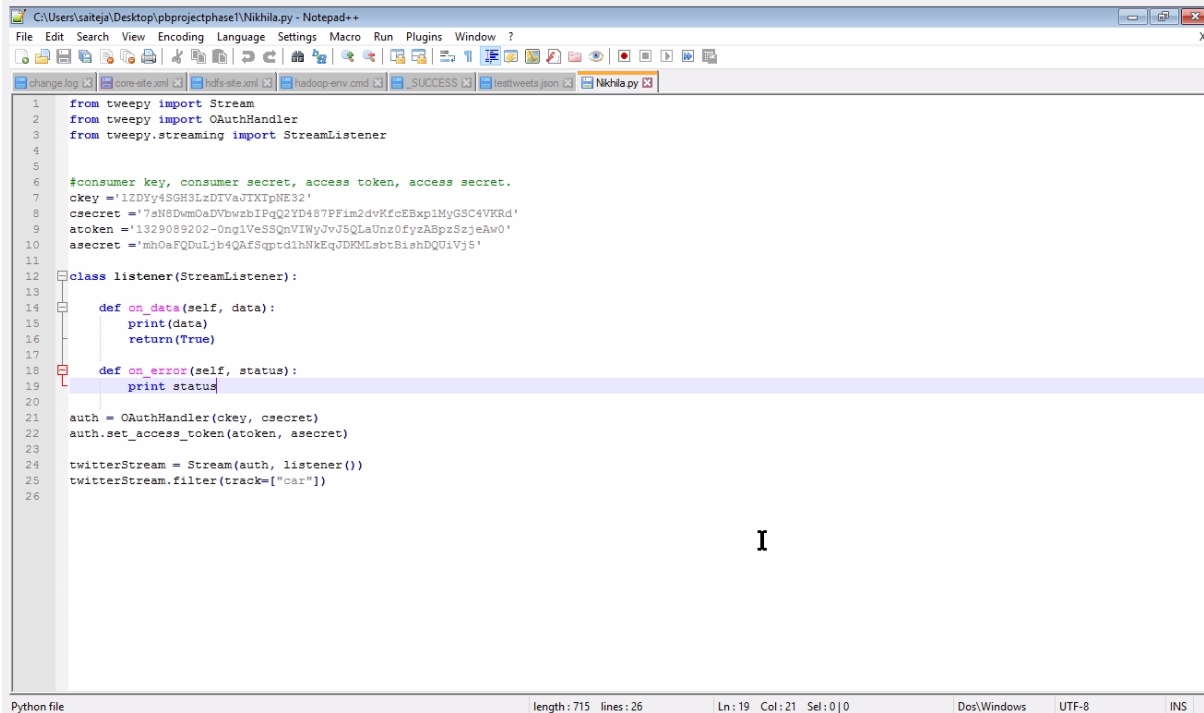
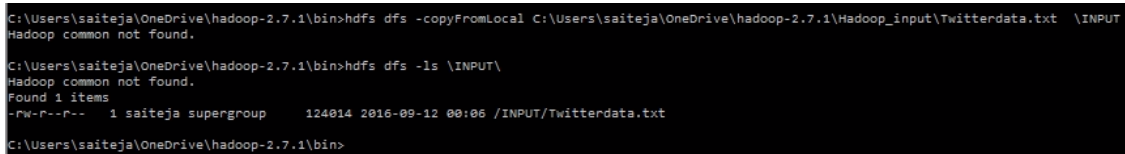


Installed apache hadoop2.7.1 , to get streaming data from twitter we used the below program and captured enough data for phase 1. Python program file is attached to the zip folder.



```
1 from tweepy import Stream
2 from tweepy import OAuthHandler
3 from tweepy.streaming import StreamListener
4
5
6 #consumer key, consumer secret, access token, access secret.
7 ckey = '1ZDYy4SGH3LzDTVaJIXTpNE32'
8 csecret = '7sN8Dwm0aDvbwzbIPqQ2YD487FFIm2dvKfcEBxp1MyGSC4VERd'
9 atoken = '1329089202-0ng1VeSSQnVIWYJvJ5QLaUnz0fyzABpzSzeAw0'
10 asecret = 'mh0aFQDuLjb4QAfSqptd1hNkEqJDRMLsbtBishDQUiVj5'
11
12 class listener(StreamListener):
13
14     def on_data(self, data):
15         print(data)
16         return(True)
17
18     def on_error(self, status):
19         print status
20
21 auth = OAuthHandler(ckey, csecret)
22 auth.set_access_token(atoken, asecret)
23
24 twitterStream = Stream(auth, listener())
25 twitterStream.filter(track=["car"])
26
```

Uploaded the extracted twitter tweets date into Hadoop HDFS system, below screen short shows the process.

