Ryan Ballenger
CSCIE63 Big Data Analytics
Assignment 4

Assignment 4 Solutions

#1
Java API Terminal Commands
Spark Java API - First 20 Results
Description of WordCount.java
WordCount.java File

#2
Scala API Terminal Commands
Spark Scala API - First 20 Results
Description of WordCount.scala
WordCount.scala

#3
Python API Terminal Commands
Spark Python API - First 20 Results
Description of WordCount.py
WordCount.py

#4
Bigram Terminal Commands
Bigram Results
Description of the bigram code
Bigram Code

Problem 1. Write a working WordCount program using Spark Java API that reads a file, e.g. Ulysis/4300.txt from an HDFS directory and writes the results of your calculations to an HDFS file. To improve your word count, remove any punctuation that might have attached itself to your words. Also transform all words into lower case so that the capitalization does not affect the word count. The original code used in lecture notes is provided in the attached mini-example-java.tar file. That archive also contains Maven's pom.xml file. Run your program and demonstrate that it works. Submit working code inside the customary MS Word Document. Describe steps in your program.

Java API Terminal Commands

// I clean, compile, and package the Java program with Maven as described in lecture.
// This converts the set of plain .java files to an executable program ready to run

```
[cloudera@localhost mini-examples-java-v2]$ mvn clean && mvn compile && mvn package
[INFO] Scanning for projects...
[INFO]
[INFO] ------------------------------------------------------------------------
[INFO] Building example 0.0.1
[INFO] ------------------------------------------------------------------------
[INFO]
[INFO] --- maven-clean-plugin:2.5:clean (default-clean) @ spark-example ---
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time: 0.930 s
[INFO] Finished at: 2016-02-26T12:32:56-08:00
[INFO] Final Memory: 5M/29M
[INFO] ------------------------------------------------------------------------
[INFO] Scanning for projects...
[INFO]
[INFO] ------------------------------------------------------------------------
[INFO] Building example 0.0.1
[INFO] ------------------------------------------------------------------------
[INFO]
[INFO] --- maven-resources-plugin:2.6:resources (default-resources) @ spark-example ---
[WARNING] Using platform encoding (UTF-8 actually) to copy filtered resources, i.e. build is
platform dependent!
[INFO] skip non existing resourceDirectory
/home/cloudera/mini-examples-java-v2/src/main/resources
[INFO]
[INFO] --- maven-compiler-plugin:3.1:compile (default-compile) @ spark-example ---
[INFO] Changes detected - recompiling the module!
[WARNING] File encoding has not been set, using platform encoding UTF-8, i.e. build is
platform dependent!
[INFO] Compiling 4 source files to /home/cloudera/mini-examples-java-v2/target/classes
[WARNING]
/home/cloudera/mini-examples-java-v2/src/main/java/edu/hu/examples/WordCount.java: Some
input files use unchecked or unsafe operations.
```

```
[WARNING]
/home/cloudera/mini-examples-java-v2/src/main/java/edu/hu/examples/WordCount.java:
Recompile with -Xlint:unchecked for details.
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time: 7.403 s
[INFO] Finished at: 2016-02-26T12:33:06-08:00
[INFO] Final Memory: 22M/54M
[INFO] ------------------------------------------------------------------------
[INFO] Scanning for projects...
[INFO]
[INFO] ------------------------------------------------------------------------
[INFO] Building example 0.0.1
[INFO] ------------------------------------------------------------------------
[INFO]
[INFO] --- maven-resources-plugin:2.6:resources (default-resources) @ spark-example ---
[WARNING] Using platform encoding (UTF-8 actually) to copy filtered resources, i.e. build is
platform dependent!
[INFO] skip non existing resourceDirectory
/home/cloudera/mini-examples-java-v2/src/main/resources
[INFO]
[INFO] --- maven-compiler-plugin:3.1:compile (default-compile) @ spark-example ---
[INFO] Nothing to compile - all classes are up to date
[INFO]
[INFO] --- maven-resources-plugin:2.6:testResources (default-testResources) @ spark-example
---
[WARNING] Using platform encoding (UTF-8 actually) to copy filtered resources, i.e. build is
platform dependent!
[INFO] skip non existing resourceDirectory
/home/cloudera/mini-examples-java-v2/src/test/resources
[INFO]
[INFO] --- maven-compiler-plugin:3.1:testCompile (default-testCompile) @ spark-example ---
[INFO] No sources to compile
[INFO]
[INFO] --- maven-surefire-plugin:2.12.4:test (default-test) @ spark-example ---
[INFO] No tests to run.
[INFO]
[INFO] --- maven-jar-plugin:2.4:jar (default-jar) @ spark-example ---
[INFO] Building jar: /home/cloudera/mini-examples-java-v2/target/spark-example-0.0.1.jar
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
```

