MAPPER

```
package bin;

import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

public class TopNMapper extends Mapper<Object, Text, Text, IntWritable> {
  private static final IntWritable one = new IntWritable(1);

  private Text word = new Text();

  private String tokens = "[_|$#<>\\^=\\[\\]\\*/\\\\,;,.\\-:()?!\"']";

  public void map(Object key, Text value, Mapper<Object, Text, Text, IntWritable>.Context
context) throws IOException, InterruptedException {
    String cleanLine = value.toString().toLowerCase().replaceAll(this.tokens, " ");
    StringTokenizer itr = new StringTokenizer(cleanLine);
    while (itr.hasMoreTokens()) {
      this.word.set(itr.nextToken().trim());
      context.write(this.word, one);
    }
  }
}
```

REDUCER

```
package bin;

import java.io.IOException;
import java.util.HashMap;
import java.util.Map;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
import utils.MiscUtils;

public class TopNReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
```

```
  private Map<Text, IntWritable> countMap = new HashMap<>();

  public void reduce(Text key, Iterable<IntWritable> values, Reducer<Text, IntWritable, Text,
IntWritable>.Context context) throws IOException, InterruptedException {
    int sum = 0;
    for (IntWritable val : values)
      sum += val.get();
    this.countMap.put(new Text(key), new IntWritable(sum));
  }

  protected void cleanup(Reducer<Text, IntWritable, Text, IntWritable>.Context context) throws
IOException, InterruptedException {
    Map<Text, IntWritable> sortedMap = MiscUtils.sortByValues(this.countMap);
    int counter = 0;
    for (Text key : sortedMap.keySet()) {
      if (counter++ == 20)
        break;
      context.write(key, sortedMap.get(key));
    }
  }
}
```

DRIVER

```
package bin;

import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;

public class TopN {
  public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    String[] otherArgs = (new GenericOptionsParser(conf, args)).getRemainingArgs();
```

```
    if (otherArgs.length != 2) {
      System.err.println("Usage: TopN <in> <out>");
      System.exit(2);
    }
    Job job = Job.getInstance(conf);
    job.setJobName("Top N");
    job.setJarByClass(TopN.class);
    job.setMapperClass(TopNMapper.class);
    job.setReducerClass(TopNReducer.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    FileInputFormat.addInputPath(job, new Path(otherArgs[0]));
    FileOutputFormat.setOutputPath(job, new Path(otherArgs[1]));
    System.exit(job.waitForCompletion(true) ? 0 : 1);
  }

  public static class TopNMapper extends Mapper<Object, Text, Text, IntWritable> {
    private static final IntWritable one = new IntWritable(1);

    private Text word = new Text();

    private String tokens = "[_|$#<>\\^=\\[\\]\\*/\\\\,;,.\\-:()?!\"']";

    public void map(Object key, Text value, Mapper<Object, Text, Text, IntWritable>.Context
context) throws IOException, InterruptedException {
      String cleanLine = value.toString().toLowerCase().replaceAll(this.tokens, " ");
      StringTokenizer itr = new StringTokenizer(cleanLine);
      while (itr.hasMoreTokens()) {
        this.word.set(itr.nextToken().trim());
        context.write(this.word, one);
      }
    }
  }
}
```

COMBINER

```
package bin;

import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
```

```
public class TopNCombiner extends Reducer<Text, IntWritable, Text, IntWritable> {
  public void reduce(Text key, Iterable<IntWritable> values, Reducer<Text, IntWritable, Text,
IntWritable>.Context context) throws IOException, InterruptedException {
    int sum = 0;
    for (IntWritable val : values)
      sum += val.get();
    context.write(key, new IntWritable(sum));
  }
}
```

```
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ cd TopN/
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228/TopN$ nano input.txt
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228/TopN$ hdfs dfs -mkdir /cs228/topn
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228/TopN$ hdfs dfs -put ~/eclipse-
workspace/cs228/TopN/input.txt /cs228/topn
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228/TopN$ hadoop jar TopN.jar bin.TopN
/cs228/topn/input.txt /cs228/topn/output
Not a valid JAR: /home/hduser/eclipse-workspace/cs228/TopN/TopN.jar
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228/TopN$ cd ..