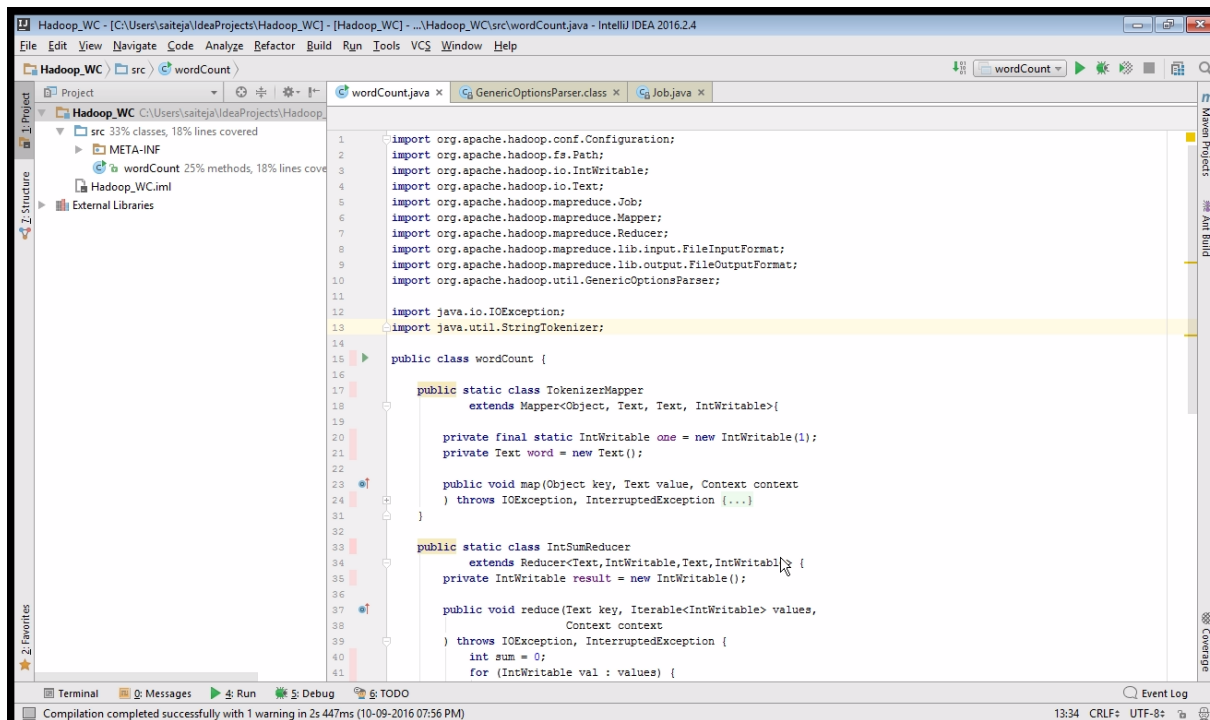


```
C:\Users\saitaja\OneDrive\hadoop-2.7.1\bin>hdfs dfs -copyFromLocal C:\Users\saitaja\OneDrive\hadoop-2.7.1\Hadoop_input\Twitterdata.txt \INPUT
Hadoop common not found.

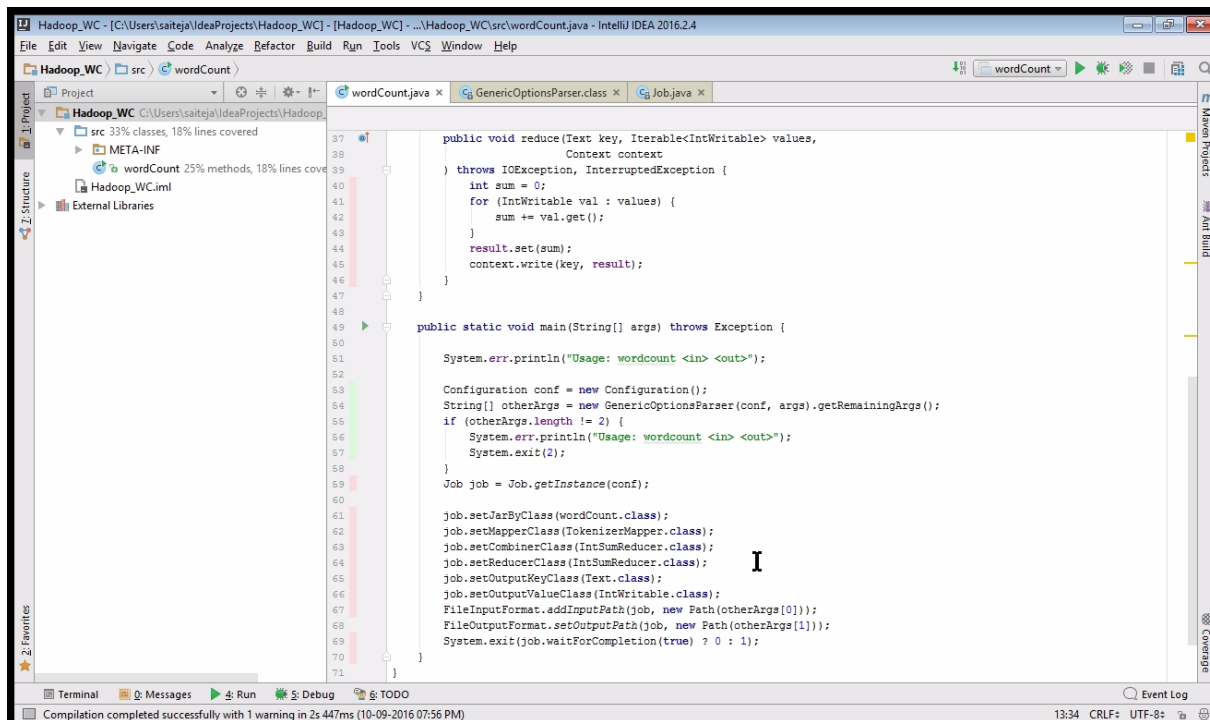
C:\Users\saitaja\OneDrive\hadoop-2.7.1\bin>hdfs dfs -ls \INPUT\
Hadoop common not found.
Found 1 items
-rw-r--r-- 1 saiteja supergroup 124014 2016-09-12 00:06 /INPUT/Twitterdata.txt

C:\Users\saitaja\OneDrive\hadoop-2.7.1\bin>
```

