Data Science

hw09 . Map Reduce Programming

2017년 5월 24일

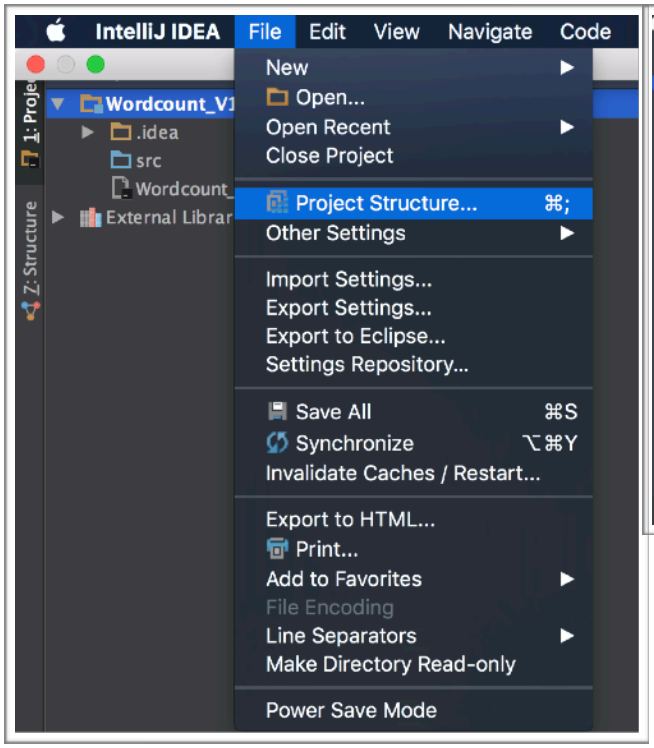
00 반
충남대학교 컴퓨터공학과
201202154
조윤재

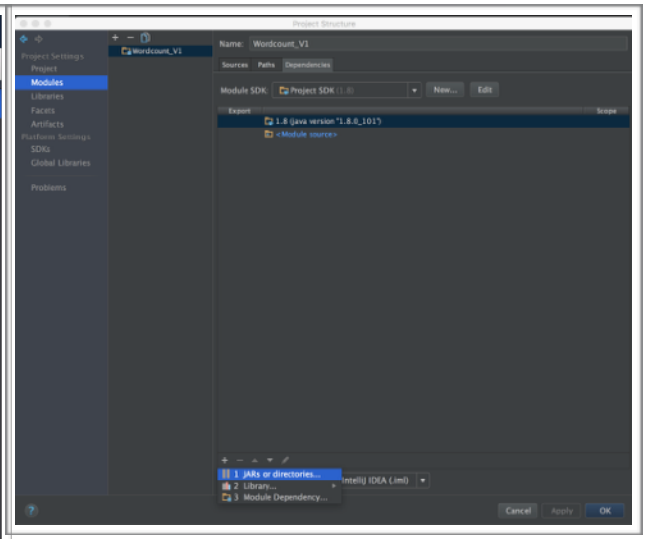
❖ Contents.

1. Map Reduce Programming Setting & Word Count Version 1

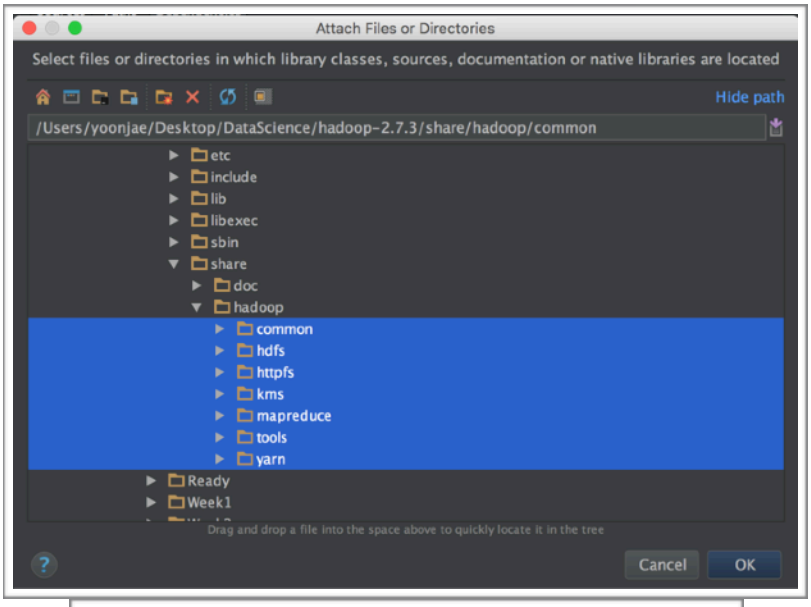
Java 컴파일은 다음과 같은 환경에서 진행하였습니다.
- OS : Mac OS Sierra Ver 10.12.4
- Editor : IntelliJ

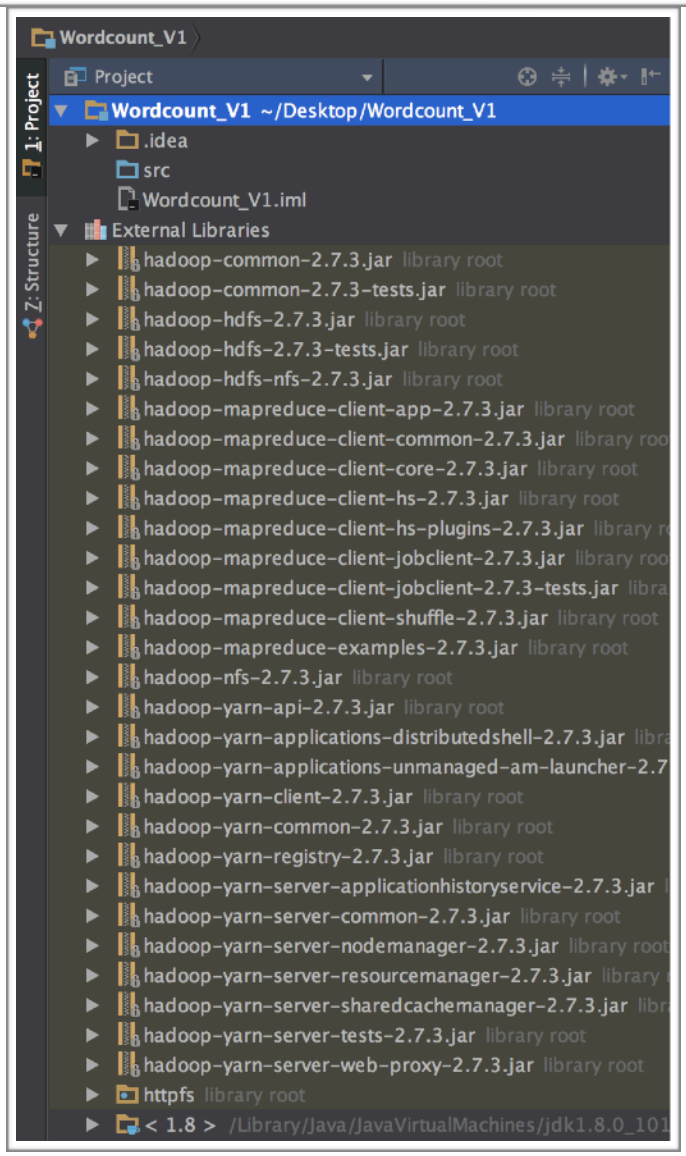
먼저 hadoop-2.7.3.tar 를 다운 받고, IntelliJ 에서 새로운 Java 프로젝트를 생성 해주었습니다. 그리고 Project Structure에서 Module/dependencies 설정을 통해 hadoop-2.7.3/share/hadoop/ 안의 모든 jar 파일을 import 해줍니다.



