NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR

Cachar, Assam

B.Tech. VIIIth Sem

Subject Code: CS-484

Subject Name: Cloud Computing

Submitted By:

Name : Subhojit Ghimire

Sch. Id. : 1912160

Branch : CSE – B

Q. Write a MapReduce program to count k-mers (28-mers or 55-mers) of a DNA sequence.

➔ **Filename: KmerCount.java**

```
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.*;
import org.apache.hadoop.mapreduce.*;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class KmerCount {

 public static class KmerMapper extends Mapper<LongWritable, Text, Text, IntWritable> {

  private final static IntWritable one = new IntWritable(1);
  private Text kmer = new Text();

  public void map(LongWritable key, Text value, Context context) throws IOException,
       InterruptedException {

   String line = value.toString().toUpperCase();
   int k = 28;
   //int k = 55; depending whether it is for 28 or 55
   // Loop over the line and extract k-mers
   for (int i = 0; i <= line.length() - k; i++) {
    String kmerStr = line.substring(i, i + k);
    kmer.set(kmerStr);
    context.write(kmer, one);
   }
  }
 }

 public static class KmerReducer extends Reducer<Text, IntWritable, Text, IntWritable> {

  private IntWritable result = new IntWritable();

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

```
  }
 }

 public static void main(String[] args) throws Exception {
  Configuration conf = new Configuration();
  Job job = Job.getInstance(conf, "kmer count");
  job.setJarByClass(KmerCount.class);
  job.setMapperClass(KmerMapper.class);
  job.setCombinerClass(KmerReducer.class);
  job.setReducerClass(KmerReducer.class);
  job.setOutputKeyClass(Text.class);
  job.setOutputValueClass(IntWritable.class);
  FileInputFormat.addInputPath(job, new Path(args[0]));
  FileOutputFormat.setOutputPath(job, new Path(args[1]));
  int k = Integer.parseInt(args[2]);
  job.getConfiguration().setInt("k", k);
  job.waitForCompletion(true);
 }
}
```

During executing, the format should be:
hadoop jar <jarAddress>.jar KmerCount <inputAddress> <outputAddress> <k-mers size arg>

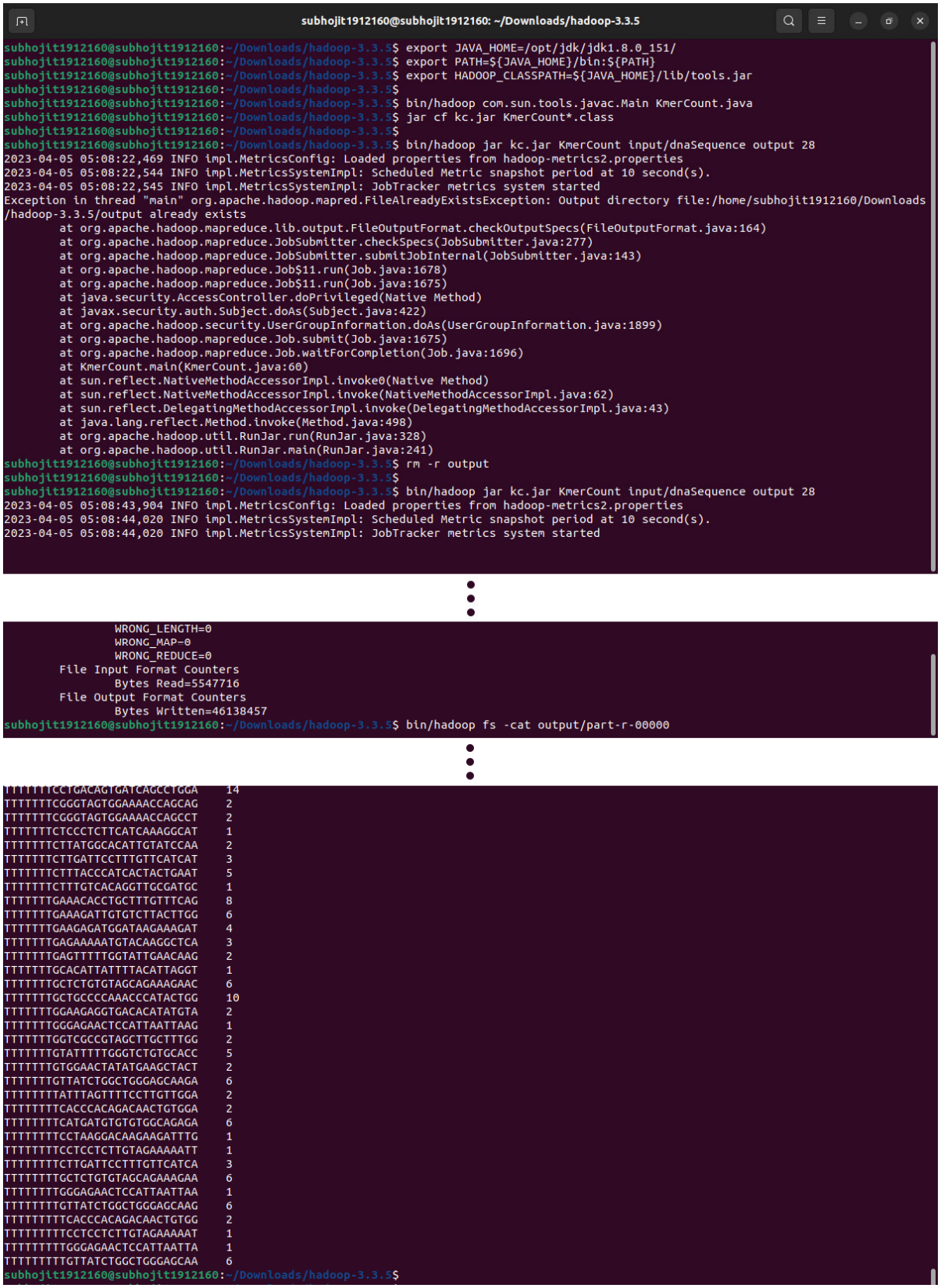
Foldername/Filename: /input/dnaSequence/human.txt
(DNA dataset: https://www.kaggle.com/datasets/nageshsingh/dna-sequence-dataset)

