreduce.job.maps
2025-01-28 15:10:40,025 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
2025-01-28 15:10:40,026 INFO streaming.PipeMapRed: Records R/W=9/1
2025-01-28 15:10:40,030 INFO streaming.PipeMapRed: MRErrorThread done
2025-01-28 15:10:40,031 INFO streaming.PipeMapRed: mapRedFinished
2025-01-28 15:10:40,258 INFO mapreduce.Job: Job job_local1225768448_0001 running in uber
mode : false
2025-01-28 15:10:40,258 INFO mapreduce.Job:  map 100% reduce 0%
2025-01-28 15:10:40,548 INFO mapred.Task: Task:attempt_local1225768448_0001_r_000000_0 is
done. And is in the process of committing
2025-01-28 15:10:40,552 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-01-28 15:10:40,552 INFO mapred.Task: Task attempt_local1225768448_0001_r_000000_0 is
allowed to commit now
```

```
2025-01-28 15:10:40,576 INFO output.FileOutputCommitter: Saved output of task
'attempt_local1225768448_0001_r_000000_0' to hdfs://localhost:9000/160122771091/output3.txt
2025-01-28 15:10:40,576 INFO mapred.LocalJobRunner: Records R/W=9/1 > reduce
2025-01-28 15:10:40,576 INFO mapred.Task: Task 'attempt_local1225768448_0001_r_000000_0'
done.
2025-01-28 15:10:40,577 INFO mapred.Task: Final Counters for
attempt_local1225768448_0001_r_000000_0: Counters: 30
        File System Counters
                FILE: Number of bytes read=1564
                FILE: Number of bytes written=650448
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=47
                HDFS: Number of bytes written=42
                HDFS: Number of read operations=10
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=6
                Reduce shuffle bytes=87
                Reduce input records=9
                Reduce output records=6
                Spilled Records=9
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=31
                Total committed heap usage (bytes)=317718528
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=42
2025-01-28 15:10:40,577 INFO mapred.LocalJobRunner: Finishing task:
attempt_local1225768448_0001_r_000000_0
2025-01-28 15:10:40,577 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-01-28 15:10:41,260 INFO mapreduce.Job:  map 100% reduce 100%
2025-01-28 15:10:41,260 INFO mapreduce.Job: Job job_local1225768448_0001 completed
successfully
2025-01-28 15:10:41,265 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=2922
                FILE: Number of bytes written=1300809
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=94
                HDFS: Number of bytes written=42
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=3
                Map output records=9
                Map output bytes=63
                Map output materialized bytes=87
                Input split bytes=100
                Combine input records=0
                Combine output records=0
                Reduce input groups=6
                Reduce shuffle bytes=87
                Reduce input records=9
                Reduce output records=6
                Spilled Records=18
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=31
                Total committed heap usage (bytes)=592969728
        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=47
        File Output Format Counters
                Bytes Written=42
2025-01-28 15:10:41,265 INFO streaming.StreamJob: Output directory: /160122771091/output3.txt
```