Below shows the java code for wordCount. We installed IntelliJ IDEA IDE and written program in it. Created the jar file for this program so that Hadoop can execute the program from CMD. program is attached to the zip folder.



```
1 import org.apache.hadoop.conf.Configuration;
2 import org.apache.hadoop.fs.Path;
3 import org.apache.hadoop.io.IntWritable;
4 import org.apache.hadoop.io.Text;
5 import org.apache.hadoop.mapreduce.Job;
6 import org.apache.hadoop.mapreduce.Mapper;
7 import org.apache.hadoop.mapreduce.Reducer;
8 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
9 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
10 import org.apache.hadoop.util.GenericOptionsParser;
11
12 import java.io.IOException;
13 import java.util.StringTokenizer;
14
15 public class wordCount {
16
17     public static class TokenizerMapper
18         extends Mapper<Object, Text, Text, IntWritable>{
19
20         private final static IntWritable one = new IntWritable(1);
21         private Text word = new Text();
22
23         public void map(Object key, Text value, Context context
24             ) throws IOException, InterruptedException {...}
25     }
26
27     public static class IntSumReducer
28         extends Reducer<Text, IntWritable, Text, IntWritable> {
29         private IntWritable result = new IntWritable();
30
31         public void reduce(Text key, Iterable<IntWritable> values,
32             Context context
33             ) throws IOException, InterruptedException {
34             int sum = 0;
35             for (IntWritable val : values) {
```



```
36
37         public void reduce(Text key, Iterable<IntWritable> values,
38             Context context
39             ) throws IOException, InterruptedException {
40             int sum = 0;
41             for (IntWritable val : values) {
42                 sum += val.get();
43             }
44             result.set(sum);
45             context.write(key, result);
46         }
47     }
48
49     public static void main(String[] args) throws Exception {
50
51         System.err.println("Usage: wordcount <in> <out>");
52
53         Configuration conf = new Configuration();
54         String[] otherArgs = new GenericOptionsParser(conf, args).getRemainingArgs();
55         if (otherArgs.length != 2) {
56             System.err.println("Usage: wordcount <in> <out>");
57             System.exit(2);
58         }
59         Job job = Job.getInstance(conf);
60
61         job.setJarByClass(wordCount.class);
62         job.setMapperClass(TokenizerMapper.class);
63         job.setCombinerClass(IntSumReducer.class);
64         job.setReducerClass(IntSumReducer.class);
65         job.setOutputKeyClass(Text.class);
66         job.setOutputValueClass(IntWritable.class);
67         FileInputFormat.addInputPath(job, new Path(otherArgs[0]));
68         FileOutputFormat.setOutputPath(job, new Path(otherArgs[1]));
69         System.exit(job.waitForCompletion(true) ? 0 : 1);
70     }
71 }
```

