Question 5. Develop a MapReduce program to count the number of occurrences of words in a given file.

WCMapper.java

```
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reporter;

public class WCMapper extends MapReduceBase implements Mapper<LongWritable,
                                    Text, Text, IntWritable> {

  // Map function
  public void map(LongWritable key, Text value, OutputCollector<Text,
          IntWritable> output, Reporter rep) throws IOException
  {

    String line = value.toString();

    // Splitting the line on spaces
    for (String word : line.split(" "))
    {
      if (word.length() > 0)
      {
        output.collect(new Text(word), new IntWritable(1));
      }
    }
  }
}
```

WCReducer.java

```
import java.io.IOException;
import java.util.Iterator;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.OutputCollector;
```

```
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;

public class WCReducer extends MapReduceBase implements Reducer<Text,
                                                                 IntWritable, Text,
IntWritable> {

        // Reduce function
        public void reduce(Text key, Iterator<IntWritable> value,
                                OutputCollector<Text, IntWritable> output,
                                                        Reporter rep) throws IOException
        {

                int count = 0;

                // Counting the frequency of each words
                while (value.hasNext())
                {
                        IntWritable i = value.next();
                        count += i.get();
                }

                output.collect(key, new IntWritable(count));
        }
}
```

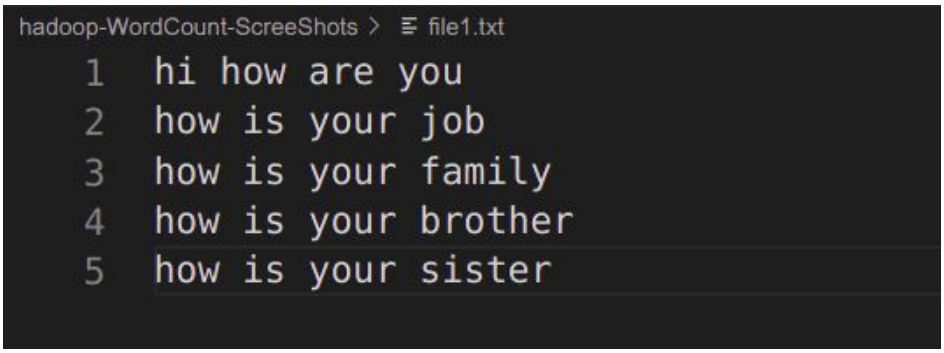
WCDriver.java

```
import java.io.IOException;
import org.apache.hadoop.conf.Configured;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.FileInputFormat;
import org.apache.hadoop.mapred.FileOutputFormat;
import org.apache.hadoop.mapred.JobClient;
import org.apache.hadoop.mapred.JobConf;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;

public class WCDriver extends Configured implements Tool {
```

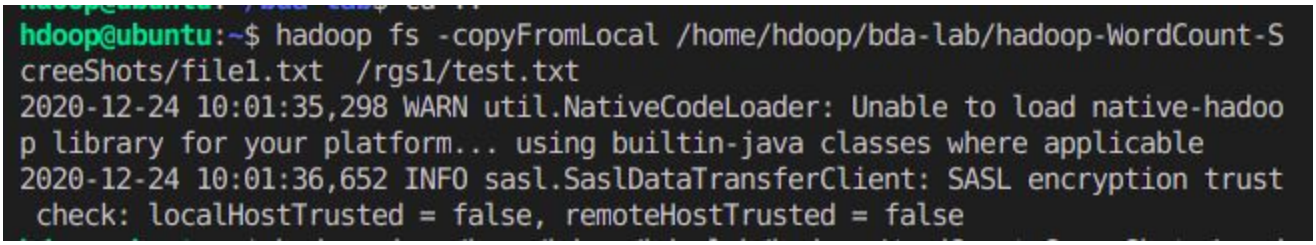
```
	public int run(String args[]) throws IOException
	{
		if (args.length < 2)
		{
			System.out.println("Please give valid inputs");
			return -1;
		}

		JobConf conf = new JobConf(WCDriver.class);
		FileInputFormat.setInputPaths(conf, new Path(args[0]));
		FileOutputFormat.setOutputPath(conf, new Path(args[1]));
		conf.setMapperClass(WCMapper.class);
		conf.setReducerClass(WCReducer.class);
		conf.setMapOutputKeyClass(Text.class);
		conf.setMapOutputValueClass(IntWritable.class);
		conf.setOutputKeyClass(Text.class);
		conf.setOutputValueClass(IntWritable.class);
		JobClient.runJob(conf);
		return 0;
	}

	// Main Method
	public static void main(String args[]) throws Exception
	{
		int exitCode = ToolRunner.run(new WCDriver(), args);
		System.out.println(exitCode);
	}
}
```