[INFO] Total time: 4.837 s
[INFO] Finished at: 2016-02-26T12:33:15-08:00
[INFO] Final Memory: 10M/29M
[INFO] --

// The java program is sent submitted and run by Spark. The WordCount class is called and text
// file 4300.txt from Ulysses is processed into the output folder wordcounts

[cloudera@localhost mini-examples-java-v2]$ spark-submit --class
edu.hu.examples.WordCount ./target/spark-example-0.0.1.jar ulysses/4300.txt wordcounts
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in
[jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in
[jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/02/26 12:35:44 INFO spark.SparkContext: Running Spark version 1.5.0-cdh5.5.2
16/02/26 12:35:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
16/02/26 12:35:53 WARN util.Utils: Your hostname, localhost.localdomain resolves to a
loopback address: 127.0.0.1; using 172.16.40.129 instead (on interface eth0)
16/02/26 12:35:53 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
16/02/26 12:35:53 INFO spark.SecurityManager: Changing view acls to: cloudera
16/02/26 12:35:53 INFO spark.SecurityManager: Changing modify acls to: cloudera
16/02/26 12:35:53 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui
acls disabled; users with view permissions: Set(cloudera); users with modify permissions:
Set(cloudera)
16/02/26 12:36:00 INFO slf4j.Slf4jLogger: Slf4jLogger started
16/02/26 12:36:01 INFO Remoting: Starting remoting
16/02/26 12:36:04 INFO Remoting: Remoting started; listening on addresses
:[akka.tcp://sparkDriver@172.16.40.129:46347]
16/02/26 12:36:04 INFO Remoting: Remoting now listens on addresses:
[akka.tcp://sparkDriver@172.16.40.129:46347]
16/02/26 12:36:04 INFO util.Utils: Successfully started service 'sparkDriver' on port 46347.
16/02/26 12:36:05 INFO spark.SparkEnv: Registering MapOutputTracker
16/02/26 12:36:06 INFO spark.SparkEnv: Registering BlockManagerMaster
16/02/26 12:36:06 INFO storage.DiskBlockManager: Created local directory at
/tmp/blockmgr-97f7514d-dd1d-4ba4-ae9d-d1ee6f8ef905
16/02/26 12:36:07 INFO storage.MemoryStore: MemoryStore started with capacity 534.5 MB

16/02/26 12:36:07 INFO spark.HttpFileServer: HTTP File server directory is /tmp/spark-55b53252-500f-4164-92b5-6aacacd482f7/httpd-4e44f174-1040-44cb-bc79-e07dc3a19038
16/02/26 12:36:07 INFO spark.HttpServer: Starting HTTP Server
16/02/26 12:36:08 INFO server.Server: jetty-8.y.z-SNAPSHOT
16/02/26 12:36:08 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0:34524
16/02/26 12:36:08 INFO util.Utils: Successfully started service 'HTTP file server' on port 34524.
16/02/26 12:36:08 INFO spark.SparkEnv: Registering OutputCommitCoordinator
16/02/26 12:36:09 INFO server.Server: jetty-8.y.z-SNAPSHOT
16/02/26 12:36:09 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
16/02/26 12:36:09 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
16/02/26 12:36:09 INFO ui.SparkUI: Started SparkUI at http://172.16.40.129:4040
16/02/26 12:36:09 INFO spark.SparkContext: Added JAR file:/home/cloudera/mini-examples-java-v2/./target/spark-example-0.0.1.jar at http://172.16.40.129:34524/jars/spark-example-0.0.1.jar with timestamp 1456518969507
16/02/26 12:36:10 WARN metrics.MetricsSystem: Using default name DAGScheduler for source because spark.app.id is not set.
16/02/26 12:36:10 INFO executor.Executor: Starting executor ID driver on host localhost
16/02/26 12:36:10 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 51368.
16/02/26 12:36:10 INFO netty.NettyBlockTransferService: Server created on 51368
16/02/26 12:36:10 INFO storage.BlockManagerMaster: Trying to register BlockManager
16/02/26 12:36:10 INFO storage.BlockManagerMasterEndpoint: Registering block manager localhost:51368 with 534.5 MB RAM, BlockManagerId(driver, localhost, 51368)
16/02/26 12:36:10 INFO storage.BlockManagerMaster: Registered BlockManager
16/02/26 12:36:16 INFO storage.MemoryStore: ensureFreeSpace(136720) called with curMem=0, maxMem=560497950
16/02/26 12:36:16 INFO storage.MemoryStore: Block broadcast_0 stored as values in memory (estimated size 133.5 KB, free 534.4 MB)
16/02/26 12:36:16 INFO storage.MemoryStore: ensureFreeSpace(15386) called with curMem=136720, maxMem=560497950
16/02/26 12:36:16 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 15.0 KB, free 534.4 MB)
16/02/26 12:36:16 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:51368 (size: 15.0 KB, free: 534.5 MB)
16/02/26 12:36:16 INFO spark.SparkContext: Created broadcast 0 from textFile at WordCount.java:31
16/02/26 12:36:20 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
16/02/26 12:36:21 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/26 12:36:22 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id