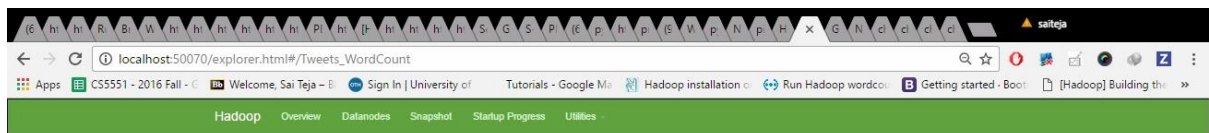
```
C:\Users\saitetja\OneDrive\Hadoop>2.7.1\bin\hadoop jar C:\Users\saitetja\IdeaProjects\Hadoop\Output\artifacts\Hadoop_MC_jar\Hadoop_MC_jar\INPUT\Twitterdata.txt /Twitter_WordCount
Usage: wordcount <in> <out>
16/09/12 00:21:07 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/09/12 00:21:07 INFO java.util.concurrent: Initializing JVM Metrics with processName=JobTracker, sessionID=
16/09/12 00:21:08 INFO Input.FileInputFormat: Total input paths to process : 1
16/09/12 00:21:08 INFO mapreduce.JobSubmitter: number of splits:1
16/09/12 00:21:09 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1661323393_0001
16/09/12 00:21:10 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/09/12 00:21:10 INFO mapreduce.Job: Running Job: job_local1661323393_0001
16/09/12 00:21:10 INFO mapreduce.LocalJobRunner: OutputCommitter set in config null
16/09/12 00:21:10 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/09/12 00:21:10 INFO mapreduce.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
16/09/12 00:21:11 INFO mapreduce.LocalJobRunner: Waiting for map tasks
16/09/12 00:21:10 INFO mapreduce.LocalJobRunner: Starting task: attempt_local1661323393_0001_m_000000_e
16/09/12 00:21:10 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/09/12 00:21:10 INFO util.ProcessBasedProcessTree: ProcessBasedProcessTree currently is supported only on Linux.
16/09/12 00:21:11 INFO mapreduce.Job: Job: job_local1661323393_0001 running in uber mode : false
16/09/12 00:21:11 INFO mapreduce.Job: map 0% reduce 0%
16/09/12 00:21:11 INFO mapreduce.Task: Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBasedProcessTree@575d26b
16/09/12 00:21:11 INFO mapreduce.HadoopMapTask: Processing split: hdfs://localhost:19000/INPUT/Twitterdata.txt:124814
16/09/12 00:21:11 INFO mapreduce.HadoopMapTask: (Equation) = k/1 2621346(18465758)
16/09/12 00:21:11 INFO mapreduce.HadoopMapTask: mapreduce.task.io.sort.mb: 100
16/09/12 00:21:11 INFO mapreduce.HadoopMapTask: sort limit at 83866000
16/09/12 00:21:11 INFO mapreduce.HadoopMapTask: bursort = 0, bufoverid = 104557600
16/09/12 00:21:11 INFO mapreduce.HadoopMapTask: kvstart = 26214396; length = 655360
16/09/12 00:21:11 INFO mapreduce.HadoopMapTask: mapreduce.task.collector.class = org.apache.hadoop.mapreduce.HadoopMapTaskOutputBuffer
16/09/12 00:21:11 INFO mapreduce.LocalJobRunner:
16/09/12 00:21:13 INFO mapreduce.HadoopMapTask: Starting flush of map output
16/09/12 00:21:13 INFO mapreduce.HadoopMapTask: Spilling map output
16/09/12 00:21:13 INFO mapreduce.HadoopMapTask: bursort = 0, bufoverid = 104557600
16/09/12 00:21:13 INFO mapreduce.HadoopMapTask: kvstart = 2621346(18465758); kvend = 26214072(184616289); length = 60321/655360
16/09/12 00:21:13 INFO mapreduce.HadoopMapTask: Finished spill 0
16/09/12 00:21:13 INFO mapreduce.Task: Taskattempt_local1661323393_0001_m_000000_e is done. And is in the process of committing