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hadoop jar TopN.jar bin.TopN
/cs228/topn/input.txt /cs228/topn/output
22/06/22 15:45:42 INFO Configuration.deprecation: session.id is deprecated. Instead, use
dfs.metrics.session-id
22/06/22 15:45:42 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
22/06/22 15:45:42 INFO input.FileInputFormat: Total input paths to process : 1
22/06/22 15:45:42 INFO mapreduce.JobSubmitter: number of splits:1
22/06/22 15:45:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1869546441_0001
22/06/22 15:45:43 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
22/06/22 15:45:43 INFO mapreduce.Job: Running job: job_local1869546441_0001
22/06/22 15:45:43 INFO mapred.LocalJobRunner: OutputCommitter set in config null
22/06/22 15:45:43 INFO mapred.LocalJobRunner: OutputCommitter is
org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
22/06/22 15:45:43 INFO mapred.LocalJobRunner: Waiting for map tasks
22/06/22 15:45:43 INFO mapred.LocalJobRunner: Starting task: attempt_local1869546441_0001_m_000000_0
22/06/22 15:45:43 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
22/06/22 15:45:43 INFO mapred.MapTask: Processing split:
hdfs://localhost:54310/cs228/topn/input.txt:0+56
22/06/22 15:45:43 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
22/06/22 15:45:43 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
22/06/22 15:45:43 INFO mapred.MapTask: soft limit at 83886080
22/06/22 15:45:43 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
22/06/22 15:45:43 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
22/06/22 15:45:43 INFO mapred.MapTask: Map output collector class =
org.apache.hadoop.mapred.MapTask$MapOutputBuffer
22/06/22 15:45:43 INFO mapred.LocalJobRunner:
22/06/22 15:45:43 INFO mapred.MapTask: Starting flush of map output
22/06/22 15:45:43 INFO mapred.MapTask: Spilling map output
22/06/22 15:45:43 INFO mapred.MapTask: bufstart = 0; bufend = 100; bufvoid = 104857600
22/06/22 15:45:43 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214356(104857424);
length = 41/6553600
22/06/22 15:45:43 INFO mapred.MapTask: Finished spill 0
22/06/22 15:45:43 INFO mapred.Task: Task:attempt_local1869546441_0001_m_000000_0 is done. And is in the
process of committing
22/06/22 15:45:43 INFO mapred.LocalJobRunner: map
```

```
22/06/22 15:45:43 INFO mapred.LocalJobRunner: map
22/06/22 15:45:43 INFO mapred.Task: Task 'attempt_local1869546441_0001_m_000000_0' done.
22/06/22 15:45:43 INFO mapred.LocalJobRunner: Finishing task: attempt_local1869546441_0001_m_000000_0
22/06/22 15:45:43 INFO mapred.LocalJobRunner: map task executor complete.
22/06/22 15:45:43 INFO mapred.LocalJobRunner: Waiting for reduce tasks
22/06/22 15:45:43 INFO mapred.LocalJobRunner: Starting task: attempt_local1869546441_0001_r_000000_0
22/06/22 15:45:43 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
22/06/22 15:45:43 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin:
org.apache.hadoop.mapreduce.task.reduce.Shuffle@4928b3df
22/06/22 15:45:43 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464,
maxSingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10,
memToMemMergeOutputsThreshold=10
22/06/22 15:45:43 INFO reduce.EventFetcher: attempt_local1869546441_0001_r_000000_0 Thread started:
EventFetcher for fetching Map Completion Events
22/06/22 15:45:43 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map
attempt_local1869546441_0001_m_000000_0 decomp: 124 len: 128 to MEMORY
22/06/22 15:45:43 INFO reduce.InMemoryMapOutput: Read 124 bytes from map-output for
attempt_local1869546441_0001_m_000000_0
22/06/22 15:45:43 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 124,
inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->124
22/06/22 15:45:43 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
22/06/22 15:45:43 INFO mapred.LocalJobRunner: 1 / 1 copied.