16/02/26 12:36:22 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/02/26 12:36:22 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
16/02/26 12:36:22 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
16/02/26 12:36:22 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/02/26 12:36:22 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/26 12:36:23 INFO spark.SparkContext: Starting job: saveAsTextFile at WordCount.java:59
16/02/26 12:36:23 INFO scheduler.DAGScheduler: Registering RDD 3 (mapToPair at WordCount.java:49)
16/02/26 12:36:23 INFO scheduler.DAGScheduler: Got job 0 (saveAsTextFile at WordCount.java:59) with 1 output partitions
16/02/26 12:36:23 INFO scheduler.DAGScheduler: Final stage: ResultStage 1(saveAsTextFile at WordCount.java:59)
16/02/26 12:36:23 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 0)
16/02/26 12:36:23 INFO scheduler.DAGScheduler: Missing parents: List(ShuffleMapStage 0)
16/02/26 12:36:23 INFO scheduler.DAGScheduler: Submitting ShuffleMapStage 0 (MapPartitionsRDD[3] at mapToPair at WordCount.java:49), which has no missing parents
16/02/26 12:36:23 INFO storage.MemoryStore: ensureFreeSpace(4744) called with curMem=152106, maxMem=560497950
16/02/26 12:36:23 INFO storage.MemoryStore: Block broadcast_1 stored as values in memory (estimated size 4.6 KB, free 534.4 MB)
16/02/26 12:36:23 INFO storage.MemoryStore: ensureFreeSpace(2669) called with curMem=156850, maxMem=560497950
16/02/26 12:36:23 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 2.6 KB, free 534.4 MB)
16/02/26 12:36:23 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:51368 (size: 2.6 KB, free: 534.5 MB)
16/02/26 12:36:23 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:861
16/02/26 12:36:23 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 0 (MapPartitionsRDD[3] at mapToPair at WordCount.java:49)
16/02/26 12:36:23 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
16/02/26 12:36:24 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, partition 0,PROCESS_LOCAL, 2215 bytes)
16/02/26 12:36:24 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
16/02/26 12:36:24 INFO executor.Executor: Fetching http://172.16.40.129:34524/jars/spark-example-0.0.1.jar with timestamp 1456518969507

16/02/26 12:36:24 INFO util.Utils: Fetching http://172.16.40.129:34524/jars/spark-example-0.0.1.jar to /tmp/spark-55b53252-500f-4164-92b5-6aacacd482f7/userFiles-7e750677-082b-42ef-9594-762854e48ce4/fetchFileTemp8585060608422053903.tmp
16/02/26 12:36:24 INFO executor.Executor: Adding file:/tmp/spark-55b53252-500f-4164-92b5-6aacacd482f7/userFiles-7e750677-082b-42ef-9594-762854e48ce4/spark-example-0.0.1.jar to class loader
16/02/26 12:36:25 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/user/cloudera/ulysses/4300.txt:0+1573079
16/02/26 12:36:45 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 2253 bytes result sent to driver
16/02/26 12:36:45 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 21766 ms on localhost (1/1)
16/02/26 12:36:45 INFO scheduler.DAGScheduler: ShuffleMapStage 0 (mapToPair at WordCount.java:49) finished in 21.831 s
16/02/26 12:36:45 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
16/02/26 12:36:45 INFO scheduler.DAGScheduler: looking for newly runnable stages
16/02/26 12:36:45 INFO scheduler.DAGScheduler: running: Set()
16/02/26 12:36:45 INFO scheduler.DAGScheduler: waiting: Set(ResultStage 1)
16/02/26 12:36:45 INFO scheduler.DAGScheduler: failed: Set()
16/02/26 12:36:46 INFO scheduler.DAGScheduler: Missing parents for ResultStage 1: List()
16/02/26 12:36:46 INFO scheduler.DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[5] at saveAsTextFile at WordCount.java:59), which is now runnable
16/02/26 12:36:46 INFO storage.MemoryStore: ensureFreeSpace(135504) called with curMem=159519, maxMem=560497950
16/02/26 12:36:46 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory (estimated size 132.3 KB, free 534.3 MB)
16/02/26 12:36:46 INFO storage.MemoryStore: ensureFreeSpace(46401) called with curMem=295023, maxMem=560497950
16/02/26 12:36:46 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 45.3 KB, free 534.2 MB)
16/02/26 12:36:46 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:51368 (size: 45.3 KB, free: 534.5 MB)
16/02/26 12:36:46 INFO spark.SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:861
16/02/26 12:36:46 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (MapPartitionsRDD[5] at saveAsTextFile at WordCount.java:59)
16/02/26 12:36:46 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
16/02/26 12:36:46 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, partition 0,PROCESS_LOCAL, 1966 bytes)
16/02/26 12:36:46 INFO executor.Executor: Running task 0.0 in stage 1.0 (TID 1)