16/09/12 00:21:13 INFO mapreduce.LocalJobRunner: map
16/09/12 00:21:13 INFO mapreduce.Task: Task attempt_local1661323393_0001_m_000000_e done.
16/09/12 00:21:13 INFO mapreduce.LocalJobRunner: Finishing task: attempt_local1661323393_0001_m_000000_e
16/09/12 00:21:13 INFO mapreduce.LocalJobRunner: map task executor complete.
16/09/12 00:21:14 INFO mapreduce.LocalJobRunner: Waiting for reduce tasks
16/09/12 00:21:14 INFO output.FileOutputCommitter: Starting task: attempt_local1661323393_0001_m_000000_e
16/09/12 00:21:14 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/09/12 00:21:14 INFO util.ProcessBasedProcessTree: ProcessBasedProcessTree currently is supported only on Linux.
16/09/12 00:21:14 INFO mapreduce.Task: Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBasedProcessTree@8a183bf
16/09/12 00:21:14 INFO mapreduce.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.ShuffleMap483ca727
16/09/12 00:21:14 INFO reduce.HergeManagerImpl: HergeManager: memoryLimit=333971456, maxSingleShuffleLimit=85492864, mergeThreshold=220631169, ioSortFactor=16, memToMergeOutputsThreshold=18
16/09/12 00:21:14 INFO reduce.HergeManagerImpl: HergeManager: memoryLimit=333971456, maxSingleShuffleLimit=85492864, mergeThreshold=220631169, ioSortFactor=16, memToMergeOutputsThreshold=18
16/09/12 00:21:14 INFO reduce.LocalFetcher: LocalFetcher#1 about to supply output of map attempt_local1661323393_0001_m_000000_e compEvent: 88344 len: 88345 to MEMORY
16/09/12 00:21:14 INFO reduce.InMemoryMapOutput: Read 88344 bytes from map-output for attempt_local1661323393_0001_m_000000_e
16/09/12 00:21:14 INFO reduce.HergeManagerImpl: closeMemoryInput() = map-output of size: 88344, inMemoryMapOutputs.size() = 1, compMemory -> 0, usedMemory -> 88344
16/09/12 00:21:14 INFO reduce.HergeManagerImpl: HergeManager: memoryLimit=333971456, maxSingleShuffleLimit=85492864, mergeThreshold=220631169, ioSortFactor=16, memToMergeOutputsThreshold=18
16/09/12 00:21:14 INFO mapreduce.LocalJobRunner: 1 / 1 copied.
16/09/12 00:21:14 INFO reduce.HergeManagerImpl: FinalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/09/12 00:21:14 INFO mapreduce.HergeManagerImpl: HergeManager: Merging 1 sorted segments
16/09/12 00:21:14 INFO mapreduce.HergeManagerImpl: Down to the last merge-pass, with 1 segments left of total size: 88344 bytes
16/09/12 00:21:14 INFO reduce.HergeManagerImpl: Merged 1 segments, 88344 bytes to disk to satisfy reduce memory limit
16/09/12 00:21:14 INFO reduce.HergeManagerImpl: Merging 1 files, 88344 bytes from disk
16/09/12 00:21:14 INFO reduce.HergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/09/12 00:21:14 INFO mapreduce.HergeManagerImpl: Merging 1 sorted segments
16/09/12 00:21:14 INFO mapreduce.HergeManagerImpl: Down to the last merge-pass, with 1 segments left of total size: 88344 bytes