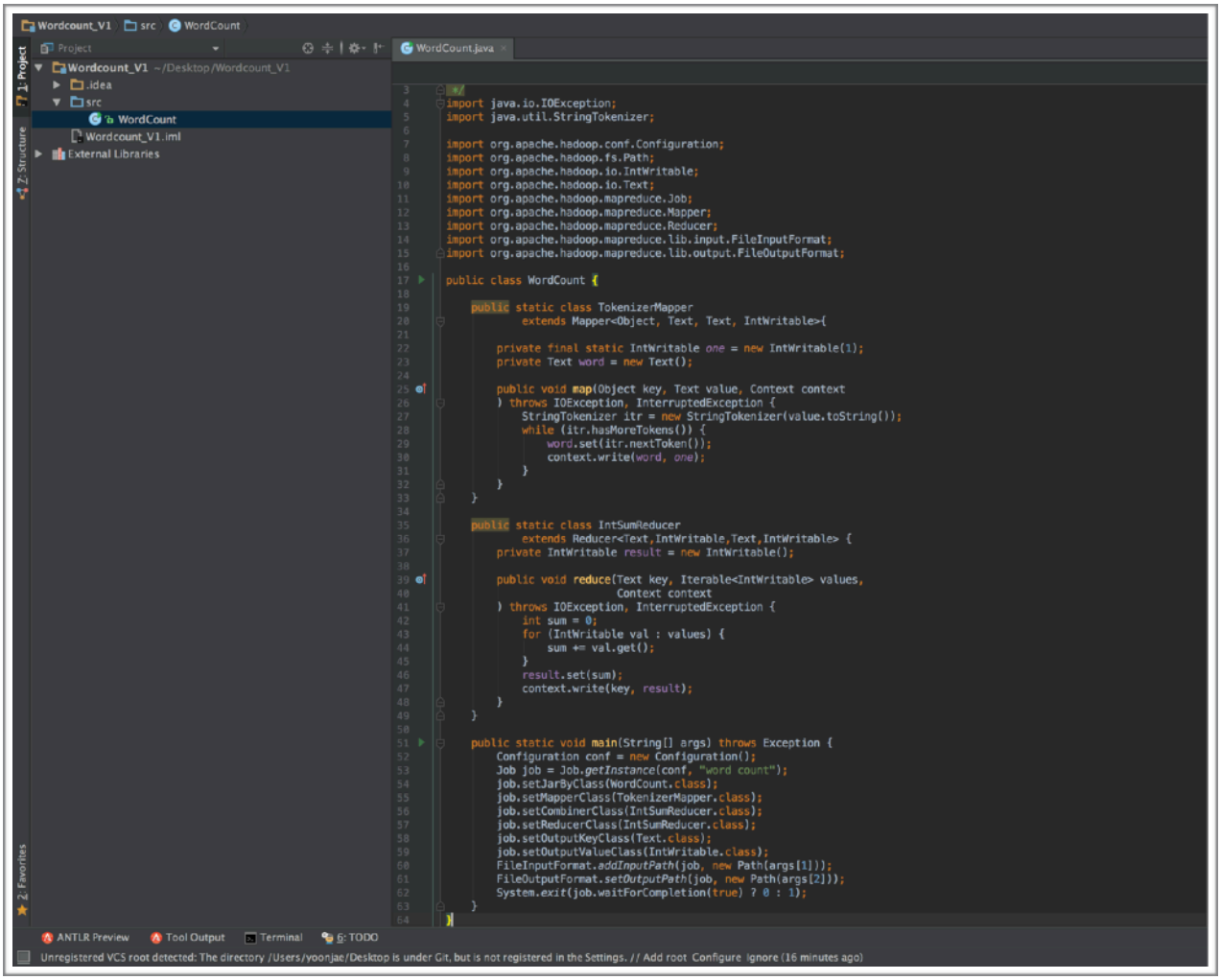


Modules Tab의 Dependencies





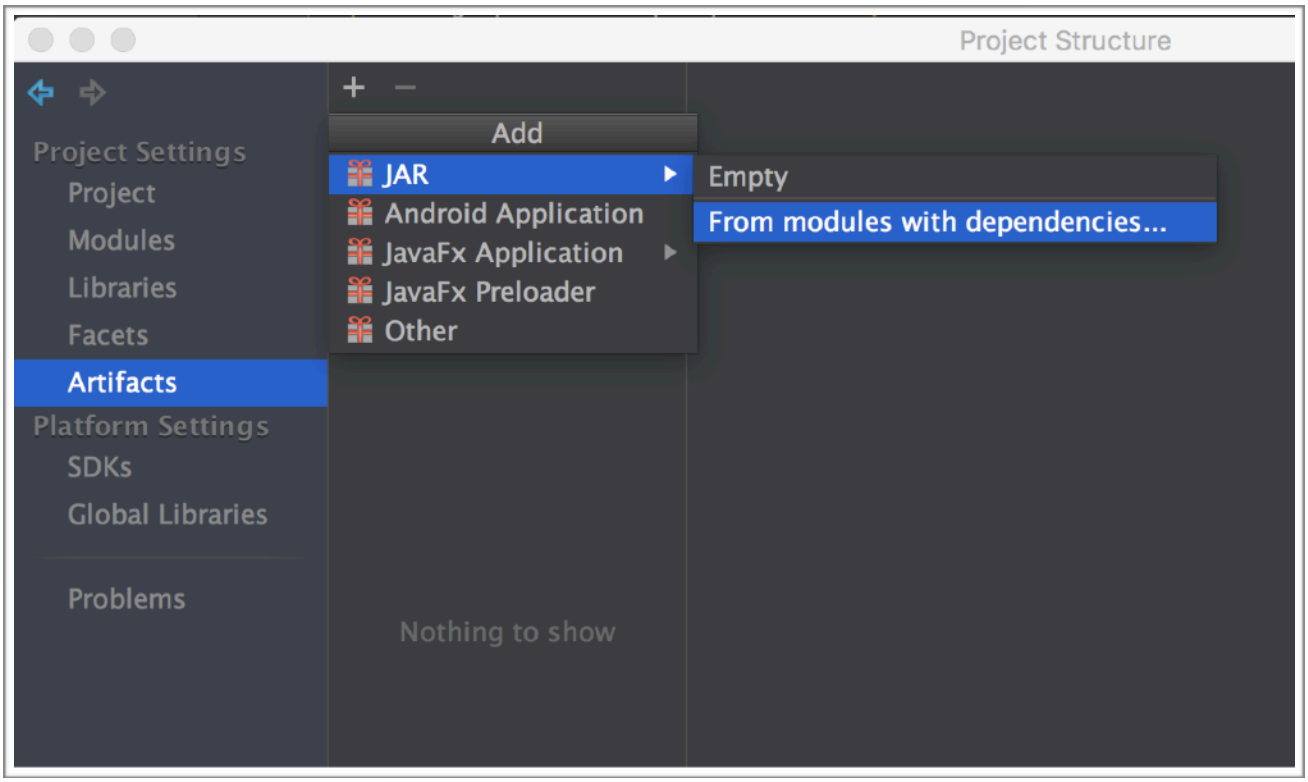
External Libraries에 hadoop에 관한 jar 파일이 import 된 것을 확인 할 수 있습니다.