22/06/22 15:45:43 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0
on-disk map-outputs
22/06/22 15:45:43 INFO mapred.Merger: Merging 1 sorted segments
22/06/22 15:45:43 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
115 bytes
22/06/22 15:45:43 INFO reduce.MergeManagerImpl: Merged 1 segments, 124 bytes to disk to satisfy reduce
memory limit
22/06/22 15:45:43 INFO reduce.MergeManagerImpl: Merging 1 files, 128 bytes from disk
22/06/22 15:45:43 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
22/06/22 15:45:43 INFO mapred.Merger: Merging 1 sorted segments
22/06/22 15:45:43 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
115 bytes
22/06/22 15:45:43 INFO mapred.LocalJobRunner: 1 / 1 copied.
22/06/22 15:45:43 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use
mapreduce.job.skiprecords
22/06/22 15:45:43 INFO mapred.Task: Task:attempt_local1869546441_0001_r_000000_0 is done. And is in the
process of committing
22/06/22 15:45:43 INFO mapred.LocalJobRunner: 1 / 1 copied.
22/06/22 15:45:43 INFO mapred.Task: Task attempt_local1869546441_0001_r_000000_0 is allowed to commit
now
22/06/22 15:45:43 INFO output.FileOutputCommitter: Saved output of task
'attempt_local1869546441_0001_r_000000_0' to
```

```
22/06/22 15:45:43 INFO output.FileOutputCommitter: Saved output of task
'attempt_local1869546441_0001_r_000000_0' to
hdfs://localhost:54310/cs228/topn/output/_temporary/0/task_local1869546441_0001_r_000000
22/06/22 15:45:43 INFO mapred.LocalJobRunner: reduce > reduce
22/06/22 15:45:43 INFO mapred.Task: Task 'attempt_local1869546441_0001_r_000000_0' done.
22/06/22 15:45:43 INFO mapred.LocalJobRunner: Finishing task: attempt_local1869546441_0001_r_000000_0
22/06/22 15:45:43 INFO mapred.LocalJobRunner: reduce task executor complete.
22/06/22 15:45:44 INFO mapreduce.Job: Job job_local1869546441_0001 running in uber mode : false
22/06/22 15:45:44 INFO mapreduce.Job:  map 100% reduce 100%
22/06/22 15:45:44 INFO mapreduce.Job: Job job_local1869546441_0001 completed successfully
22/06/22 15:45:44 INFO mapreduce.Job: Counters: 38
    File System Counters
        FILE: Number of bytes read=18032
        FILE: Number of bytes written=519208
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=112
        HDFS: Number of bytes written=59
        HDFS: Number of read operations=13
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=4
    Map-Reduce Framework
        Map input records=1
        Map output records=11
        Map output bytes=100
        Map output materialized bytes=128
        Input split bytes=108
        Combine input records=0
        Combine output records=0
        Reduce input groups=8
        Reduce shuffle bytes=128
        Reduce input records=11
        Reduce output records=8
        Spilled Records=22
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=42
        CPU time spent (ms)=0
        Physical memory (bytes) snapshot=0
        Virtual memory (bytes) snapshot=0
        Total committed heap usage (bytes)=577241088
    Shuffle Errors
```

```
        Combine input records=0
        Combine output records=0
        Reduce input groups=8
        Reduce shuffle bytes=128
        Reduce input records=11
        Reduce output records=8
        Spilled Records=22
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=42
        CPU time spent (ms)=0
        Physical memory (bytes) snapshot=0
        Virtual memory (bytes) snapshot=0
        Total committed heap usage (bytes)=577241088
    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=56
    File Output Format Counters
        Bytes Written=59
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hdfs dfs -cat /cs228/topn/output/part-00000
cat: `/cs228/topn/output/part-00000': No such file or directory
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hdfs dfs -ls /cs228/topn/output
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2022-06-22 15:45 /cs228/topn/output/_SUCCESS
-rw-r--r--   1 hduser supergroup         59 2022-06-22 15:45 /cs228/topn/output/part-r-00000
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hdfs dfs -cat /cs228/topn/output/part-r-00000
this    3
is  2
world   1
me  1
there   1
change  1
hello   1
pranav  1
```