16/02/26 12:36:46 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
16/02/26 12:36:46 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 30 ms
16/02/26 12:36:46 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/26 12:36:49 INFO output.FileOutputCommitter: Saved output of task 'attempt_201602261236_0001_m_000000_1' to hdfs://localhost:8020/user/cloudera/wordcounts/_temporary/0/task_201602261236_0001_m_000000
16/02/26 12:36:49 INFO mapred.SparkHadoopMapRedUtil: attempt_201602261236_0001_m_000000_1: Committed
16/02/26 12:36:49 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 1). 2080 bytes result sent to driver
16/02/26 12:36:49 INFO scheduler.DAGScheduler: ResultStage 1 (saveAsTextFile at WordCount.java:59) finished in 2.906 s
16/02/26 12:36:49 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 2909 ms on localhost (1/1)
16/02/26 12:36:49 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
16/02/26 12:36:49 INFO scheduler.DAGScheduler: Job 0 finished: saveAsTextFile at WordCount.java:59, took 25.901224 s
16/02/26 12:36:49 INFO spark.SparkContext: Invoking stop() from shutdown hook
16/02/26 12:36:49 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/metrics/json,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/kill,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/api,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/static,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/threadDump/json,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/threadDump,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/json,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment/json,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment,null}

16/02/26 12:36:49 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage/rdd/json,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage/rdd,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage/json,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/pool/json,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/pool,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/stage/json,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/stage,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/json,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs/job/json,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs/job,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs/json,null}
16/02/26 12:36:49 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs,null}
16/02/26 12:36:49 INFO ui.SparkUI: Stopped Spark web UI at http://172.16.40.129:4040
16/02/26 12:36:49 INFO scheduler.DAGScheduler: Stopping DAGScheduler
16/02/26 12:36:49 INFO spark.MapOutputTrackerMasterEndpoint:
MapOutputTrackerMasterEndpoint stopped!
16/02/26 12:36:49 INFO storage.MemoryStore: MemoryStore cleared
16/02/26 12:36:49 INFO storage.BlockManager: BlockManager stopped
16/02/26 12:36:49 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
16/02/26 12:36:49 INFO
scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:
OutputCommitCoordinator stopped!
16/02/26 12:36:49 INFO spark.SparkContext: Successfully stopped SparkContext
16/02/26 12:36:49 INFO util.ShutdownHookManager: Shutdown hook called
16/02/26 12:36:49 INFO util.ShutdownHookManager: Deleting directory
/tmp/spark-55b53252-500f-4164-92b5-6aacacd482f7

16/02/26 12:36:49 INFO remote.RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
16/02/26 12:36:49 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.

// Lastly I copy the output to the VM_shared folder to be analyzed.

[cloudera@localhost mini-examples-java-v2]$ hadoop fs -copyToLocal wordcounts/* /mnt/hgfs/VM_shared/wordcounts/

Spark Java API - First 20 Results

(reunion,2)
(bone,11)
(luminary,2)
(monologue,1)
(rinbad,1)
(alavatar,1)
(blandly,2)
(gavin,2)
(serfs,1)
(bombast,1)
(hem,2)
(fred,2)
(stinks,4)
(flutiest,1)
(fuller,2)
(tough,3)
(jade,1)
(twinging,1)
(jove,5)
(crying,10)

Description of WordCount.java

I used the template mini-examples-java-v2. We are given code that initiates the Spark Context, initiates the Java Spark Context, loads the text file into a JavaRDD, and then begins the mapreduce process.

As a list of words are made from each string in the file, I inserted code to add on to this process. Instead of returning the plain list, a for-loop iterates through the list and changes the word's characters to all lowercase along with eliminating punctuation. It does this by using the replaceAll() function and replaces all non-alphanumerics with an empty string. After the list has

been adjusted to lowercase and punctuation-free, it is returned for the Iterable<String> method named call() inside the FlatMapFunction.

Afterwards, the reduce step proceeds as designed by the template. It transforms the data into a word and count and lastly, saves the processed data as a text file.