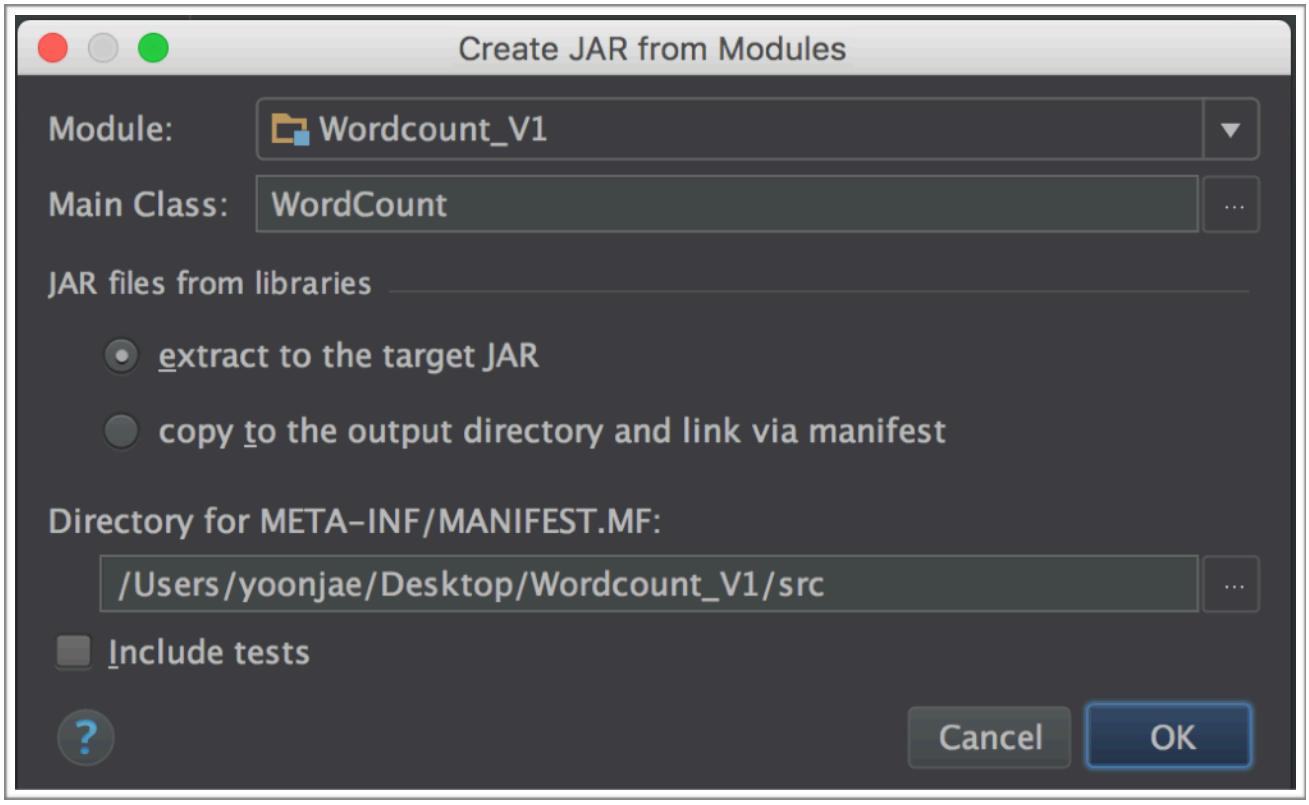
Hadoop Reference 의 word count 예제를 그대로 가져왔습니다.

```
public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = Job.getInstance(conf, "word count");
    job.setJarByClass(WordCount.class);
    job.setMapperClass(TokenizerMapper.class);
    job.setCombinerClass(IntSumReducer.class);
    job.setReducerClass(IntSumReducer.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    FileInputFormat.addInputPath(job, new Path(args[1]));
    FileOutputFormat.setOutputPath(job, new Path(args[2]));
    System.exit(job.waitForCompletion(true) ? 0 : 1);
}
```

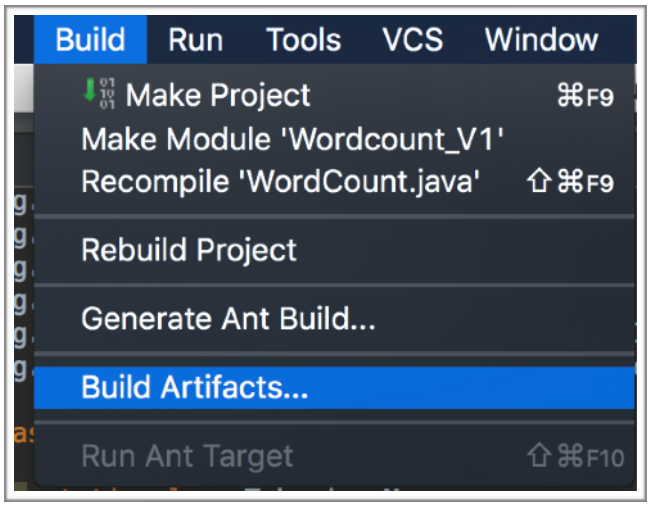
여기서 input/output의 path를 args[1], args[2]로 바꾸어줍니다.

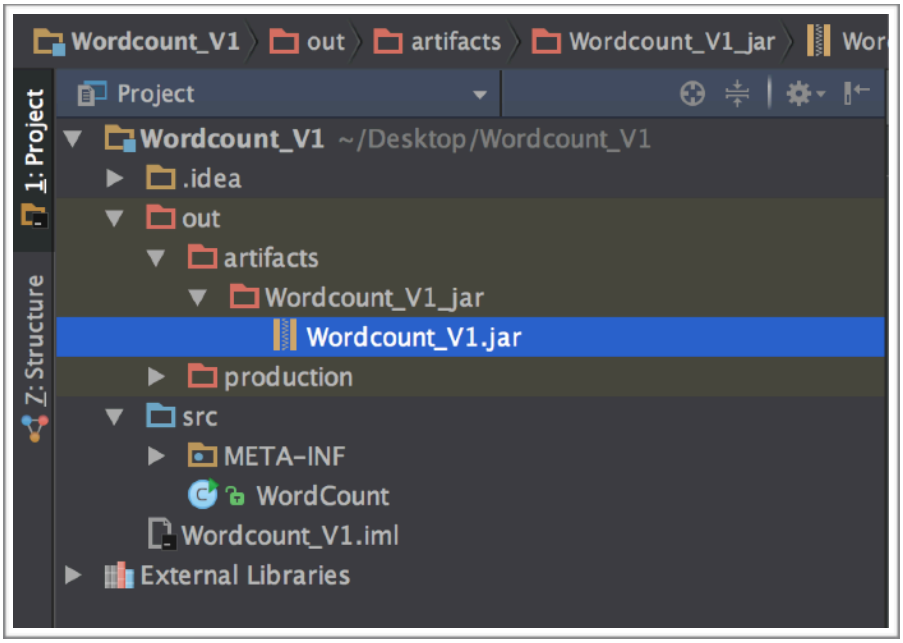


다음으로는 jar 형태로 export하기 위해 Project Structure / Artifacts / JAR 를 선택합니다.