HADOOP COMMANDS :

file1.txt



hadoop fs -copyFromLocal /home/hdoop/bda-lab/hadoop-WordCount-ScreeShots/file1.txt /rgs1/test.txt



hadoop jar /home/hdoop/bda-lab/hadoop-WordCount-ScreeShots/wordcount.jar WordCount /rgs1/test.txt /rgs1/output

```
hdoop@ubuntu:~$ hadoop jar /home/hdoop/bda-lab/hadoop-WordCount-ScreeShots/wordcount.jar WordCount  /rgs1/te
st.txt /rgs1/output
2020-12-24 10:17:06,637 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform..
. using builtin-java classes where applicable
2020-12-24 10:17:07,156 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2020-12-24 10:17:07,242 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2020-12-24 10:17:07,242 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2020-12-24 10:17:07,453 INFO input.FileInputFormat: Total input files to process : 1
2020-12-24 10:17:08,160 INFO mapreduce.JobSubmitter: number of splits:1
2020-12-24 10:17:08,763 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1012990088_0001
2020-12-24 10:17:08,763 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-12-24 10:17:08,977 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2020-12-24 10:17:08,978 INFO mapreduce.Job: Running job: job_local1012990088_0001
2020-12-24 10:17:09,013 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2020-12-24 10:17:09,026 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2020-12-24 10:17:09,026 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders
 under output directory:false, ignore cleanup failures: false
2020-12-24 10:17:09,026 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.outpu
t.FileOutputCommitter
2020-12-24 10:17:09,272 INFO mapred.LocalJobRunner: Waiting for map tasks
2020-12-24 10:17:09,273 INFO mapred.LocalJobRunner: Starting task: attempt_local1012990088_0001_m_000000_0
2020-12-24 10:17:09,334 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2020-12-24 10:17:09,334 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders
 under output directory:false, ignore cleanup failures: false
2020-12-24 10:17:09,423 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2020-12-24 10:17:09,426 INFO mapred.MapTask: Processing split: hdfs://127.0.0.1:9000/rgs1/test.txt:0+88
2020-12-24 10:17:09,601 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2020-12-24 10:17:09,601 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2020-12-24 10:17:09,601 INFO mapred.MapTask: soft limit at 83886080
2020-12-24 10:17:09,601 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2020-12-24 10:17:09,601 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2020-12-24 10:17:09,605 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$M
apOutputBuffer
2020-12-24 10:17:09,693 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = fa
lse, remoteHostTrusted = false
2020-12-24 10:17:10,001 INFO mapreduce.Job: Job job_local1012990088_0001 running in uber mode : false
2020-12-24 10:17:10,004 INFO mapreduce.Job:  map 0% reduce 0%
2020-12-24 10:17:10,047 INFO mapred.LocalJobRunner:
2020-12-24 10:17:10,055 INFO mapred.MapTask: Starting flush of map output
2020-12-24 10:17:10,057 INFO mapred.MapTask: Spilling map output
2020-12-24 10:17:10,057 INFO mapred.MapTask: bufstart = 0; bufend = 169; bufvoid = 104857600
2020-12-24 10:17:10,058 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214320(104857280); len
gth = 77/6553600
2020-12-24 10:17:10,144 INFO mapred.MapTask: Finished spill 0
2020-12-24 10:17:10,169 INFO mapred.Task: Task:attempt_local1012990088_0001_m_000000_0 is done. And is in th
e process of committing
2020-12-24 10:17:10,186 INFO mapred.LocalJobRunner: map
2020-12-24 10:17:10,186 INFO mapred.Task: Task 'attempt_local1012990088_0001_m_000000_0' done.
2020-12-24 10:17:10,198 INFO mapred.Task: Final Counters for attempt_local1012990088_0001_m_000000_0: Counte
```

---

hadoop fs -ls /rgs1/output/part-r-00000

```
hdoop@ubuntu:~$ hadoop fs -cat /rgs1/output/part-r-00000
2020-12-24 10:20:13,595 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform..
. using builtin-java classes where applicable
2020-12-24 10:20:14,272 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = fa
lse, remoteHostTrusted = false
are     1
brother 1
family  1
hi      1
how     5
is      4
job     1
sister  1
you     1
your    4
hdoop@ubuntu:~$
```