16/09/12 00:21:14 INFO mapreduce.LocalJobRunner: 1 / 1 copied.
16/09/12 00:21:14 INFO mapreduce.Job: map 100% reduce 0%
```

```

16/09/12 00:21:14 INFO mapreduce.Job: map 100% reduce 0%
16/09/12 00:21:14 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
16/09/12 00:21:20 INFO mapred.LocalJobRunner: reduce > reduce
16/09/12 00:21:20 INFO mapreduce.Job: map 100% reduce 100%
16/09/12 00:21:21 INFO mapred.Task: Task:attempt_local1661323393_0001_r_000000_0 is done. And is in the process of committing
16/09/12 00:21:21 INFO mapred.LocalJobRunner: reduce > reduce
16/09/12 00:21:21 INFO mapred.Task: Task attempt_local1661323393_0001_r_000000_0 is allowed to commit now
16/09/12 00:21:21 INFO FileOutputCommitter: Saved output of task 'attempt_local1661323393_0001_r_000000_0' to hdfs://localhost:9000/Tweets_WordCount/_temporary/0/task_local1661323393_0001_r_000000
16/09/12 00:21:21 INFO mapred.LocalJobRunner: reduce > reduce
16/09/12 00:21:21 INFO mapred.Task: Task 'attempt_local1661323393_0001_r_000000_0' done.
16/09/12 00:21:21 INFO mapred.LocalJobRunner: Finishing task: attempt_local1661323393_0001_r_000000_0
16/09/12 00:21:21 INFO mapred.LocalJobRunner: reduce task execution complete
16/09/12 00:21:22 INFO mapreduce.Job: Job Job_local1661323393_0001 completed successfully
16/09/12 00:21:22 INFO mapreduce.Job: Counters: 35
File System Counters
  FILE: Number of bytes read=6358654
  FILE: Number of bytes written=7856342
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=248825
  HDFS: Number of bytes written=67706
  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=5253
  Map output records=15802
  Map output bytes=179759
  Map output materialized bytes=88345
  Input split bytes=108
  Combine input records=15802
  Combine output records=5206
  Reduce input groups=5206
  Reduce shuffle bytes=88345
  Reduce input records=5206
  Reduce output records=5206
  Spilled Records=10412
  Shuffled Hops=1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=12
  Total committed heap usage (bytes)=557842432
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=124014
File Output Format Counters
  Bytes Written=67706
C:\Users\saitaja\OneDrive\hadoop-2.7.1\bin>hdfs dfs -ls /Tweets_WordCount/*
hadoop common not found.
#W-r-r-r-- 1 saiteja supergroup          0 2016-09-12 00:21 /Tweets_WordCount/_SUCCESS
#W-r-r-r-- 1 saiteja supergroup        67706 2016-09-12 00:21 /Tweets_WordCount/part-r-000000
C:\Users\saitaja\OneDrive\hadoop-2.7.1\bin>

```

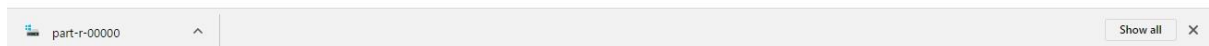
Below showing the log files from Hadoop localhost and the same attached to zip folder.



## Browse Directory

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	saiteja	supergroup	0 B	12/09/2016, 00:21:22	1	128 MB	_SUCCESS
-rw-r--r--	saiteja	supergroup	66.12 KB	12/09/2016, 00:21:21	1	128 MB	part-r-00000

Hadoop, 2015.



Output file is opened and shown word and number of occurrences.