Main Class를 선택해주고 설정을 완료합니다.

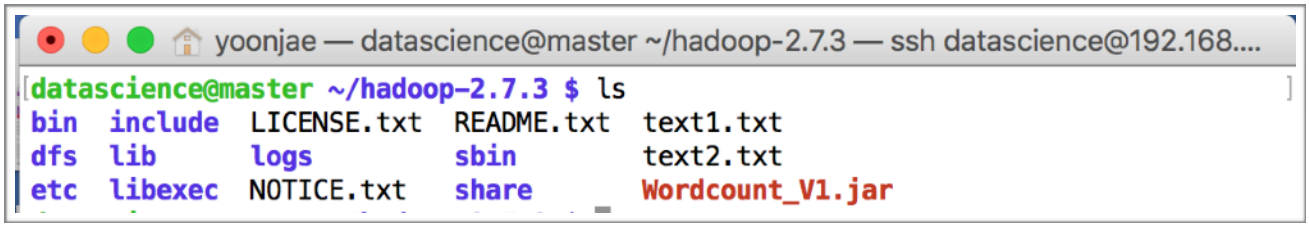




그리고 Build Artifacts를 해주면 위와 같이 Wordcount_V1.jar가 생긴 것을 확인 할 수 있습니다.

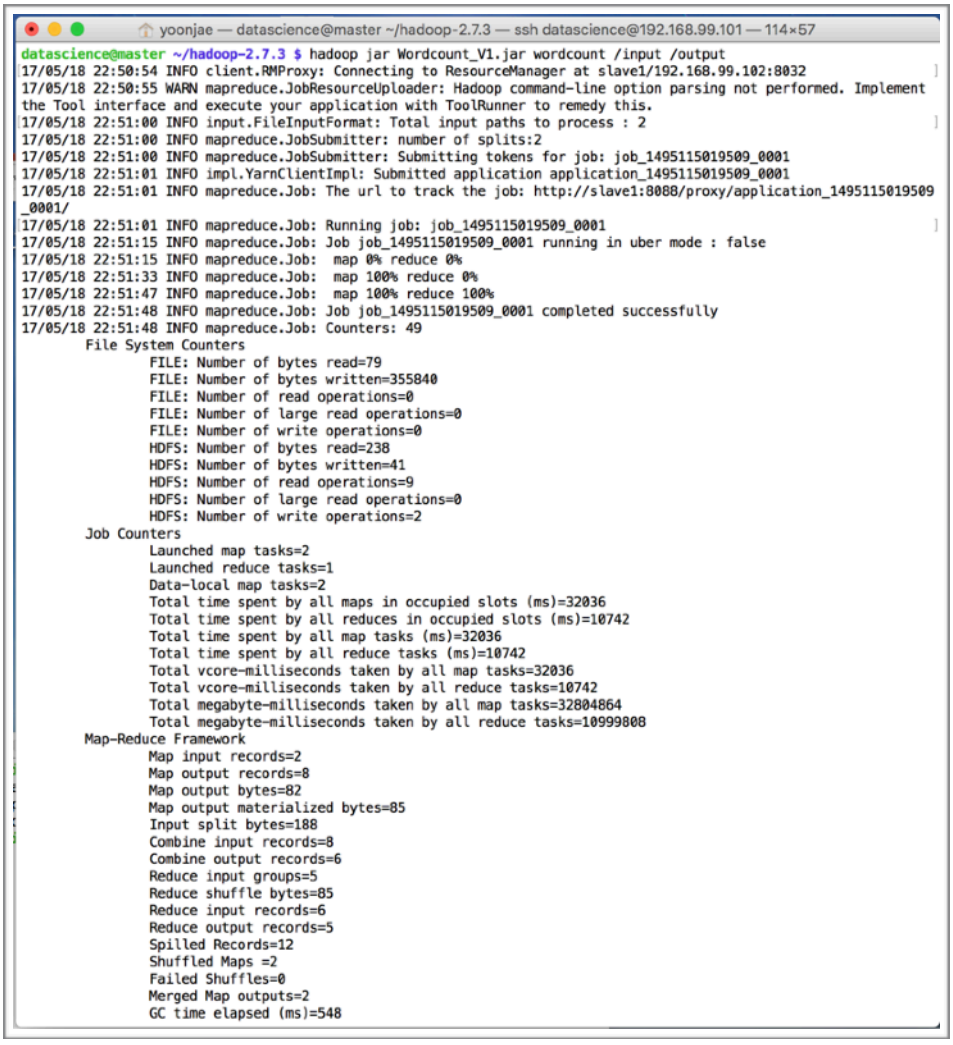
이제 Build된 jar파일을 rsync 명령어를 통해 각 Node에 전달 해줍니다.

rsync -avz /Users/yoonjae/Desktop/Wordcount_V1/out/artifacts/Wordcount_V1_jar/ Wordcount_V1.jar datascience@192.168.99.100:/home/datascience/hadoop-2.7.3/



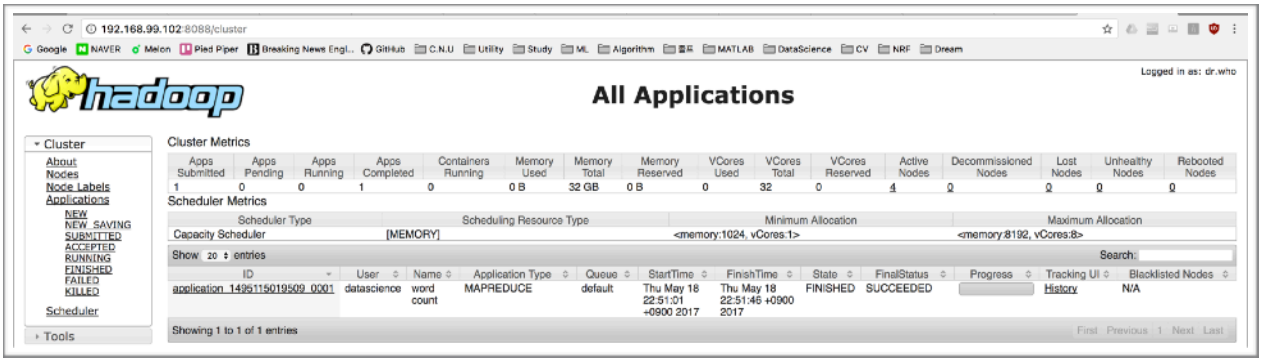
```
datascience@master ~/hadoop-2.7.3 $ jps
2529 DataNode
2674 SecondaryNameNode
2388 NameNode
3117 Jps
2863 NodeManager
datascience@master ~/hadoop-2.7.3 $ hdfs dfs -mkdir /input
datascience@master ~/hadoop-2.7.3 $ hdfs dfs -put text* /input
datascience@master ~/hadoop-2.7.3 $ hdfs dfs -ls /input
Found 2 items
-rw-r--r--   4 datascience supergroup         22 2017-05-18 22:47 /input/text1.txt
-rw-r--r--   4 datascience supergroup         28 2017-05-18 22:47 /input/text2.txt
```

start-dfs.sh 와 start-yarn.sh를 통해 각각 프로세스를 실행 해준 후에, hdfs에 input 폴더를 생성 해주고, word count 할 text 파일을 put 해줍니다.