WordCount.java File

```
package edu.hu.examples;

import java.util.*;
import java.util.Arrays;
import java.util.List;
import java.lang.Iterable;

import scala.Tuple2;

import org.apache.commons.lang.StringUtils;

import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaRDD;
import org.apache.spark.api.java.JavaPairRDD;
import org.apache.spark.api.java.JavaSparkContext;
import org.apache.spark.api.java.function.FlatMapFunction;
import org.apache.spark.api.java.function.Function2;
import org.apache.spark.api.java.function.PairFunction;



public class WordCount {
	public static void main(String[] args) throws Exception {
		String inputFile = args[0];
		String outputFile = args[1];
		// Create a Java Spark Context.
		SparkConf conf = new SparkConf().setAppName("wordCount");
		JavaSparkContext sc = new JavaSparkContext(conf);

		// Load our input data.
		JavaRDD<String> input = sc.textFile(inputFile);
		// Split up into words.
		JavaRDD<String> words = input.flatMap(new FlatMapFunction<String, String>() {
			public Iterable<String> call(String x) {
				// create list of words from string by splitting on spaces
```

```
                        List<String> list = Arrays.asList(x.split(" "));

                        // iteratre through the list and convert to lowercase
                        // also eliminate punctuation by replacing it with an empty string
                for(int i=0; i<list.size(); i++){
                        String word = list.get(i);
                        word = word.toLowerCase().replaceAll("[^A-Za-z0-9]", "");
                        list.set(i, word);
                }

                // return the list
                return list;
                        }
                });
                // Transform into word and count.
                JavaPairRDD<String, Integer> counts = words.mapToPair(new
PairFunction<String, String, Integer>() {
                        public Tuple2<String, Integer> call(String x) {
                                return new Tuple2(x, 1);
                        }
                }).reduceByKey(new Function2<Integer, Integer, Integer>() {
                        public Integer call(Integer x, Integer y) {
                                return x + y;
                        }
                });
                // Save the word count back out to a text file, causing evaluation.
                counts.saveAsTextFile(outputFile);
        }
}
```

Problem 2. Write a working WordCount program using Spark Scala API that reads a file, e.g. Ulysis/4300.txt from a local file system directory and writes the results of your calculations to a local file. To improve your word count, remove any punctuation that might have attached itself to your words. Also transform all words into lower case so that the capitalization does not affect the word count. The original code is provided in the attached mini-example-scala.tar file. That archive also contains Scala Build Tool build.sbt file. Run your program and demonstrate that it works. Submit working code inside the customary MS Word Document. Describe steps in your program.

Scala API Terminal Commands

```
// First I build the Spark application from WordCount.scala. The sbt tool is used and compiles
// the files into an executable program.
[cloudera@localhost mini-examples-scala]$ sbt clean package
[info] Set current project to mini-example (in build file:/home/cloudera/mini-examples-scala/)
[success] Total time: 2 s, completed Feb 26, 2016 1:13:20 PM
[info] Updating {file:/home/cloudera/mini-examples-scala/}mini-examples-scala...
[info] Resolving org.fusesource.jansi#jansi;1.4 ...
[info] Done updating.
[info] Compiling 2 Scala sources to
/home/cloudera/mini-examples-scala/target/scala-2.10/classes...
[warn] Multiple main classes detected.  Run 'show discoveredMainClasses' to see the list
[info] Packaging
/home/cloudera/mini-examples-scala/target/scala-2.10/mini-example_2.10-0.0.1.jar ...
[info] Done packaging.
[success] Total time: 70 s, completed Feb 26, 2016 1:14:30 PM

// Next I submit the scala application to Spark. It counts the words in 4300.txt and outputs the
// results to the scalacounts folder

[cloudera@localhost mini-examples-scala]$ spark-submit --class edu.hu.examples.WordCount
./target/scala-2.10/mini-example_2.10-0.0.1.jar ulysses/4300.txt scalacounts
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in
[jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in
[jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/02/26 13:21:03 INFO spark.SparkContext: Running Spark version 1.5.0-cdh5.5.2
16/02/26 13:21:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
16/02/26 13:21:05 WARN util.Utils: Your hostname, localhost.localdomain resolves to a
loopback address: 127.0.0.1; using 172.16.40.129 instead (on interface eth0)
16/02/26 13:21:05 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
16/02/26 13:21:05 INFO spark.SecurityManager: Changing view acls to: cloudera
16/02/26 13:21:05 INFO spark.SecurityManager: Changing modify acls to: cloudera
16/02/26 13:21:05 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui
acls disabled; users with view permissions: Set(cloudera); users with modify permissions:
Set(cloudera)
```