```
sequence        class
ATGCCCCAACTAAATACTACCGTATGGCCCACCATAATTACCCCCATACTCCTTACACTATTCCTCATC
      ACCCAACTAAAAATATTAAACACAAACTACCACCTACCTCCCTCACCAAAGCCCATAAAAATAAA
      AAATTATAACAAACCCTGAGAACCAAAATGAACGAAAATCTGTTCGCTTCATTCATTGCCCCCAC
      AATCCTAG…
… (5,547,716 characters long)
```

Execution:

```
$ bin/hadoop com.sun.tools.javac.Main KmerCount.java
$ jar cf kc.jar KmerCount*.class
$ bin/hadoop jar kc.jar KmerCount input/dnaSequence output 28
$ $ bin/hadoop fs -cat output/part-r-00000
```

Output:


subhojit1912160@subhojit1912160: ~/Downloads/hadoop-3.3.5
subhojit1912160@subhojit1912160:~/Downloads/hadoop-3.3.5$ export JAVA_HOME=/opt/jdk/jdk1.8.0_151/
subhojit1912160@subhojit1912160:~/Downloads/hadoop-3.3.5$ export PATH=${JAVA_HOME}/bin:${PATH}
subhojit1912160@subhojit1912160:~/Downloads/hadoop-3.3.5$ export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
subhojit1912160@subhojit1912160:~/Downloads/hadoop-3.3.5$
subhojit1912160@subhojit1912160:~/Downloads/hadoop-3.3.5$ bin/hadoop com.sun.tools.javac.Main KmerCount.java
subhojit1912160@subhojit1912160:~/Downloads/hadoop-3.3.5$ jar cf kc.jar KmerCount*.class
subhojit1912160@subhojit1912160:~/Downloads/hadoop-3.3.5$
subhojit1912160@subhojit1912160:~/Downloads/hadoop-3.3.5$ bin/hadoop jar kc.jar KmerCount input/dnaSequence output 28
2023-04-05 05:08:22,469 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-04-05 05:08:22,544 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2023-04-05 05:08:22,545 INFO impl.MetricsSystemImpl: JobTracker metrics system started
Exception in thread "main" org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory file:/home/subhojit1912160/Downloads
/hadoop-3.3.5/output already exists
at org.apache.hadoop.mapreduce.lib.output.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:164)
at org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:277)
at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:143)
at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1678)
at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1675)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1899)
at org.apache.hadoop.mapreduce.Job.submit(Job.java:1675)
at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1696)
at KmerCount.main(KmerCount.java:60)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:328)
at org.apache.hadoop.util.RunJar.main(RunJar.java:241)
subhojit1912160@subhojit1912160:~/Downloads/hadoop-3.3.5$ rm -r output
subhojit1912160@subhojit1912160:~/Downloads/hadoop-3.3.5$
subhojit1912160@subhojit1912160:~/Downloads/hadoop-3.3.5$ bin/hadoop jar kc.jar KmerCount input/dnaSequence output 28
2023-04-05 05:08:43,904 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-04-05 05:08:44,020 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2023-04-05 05:08:44,020 INFO impl.MetricsSystemImpl: JobTracker metrics system started
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=5547716
File Output Format Counters
Bytes Written=46138457
subhojit1912160@subhojit1912160:~/Downloads/hadoop-3.3.5$ bin/hadoop fs -cat output/part-r-00000
TTTTTTTCCTGACAGTGATCAGCCTGGA 14
TTTTTTTCGGGTAGTGGAAAACCAGCAG 2
TTTTTTTCGGGTAGTGGAAAACCAGCCT 2
TTTTTTTCTCCCTCTTCATCAAAGGCAT 1
TTTTTTTCTTATGGCACATTGTATCCAA 2
TTTTTTTCTTGATTCCTTTGTTCATCAT 3
TTTTTTTCTTTACCCATCACTACTGAAT 5
TTTTTTTCTTTGTCACAGGTTGCGATGC 1
TTTTTTTGAAACACCTGCTTTGTTTCAG 8
TTTTTTTGAAAGATTGTGTCTTACTTGG 6
TTTTTTTGAAGAGATGGATAAGAAAGAT 4
TTTTTTTGAGAAAAATGTACAAGGCTCA 3
TTTTTTTGAGTTTTTGGTATTGAACAAG 2
TTTTTTTGCACATTATTTTACATTAGGT 1
TTTTTTTGCTCTGTGTAGCAGAAAGAAC 6
TTTTTTTGCTGCCCCAAACCCATACTGG 10
TTTTTTTGGAAGAGGTGACACATATGTA 2
TTTTTTTGGGAGAACTCCATTAATTAAG 1
TTTTTTTGGTCGCCGTAGCTTGCTTTGG 2
TTTTTTTGTATTTTTGGGTCTGTGCACC 5
TTTTTTTGTGGAACTATATGAAGCTACT 2
TTTTTTTGTTATCTGGCTGGGAGCAAGA 6
TTTTTTTTATTTAGTTTTCCTTGTTGGA 2
TTTTTTTTCACCCACAGACAACTGTGGA 2
TTTTTTTTCATGATGTGTGTGGCAGAGA 6
TTTTTTTTCCTAAGGACAAGAAGATTTG 1
TTTTTTTTCCTCCTCTTGTAGAAAAATT 1
TTTTTTTTCTTGATTCCTTTGTTCATCA 3
TTTTTTTTGCTCTGTGTAGCAGAAAGAA 6
TTTTTTTTGGGAGAACTCCATTAATTAA 1
TTTTTTTTGTTATCTGGCTGGGAGCAAG 6
TTTTTTTTTCACCCACAGACAACTGTGG 2
TTTTTTTTTCCTCCTCTTGTAGAAAAAT 1
TTTTTTTTTGGGAGAACTCCATTAATTA 1
TTTTTTTTTGTTATCTGGCTGGGAGCAA 6
subhojit1912160@subhojit1912160:~/Downloads/hadoop-3.3.5$