hadoop jar Wordcount_V1.jar word count /input /output 명령어를 통해 wordcount를 실행합니다.

```
datascience@master ~/hadoop-2.7.3 $ hdfs dfs -cat /output/part-r-00000
Bye     1
Goodbye 1
Hadoop, 1
Hello   2
World!  1
World,  1
hadoop. 1
to      1
```



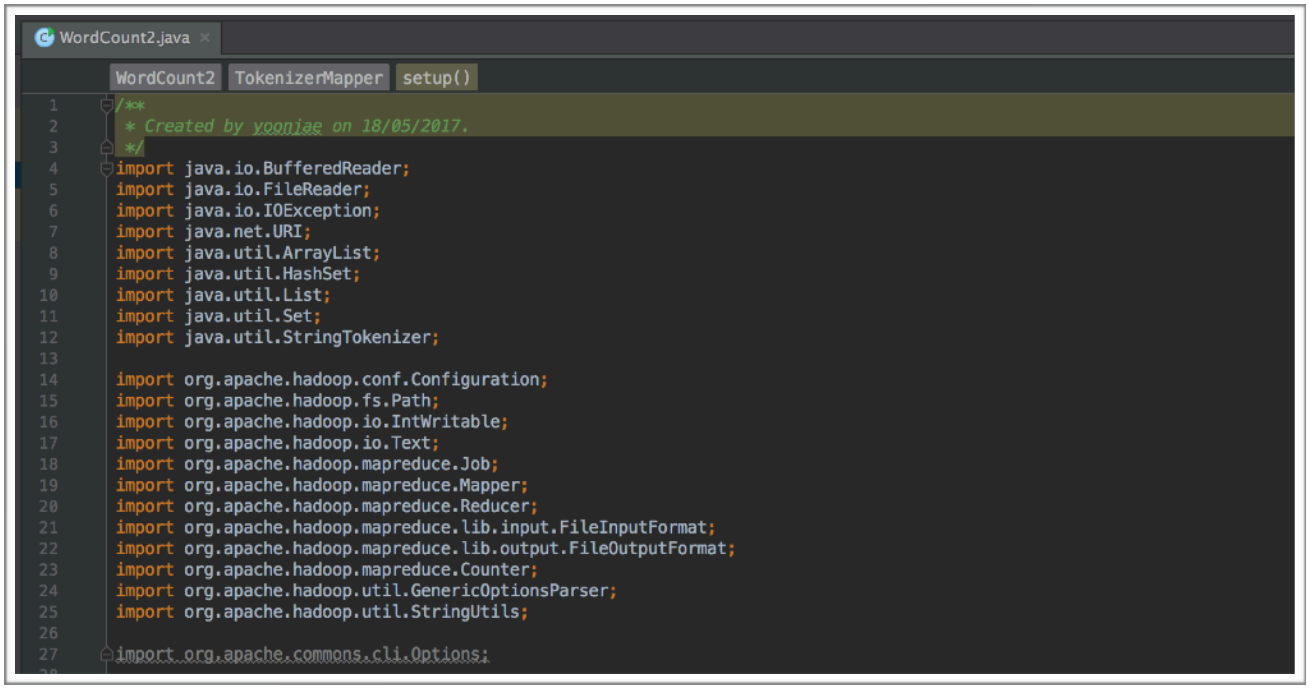

```
yoonjae — datascience@master ~/hadoop-2.7.3 — ssh
datascience@master ~/hadoop-2.7.3 $ hdfs dfsadmin -report
Configured Capacity: 109774323712 (102.24 GB)
Present Capacity: 64269320192 (59.86 GB)
DFS Remaining: 64268369920 (59.85 GB)
DFS Used: 950272 (928 KB)
DFS Used%: 0.00%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0

-------------------------------------------------
Live datanodes (4):

Name: 192.168.99.101:50010 (master)
Hostname: master
Decommission Status : Normal
Configured Capacity: 27443580928 (25.56 GB)
DFS Used: 237568 (232 KB)
Non DFS Used: 11377348608 (10.60 GB)
DFS Remaining: 16065994752 (14.96 GB)
DFS Used%: 0.00%
DFS Remaining%: 58.54%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Thu May 18 22:58:15 KST 2017


Name: 192.168.99.102:50010 (slave1)
Hostname: slave1
Decommission Status : Normal
Configured Capacity: 27443580928 (25.56 GB)
DFS Used: 237568 (232 KB)
Non DFS Used: 11376230400 (10.59 GB)
DFS Remaining: 16067112960 (14.96 GB)
DFS Used%: 0.00%
DFS Remaining%: 58.55%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Thu May 18 22:58:17 KST 2017
```

```
Name: 192.168.99.103:50010 (slave3)
Hostname: slave3
Decommission Status : Normal
Configured Capacity: 27443580928 (25.56 GB)
DFS Used: 237568 (232 KB)
Non DFS Used: 11375800320 (10.59 GB)
DFS Remaining: 16067543040 (14.96 GB)
DFS Used%: 0.00%
DFS Remaining%: 58.55%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Thu May 18 22:58:17 KST 2017


Name: 192.168.99.100:50010 (slave2)
Hostname: slave2
Decommission Status : Normal
Configured Capacity: 27443580928 (25.56 GB)
DFS Used: 237568 (232 KB)
Non DFS Used: 11375624192 (10.59 GB)
DFS Remaining: 16067719168 (14.96 GB)
DFS Used%: 0.00%
DFS Remaining%: 58.55%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Thu May 18 22:58:17 KST 2017
```