16/02/26 13:21:08 INFO slf4j.Slf4jLogger: Slf4jLogger started
16/02/26 13:21:08 INFO Remoting: Starting remoting
16/02/26 13:21:09 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@172.16.40.129:42903]
16/02/26 13:21:09 INFO Remoting: Remoting now listens on addresses: [akka.tcp://sparkDriver@172.16.40.129:42903]
16/02/26 13:21:09 INFO util.Utils: Successfully started service 'sparkDriver' on port 42903.
16/02/26 13:21:09 INFO spark.SparkEnv: Registering MapOutputTracker
16/02/26 13:21:09 INFO spark.SparkEnv: Registering BlockManagerMaster
16/02/26 13:21:09 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-e02aadbd-980b-47f6-a391-57b9b58a8a9d
16/02/26 13:21:09 INFO storage.MemoryStore: MemoryStore started with capacity 534.5 MB
16/02/26 13:21:09 INFO spark.HttpFileServer: HTTP File server directory is /tmp/spark-ae257e9a-4dd7-4787-8e70-2229cc78b84f/httpd-de943a78-5a48-4fe1-8c64-9ac5b38a4a8a
16/02/26 13:21:09 INFO spark.HttpServer: Starting HTTP Server
16/02/26 13:21:09 INFO server.Server: jetty-8.y.z-SNAPSHOT
16/02/26 13:21:09 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0:37571
16/02/26 13:21:09 INFO util.Utils: Successfully started service 'HTTP file server' on port 37571.
16/02/26 13:21:09 INFO spark.SparkEnv: Registering OutputCommitCoordinator
16/02/26 13:21:10 INFO server.Server: jetty-8.y.z-SNAPSHOT
16/02/26 13:21:10 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
16/02/26 13:21:10 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
16/02/26 13:21:10 INFO ui.SparkUI: Started SparkUI at http://172.16.40.129:4040
16/02/26 13:21:10 INFO spark.SparkContext: Added JAR file:/home/cloudera/mini-examples-scala/./target/scala-2.10/mini-example_2.10-0.0.1.jar at http://172.16.40.129:37571/jars/mini-example_2.10-0.0.1.jar with timestamp 1456521670485
16/02/26 13:21:10 WARN metrics.MetricsSystem: Using default name DAGScheduler for source because spark.app.id is not set.
16/02/26 13:21:10 INFO executor.Executor: Starting executor ID driver on host localhost
16/02/26 13:21:11 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 45780.
16/02/26 13:21:11 INFO netty.NettyBlockTransferService: Server created on 45780
16/02/26 13:21:11 INFO storage.BlockManagerMaster: Trying to register BlockManager
16/02/26 13:21:11 INFO storage.BlockManagerMasterEndpoint: Registering block manager localhost:45780 with 534.5 MB RAM, BlockManagerId(driver, localhost, 45780)
16/02/26 13:21:11 INFO storage.BlockManagerMaster: Registered BlockManager
16/02/26 13:21:14 INFO storage.MemoryStore: ensureFreeSpace(136720) called with curMem=0, maxMem=560497950
16/02/26 13:21:14 INFO storage.MemoryStore: Block broadcast_0 stored as values in memory (estimated size 133.5 KB, free 534.4 MB)

16/02/26 13:21:14 INFO storage.MemoryStore: ensureFreeSpace(15386) called with curMem=136720, maxMem=560497950
16/02/26 13:21:14 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 15.0 KB, free 534.4 MB)
16/02/26 13:21:14 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:45780 (size: 15.0 KB, free: 534.5 MB)
16/02/26 13:21:14 INFO spark.SparkContext: Created broadcast 0 from textFile at WordCount.scala:17
16/02/26 13:21:15 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
16/02/26 13:21:15 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/26 13:21:16 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/02/26 13:21:16 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/02/26 13:21:16 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
16/02/26 13:21:16 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
16/02/26 13:21:16 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/02/26 13:21:16 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/26 13:21:16 INFO spark.SparkContext: Starting job: saveAsTextFile at WordCount.scala:24
16/02/26 13:21:16 INFO scheduler.DAGScheduler: Registering RDD 3 (map at WordCount.scala:22)
16/02/26 13:21:16 INFO scheduler.DAGScheduler: Got job 0 (saveAsTextFile at WordCount.scala:24) with 1 output partitions
16/02/26 13:21:16 INFO scheduler.DAGScheduler: Final stage: ResultStage 1(saveAsTextFile at WordCount.scala:24)
16/02/26 13:21:16 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 0)
16/02/26 13:21:16 INFO scheduler.DAGScheduler: Missing parents: List(ShuffleMapStage 0)
16/02/26 13:21:17 INFO scheduler.DAGScheduler: Submitting ShuffleMapStage 0 (MapPartitionsRDD[3] at map at WordCount.scala:22), which has no missing parents
16/02/26 13:21:17 INFO storage.MemoryStore: ensureFreeSpace(4056) called with curMem=152106, maxMem=560497950
16/02/26 13:21:17 INFO storage.MemoryStore: Block broadcast_1 stored as values in memory (estimated size 4.0 KB, free 534.4 MB)
16/02/26 13:21:17 INFO storage.MemoryStore: ensureFreeSpace(2305) called with curMem=156162, maxMem=560497950

16/02/26 13:21:17 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 2.3 KB, free 534.4 MB)
16/02/26 13:21:17 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:45780 (size: 2.3 KB, free: 534.5 MB)
16/02/26 13:21:17 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:861
16/02/26 13:21:17 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 0 (MapPartitionsRDD[3] at map at WordCount.scala:22)
16/02/26 13:21:17 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
16/02/26 13:21:17 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, partition 0,PROCESS_LOCAL, 2219 bytes)
16/02/26 13:21:17 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
16/02/26 13:21:17 INFO executor.Executor: Fetching http://172.16.40.129:37571/jars/mini-example_2.10-0.0.1.jar with timestamp 1456521670485
16/02/26 13:21:18 INFO util.Utils: Fetching http://172.16.40.129:37571/jars/mini-example_2.10-0.0.1.jar to /tmp/spark-ae257e9a-4dd7-4787-8e70-2229cc78b84f/userFiles-08ccbeb6-4076-417f-b478-5a693d8fc2e1/fetchFileTemp3131626345619925944.tmp
16/02/26 13:21:18 INFO executor.Executor: Adding file:/tmp/spark-ae257e9a-4dd7-4787-8e70-2229cc78b84f/userFiles-08ccbeb6-4076-417f-b478-5a693d8fc2e1/mini-example_2.10-0.0.1.jar to class loader
16/02/26 13:21:18 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/user/cloudera/ulysses/4300.txt:0+1573079
16/02/26 13:21:26 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 2253 bytes result sent to driver
16/02/26 13:21:26 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 8815 ms on localhost (1/1)
16/02/26 13:21:26 INFO scheduler.DAGScheduler: ShuffleMapStage 0 (map at WordCount.scala:22) finished in 8.852 s
16/02/26 13:21:26 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
16/02/26 13:21:26 INFO scheduler.DAGScheduler: looking for newly runnable stages
16/02/26 13:21:26 INFO scheduler.DAGScheduler: running: Set()
16/02/26 13:21:26 INFO scheduler.DAGScheduler: waiting: Set(ResultStage 1)
16/02/26 13:21:26 INFO scheduler.DAGScheduler: failed: Set()
16/02/26 13:21:26 INFO scheduler.DAGScheduler: Missing parents for ResultStage 1: List()
16/02/26 13:21:26 INFO scheduler.DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[5] at saveAsTextFile at WordCount.scala:24), which is now runnable
16/02/26 13:21:26 INFO storage.MemoryStore: ensureFreeSpace(135376) called with curMem=158467, maxMem=560497950
16/02/26 13:21:26 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory (estimated size 132.2 KB, free 534.3 MB)

16/02/26 13:21:26 INFO storage.MemoryStore: ensureFreeSpace(46356) called with curMem=293843, maxMem=560497950
16/02/26 13:21:26 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 45.3 KB, free 534.2 MB)
16/02/26 13:21:26 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:45780 (size: 45.3 KB, free: 534.5 MB)
16/02/26 13:21:26 INFO spark.SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:861
16/02/26 13:21:26 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (MapPartitionsRDD[5] at saveAsTextFile at WordCount.scala:24)
16/02/26 13:21:27 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
16/02/26 13:21:27 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, partition 0,PROCESS_LOCAL, 1970 bytes)
16/02/26 13:21:27 INFO executor.Executor: Running task 0.0 in stage 1.0 (TID 1)
16/02/26 13:21:27 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
16/02/26 13:21:27 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 12 ms
16/02/26 13:21:28 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/26 13:21:30 INFO output.FileOutputCommitter: Saved output of task 'attempt_201602261321_0001_m_000000_1' to hdfs://localhost:8020/user/cloudera/scalacounts/_temporary/0/task_201602261321_0001_m_000000
16/02/26 13:21:30 INFO mapred.SparkHadoopMapRedUtil: attempt_201602261321_0001_m_000000_1: Committed
16/02/26 13:21:30 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 1). 2080 bytes result sent to driver
16/02/26 13:21:30 INFO scheduler.DAGScheduler: ResultStage 1 (saveAsTextFile at WordCount.scala:24) finished in 3.076 s
16/02/26 13:21:30 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 3077 ms on localhost (1/1)
16/02/26 13:21:30 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
16/02/26 13:21:30 INFO scheduler.DAGScheduler: Job 0 finished: saveAsTextFile at WordCount.scala:24, took 13.287871 s
16/02/26 13:21:30 INFO spark.SparkContext: Invoking stop() from shutdown hook
16/02/26 13:21:30 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/metrics/json,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/kill,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/api,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/,null}

16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/static,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/executors/threadDump/json,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/executors/threadDump,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/executors/json,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/executors,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/environment/json,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/environment,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage/rdd/json,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage/rdd,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage/json,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/pool/json,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/pool,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/stage/json,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/stage,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/json,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs/job/json,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs/job,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs/json,null}
16/02/26 13:21:30 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs,null}
16/02/26 13:21:30 INFO ui.SparkUI: Stopped Spark web UI at http://172.16.40.129:4040