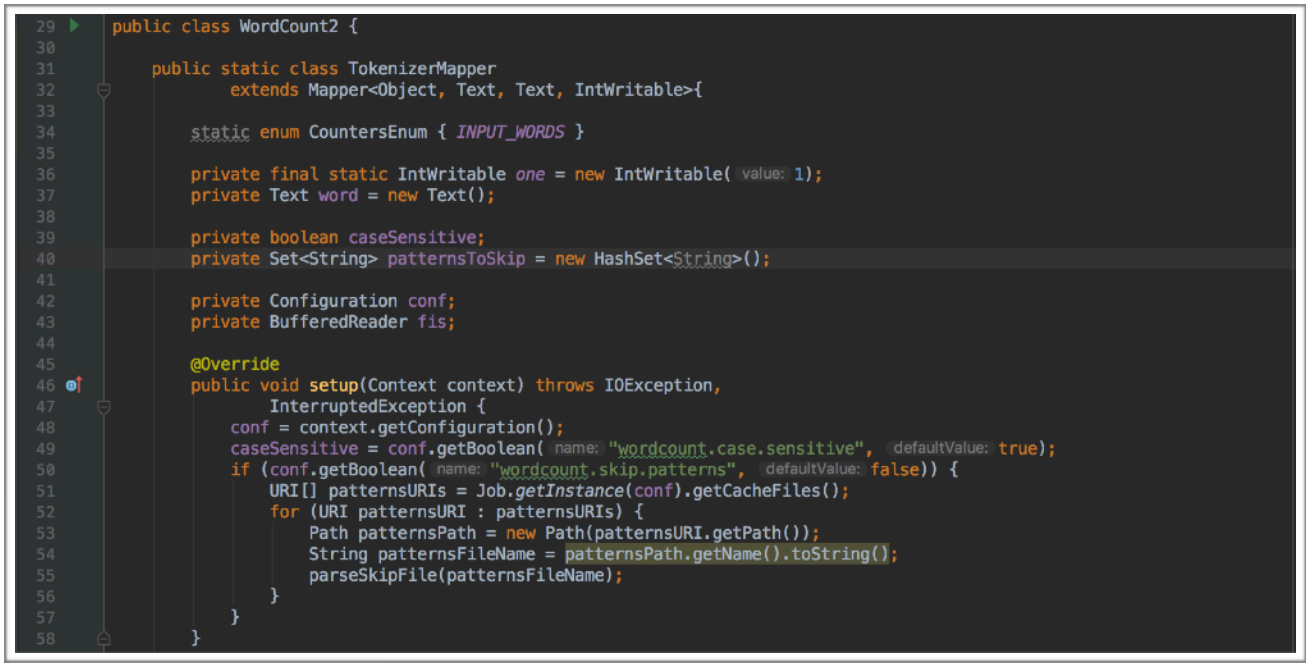
Word count Version 1 을 진행한 결과 입니다.

2. Word Count Version 2

Word count Version 2 에서는 Reference 에 있는 Code를 그대로 컴파일 하면 오류가 있어서 몇가지 수정을 해야 했습니다.



첫번째 오류는 import가 되어있지 않은 module이었습니다. 따라서 컴파일에 필요한 commons-cli-1.4를 다운받아서 Module을 import 해주었습니다.



TokenizeMapper Class의 코드 입니다. setup이라는 함수가 호출되며 처음에 여러가지 설정을 하는데 그 중, case sensitive와 patterns에 대한 모습입니다. default값이 true, false로 되어있는 것을 확인 할 수 있습니다. 따라서 case sensitive 설정을 따로 하지않으면 true로 되어 대/소문자를 구별하게 됩니다.

```
60          private void parseSkipFile(String fileName) {
61              try {
62                  fis = new BufferedReader(new FileReader(fileName));
63                  String pattern = null;
64                  while ((pattern = fis.readLine()) != null) {
65                      patternsToSkip.add(pattern);
66                  }
67              } catch (IOException ioe) {
68                  System.err.println("Caught exception while parsing the cached file '"
69                          + StringUtils.stringifyException(ioe));
70              }
71          }
72
73          @Override
74          public void map(Object key, Text value, Context context
75          ) throws IOException, InterruptedException {
76              String line = (caseSensitive) ?
77                      value.toString() : value.toString().toLowerCase();
78              for (String pattern : patternsToSkip) {
79                  line = line.replaceAll(pattern,  replacement: "");
80              }
81              StringTokenizer itr = new StringTokenizer(line);
82              while (itr.hasMoreTokens()) {
83                  word.set(itr.nextToken());
84                  context.write(word, one);
85                  Counter counter = context.getCounter(CountersEnum.class.getName(),
86                          CountersEnum.INPUT_WORDS.toString());
87                  counter.increment( l: 1);
88              }
89          }
90      }
```

TokenizeMapper Class의 나머지 코드 입니다. pattern file에 대한 처리와 map 함수를 보면 caseSensitive의 값을 통해 소문자로 변환 할 것인가의 유무를 정하게 됩니다.

```
108     public static void main(String[] args) throws Exception {
109         Configuration conf = new Configuration();
110         GenericOptionsParser optionParser = new GenericOptionsParser(conf, args);
111         String[] remainingArgs = optionParser.getRemainingArgs();
112         if ((remainingArgs.length != 2) && (remainingArgs.length != 5)) {
113             System.err.println("Usage: wordcount <in> <out> [-skip skipPatternFile]");
114             System.exit( status: 2);
115         }
116         Job job = Job.getInstance(conf,  jobName: "word count");
117         job.setJarByClass(WordCount2.class);
118         job.setMapperClass(TokenizerMapper.class);
119         job.setCombinerClass(IntSumReducer.class);
120         job.setReducerClass(IntSumReducer.class);
121         job.setOutputKeyClass(Text.class);
122         job.setOutputValueClass(IntWritable.class);
123
124         List<String> otherArgs = new ArrayList<String>();
125         for (int i=0; i < remainingArgs.length; ++i) {
126             if ("-skip".equals(remainingArgs[i])) {
127                 job.addCacheFile(new Path(remainingArgs[++i]).toUri());
128                 job.getConfiguration().setBoolean( name: "wordcount.skip.patterns",  value: true);
129             } else {
130                 otherArgs.add(remainingArgs[i]);
131             }
132         }
133         FileInputFormat.addInputPath(job, new Path(otherArgs.get(1)));
134         FileOutputFormat.setOutputPath(job, new Path(otherArgs.get(2)));
135
136         System.exit(job.waitForCompletion( verbose: true) ? 0 : 1);
137     }
138 }
```

Main 함수의 모습입니다. 두번째 오류는 hadoop jar 명령어를 실행 할 때, 여러가지 인자를 전달 하는데 그 인자의 개수가 2개 또는 4개가 아니면 오류라고 되어있었는데, 4개를 5개로 고치고 여러가지 옵션을 넣을수 있도록 설정하였습니다.

코드를 수정 한 후, 1번과 같이 jar 파일로 변환 한 후, 모든 Node에 전달합니다.

```
yoonjae — datascience@master ~/hadoop-2.7.3 — ssh datascience@192.168....
[datascience@master ~/hadoop-2.7.3 $ cat pattern.txt
\.
\,
\!
to
```

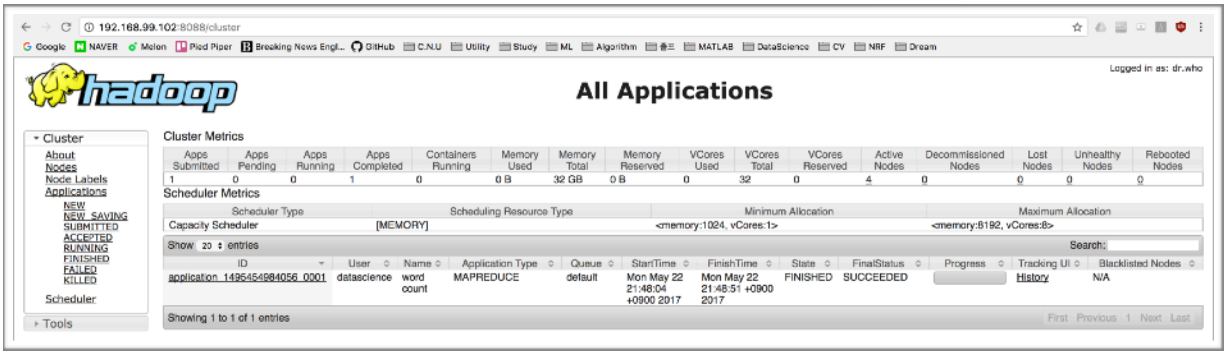
```
[datascience@master ~/hadoop-2.7.3 $ hdfs dfs -put pattern.txt /patterns
[datascience@master ~/hadoop-2.7.3 $ hdfs dfs -ls /patterns
Found 1 items
-rw-r--r--   4 datascience supergroup         12 2017-05-22 21:45 /patterns/patt
ern.txt
```

다음으로는 Word count시 제외할 pattern을 기록 해두는 text파일을 만들고, hdfs에 올려 줍니다.

```
yoonjae — datascience@master ~/hadoop-2.7.3 — ssh datascience@192.168.99.100 — 122×59
datascience@master ~/hadoop-2.7.3 $ hadoop jar Wordcount_V2.jar -Dwordcount.case.sensitive=true Wordcount2 /input /output
-skip /patterns/pattern.txt
17/05/22 21:47:58 INFO client.RMProxy: Connecting to ResourceManager at slave1/192.168.99.102:8032
17/05/22 21:48:03 INFO input.FileInputFormat: Total input paths to process : 2
17/05/22 21:48:03 INFO mapreduce.JobSubmitter: number of splits:2
17/05/22 21:48:04 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1495454984056_0001
17/05/22 21:48:04 INFO impl.YarnClientImpl: Submitted application application_1495454984056_0001
17/05/22 21:48:05 INFO mapreduce.Job: The url to track the job: http://slave1:8088/proxy/application_1495454984056_0001/
17/05/22 21:48:05 INFO mapreduce.Job: Running job: job_1495454984056_0001
17/05/22 21:48:19 INFO mapreduce.Job: Job job_1495454984056_0001 running in uber mode : false
17/05/22 21:48:19 INFO mapreduce.Job:  map 0% reduce 0%
17/05/22 21:48:40 INFO mapreduce.Job:  map 100% reduce 0%
17/05/22 21:48:52 INFO mapreduce.Job:  map 100% reduce 100%
17/05/22 21:48:53 INFO mapreduce.Job: Job job_1495454984056_0001 completed successfully
17/05/22 21:48:53 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=92
                FILE: Number of bytes written=359754
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=245
                HDFS: Number of bytes written=50
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=37899
                Total time spent by all reduces in occupied slots (ms)=7793
                Total time spent by all map tasks (ms)=37899
                Total time spent by all reduce tasks (ms)=7793
                Total vcore-milliseconds taken by all map tasks=37899
                Total vcore-milliseconds taken by all reduce tasks=7793
                Total megabyte-milliseconds taken by all map tasks=38808576
                Total megabyte-milliseconds taken by all reduce tasks=7980032
        Map-Reduce Framework
                Map input records=2
                Map output records=8
                Map output bytes=82
                Map output materialized bytes=98
                Input split bytes=188
                Combine input records=8
                Combine output records=7
                Reduce input groups=6
                Reduce shuffle bytes=98
                Reduce input records=7
                Reduce output records=6
                Spilled Records=14
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=516
                CPU time spent (ms)=2960
                Physical memory (bytes) snapshot=481865728
                Virtual memory (bytes) snapshot=5658157056
                Total committed heap usage (bytes)=262500352
```

hadoop jar Wordcount_V2.jar -Dwordcount.case.sensitive=true WordCount2 /input /output -skip /patterns/pattern.txt 명령어를 통해 word count를 실행합니다.

```
datascience@master ~/hadoop-2.7.3 $ hdfs dfs -cat /output/part-r-00000
Bye     1
Goodbye 1
Hadoop  1
Hello   2
World   2
hadoop  1
```

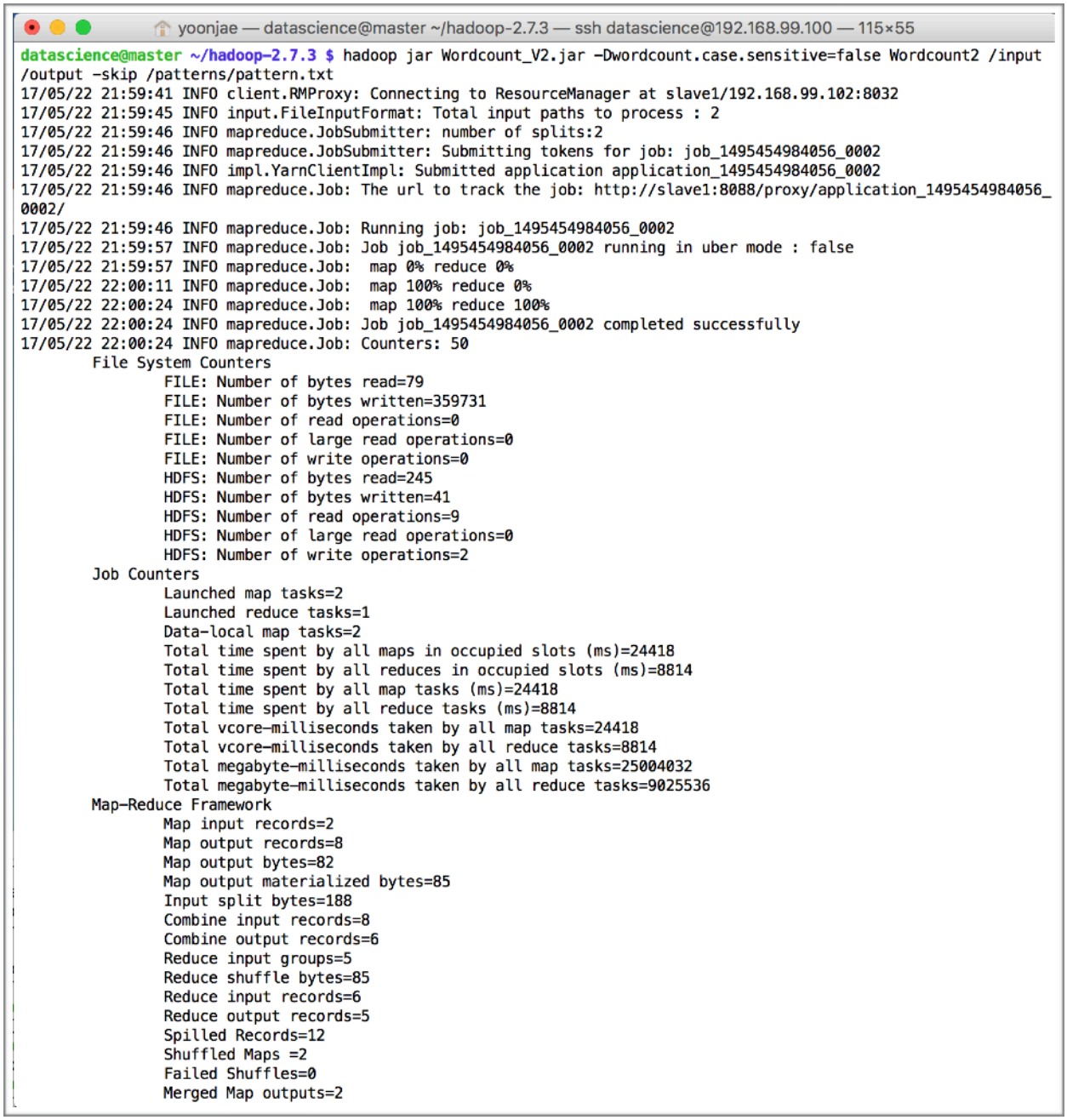



```
datascience@master ~/hadoop-2.7.3 $ hdfs dfsadmin -report
Configured Capacity: 109774323712 (102.24 GB)
Present Capacity: 64005353472 (59.61 GB)
DFS Remaining: 64004354048 (59.61 GB)
DFS Used: 999424 (976 KB)
DFS Used%: 0.00%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0

-------------------------------------------------
Live datanodes (4):

Name: 192.168.99.101:50010 (slave2)
Hostname: slave2
Decommission Status : Normal
Configured Capacity: 27443580928 (25.56 GB)
DFS Used: 249856 (244 KB)
Non DFS Used: 11441618944 (10.66 GB)
DFS Remaining: 16001712128 (14.90 GB)
DFS Used%: 0.00%
DFS Remaining%: 58.31%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Mon May 22 21:56:11 KST 2017


Name: 192.168.99.100:50010 (master)
Hostname: master
Decommission Status : Normal
Configured Capacity: 27443580928 (25.56 GB)
DFS Used: 249856 (244 KB)
Non DFS Used: 11443470336 (10.66 GB)
DFS Remaining: 15999860736 (14.90 GB)
DFS Used%: 0.00%
DFS Remaining%: 58.30%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Mon May 22 21:56:13 KST 2017
```

```
Name: 192.168.99.103:50010 (slave3)
Hostname: slave3
Decommission Status : Normal
Configured Capacity: 27443580928 (25.56 GB)
DFS Used: 249856 (244 KB)
Non DFS Used: 11441623040 (10.66 GB)
DFS Remaining: 16001708032 (14.90 GB)
DFS Used%: 0.00%
DFS Remaining%: 58.31%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Mon May 22 21:56:11 KST 2017


Name: 192.168.99.102:50010 (slave1)
Hostname: slave1
Decommission Status : Normal
Configured Capacity: 27443580928 (25.56 GB)
DFS Used: 249856 (244 KB)
Non DFS Used: 11442257920 (10.66 GB)
DFS Remaining: 16001073152 (14.90 GB)
DFS Used%: 0.00%
DFS Remaining%: 58.31%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Mon May 22 21:56:11 KST 2017
```

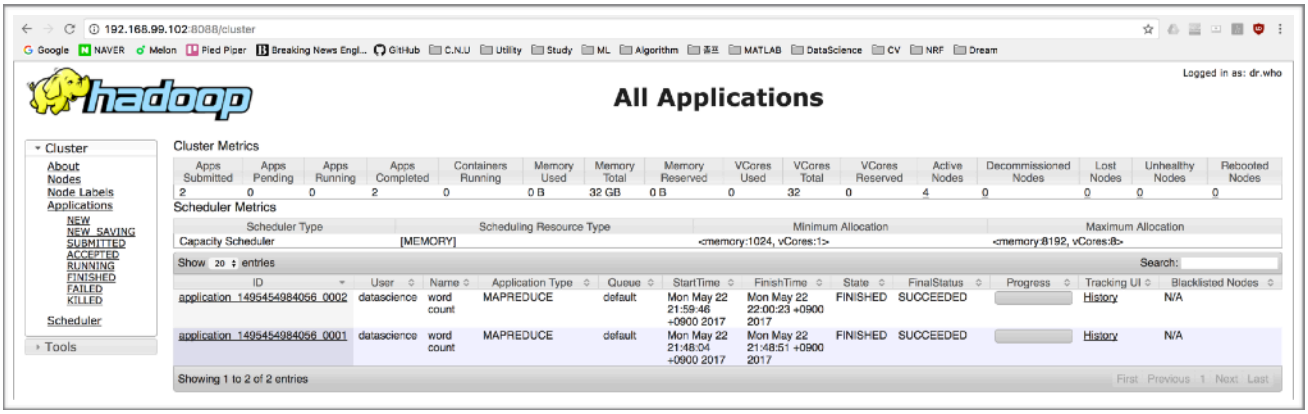
Word count Version 2 의 case.sensitive=true를 진행한 결과 입니다.

이제 hdfs의 /output file을 비워두고, 이번에는 case.sensitive=false 명령어를 주어 wordcount를 진행합니다.



hadoop jar Wordcount_V2.jar -Dwordcount.case.sensitive=false Wordcount2 /input /output -skip /patterns/pattern.txt

```
datascience@master ~/hadoop-2.7.3 $ hdfs dfs -cat /output/part-r-00000
bye     1
goodbye 1
hadoop  2
hello   2
world   2
```




```
datascience@master ~/hadoop-2.7.3 $ hdfs dfsadmin -report
Configured Capacity: 109774323712 (102.24 GB)
Present Capacity: 64004993236 (59.61 GB)
DFS Remaining: 64003350528 (59.61 GB)
DFS Used: 1642708 (1.57 MB)
DFS Used%: 0.00%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0

-------------------------------------------------
Live datanodes (4):

Name: 192.168.99.101:50010 (slave2)
Hostname: slave2
Decommission Status : Normal
Configured Capacity: 27443580928 (25.56 GB)
DFS Used: 410677 (401.05 KB)
Non DFS Used: 11441650635 (10.66 GB)
DFS Remaining: 16001519616 (14.90 GB)
DFS Used%: 0.00%
DFS Remaining%: 58.31%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Mon May 22 22:01:50 KST 2017


Name: 192.168.99.100:50010 (master)
Hostname: master
Decommission Status : Normal
Configured Capacity: 27443580928 (25.56 GB)
DFS Used: 410677 (401.05 KB)
Non DFS Used: 11443538891 (10.66 GB)
DFS Remaining: 15999631360 (14.90 GB)
DFS Used%: 0.00%
DFS Remaining%: 58.30%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Mon May 22 22:01:49 KST 2017
```

```
Name: 192.168.99.103:50010 (slave3)
Hostname: slave3
Decommission Status : Normal
Configured Capacity: 27443580928 (25.56 GB)
DFS Used: 410677 (401.05 KB)
Non DFS Used: 11441740747 (10.66 GB)
DFS Remaining: 16001429504 (14.90 GB)
DFS Used%: 0.00%
DFS Remaining%: 58.31%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Mon May 22 22:01:50 KST 2017


Name: 192.168.99.102:50010 (slave1)
Hostname: slave1
Decommission Status : Normal
Configured Capacity: 27443580928 (25.56 GB)
DFS Used: 410677 (401.05 KB)
Non DFS Used: 11442400203 (10.66 GB)
DFS Remaining: 16000770048 (14.90 GB)
DFS Used%: 0.00%
DFS Remaining%: 58.30%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Mon May 22 22:01:50 KST 2017
```

Word count Version 2 의 case.sensitive=false를 진행한 결과 입니다.