16/02/26 13:21:30 INFO scheduler.DAGScheduler: Stopping DAGScheduler
16/02/26 13:21:30 INFO spark.MapOutputTrackerMasterEndpoint:
MapOutputTrackerMasterEndpoint stopped!
16/02/26 13:21:30 INFO storage.MemoryStore: MemoryStore cleared
16/02/26 13:21:30 INFO storage.BlockManager: BlockManager stopped
16/02/26 13:21:30 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
16/02/26 13:21:30 INFO
scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:
OutputCommitCoordinator stopped!
16/02/26 13:21:30 INFO spark.SparkContext: Successfully stopped SparkContext
16/02/26 13:21:30 INFO util.ShutdownHookManager: Shutdown hook called
16/02/26 13:21:30 INFO util.ShutdownHookManager: Deleting directory
/tmp/spark-ae257e9a-4dd7-4787-8e70-2229cc78b84f

Spark Scala API - First 20 Results

(reunion,2)
(bone,11)
(luminary,2)
(monologue,1)
(rinbad,1)
(alavatar,1)
(blandly,2)
(gavin,2)
(serfs,1)
(bombast,1)
(hem,2)
(fred,2)
(stinks,4)
(flutiest,1)
(fuller,2)
(tough,3)
(jade,1)
(twinging,1)
(jove,5)
(crying,10)

Description of WordCount.scala

I used the provided code in mini-examples-scala.tar. The inputs locations are set, a Spark Configuration is initiated, a Spark Context is initiated, the text file is loaded by the Spark Context, and the lines are split by spaces.

Next I added code to eliminate punctuation and convert to lowercase. When the words are mapped, the original file maps to the word and integer 1. My code instead maps to a lowercase conversion of the word that has the punctuation replaced with empty strings. This cleans the words by removing the punctuation that previously differentiated equivalent words and removes case differences too by converting to lowercase.

After the mapping and cleaning, the reduction is done to count the occurrences of the words. The new counts are saved as a text file.

WordCount.scala

```
/**
 * Illustrates flatMap + countByValue for wordcount.
 */
package edu.hu.examples

import org.apache.spark._
import org.apache.spark.SparkContext._

object WordCount {
   def main(args: Array[String]) {
     val inputFile = args(0)
     val outputFile = args(1)
     val conf = new SparkConf().setAppName("wordCount")
     // Create a Scala Spark Context.
     val sc = new SparkContext(conf)
     // Load our input data.
     val input =  sc.textFile(inputFile)
     // Split up into words.
     val words = input.flatMap(line => line.split(" "))

     // Transform into word and count.
     val counts = words.map(word => (word.toLowerCase.replaceAll("[^A-Za-z0-9]", ""),
1)).reduceByKey{case (x, y) => x + y}
     // Save the word count back out to a text file, causing evaluation.
     counts.saveAsTextFile(outputFile)
   }
}
```

Problem 3. Write a working WordCount script using Spark Python API. Read Ulysis (4300.txt) file from an HDFS directory and write the results of your calculations to an HDFS file. To

improve your word count, remove any punctuation that might have attached itself to your words. Also transform all words into lower case so that the capitalization does not affect the word count. Run your script using submit-spark tool and demonstrate that it works. Submit working code. Describe steps in your program in the MS Word document.

Python API Terminal Commands

```
[cloudera@localhost mini-examples-python]$ spark-submit my_script.py
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in
[jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in
[jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/02/26 14:38:27 INFO spark.SparkContext: Running Spark version 1.5.0-cdh5.5.2
16/02/26 14:38:28 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
16/02/26 14:38:29 WARN util.Utils: Your hostname, localhost.localdomain resolves to a
loopback address: 127.0.0.1; using 172.16.40.129 instead (on interface eth0)
16/02/26 14:38:29 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
16/02/26 14:38:29 INFO spark.SecurityManager: Changing view acls to: cloudera
16/02/26 14:38:29 INFO spark.SecurityManager: Changing modify acls to: cloudera
16/02/26 14:38:29 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui
acls disabled; users with view permissions: Set(cloudera); users with modify permissions:
Set(cloudera)
16/02/26 14:38:31 INFO slf4j.Slf4jLogger: Slf4jLogger started
16/02/26 14:38:31 INFO Remoting: Starting remoting
16/02/26 14:38:31 INFO Remoting: Remoting started; listening on addresses
:[akka.tcp://sparkDriver@172.16.40.129:39733]
16/02/26 14:38:31 INFO Remoting: Remoting now listens on addresses:
[akka.tcp://sparkDriver@172.16.40.129:39733]
16/02/26 14:38:31 INFO util.Utils: Successfully started service 'sparkDriver' on port 39733.
16/02/26 14:38:31 INFO spark.SparkEnv: Registering MapOutputTracker
16/02/26 14:38:31 INFO spark.SparkEnv: Registering BlockManagerMaster
16/02/26 14:38:31 INFO storage.DiskBlockManager: Created local directory at
/tmp/blockmgr-2699a9dc-b2ab-4497-a5f4-f08e1238cf39
16/02/26 14:38:31 INFO storage.MemoryStore: MemoryStore started with capacity 534.5 MB
```

16/02/26 14:38:31 INFO spark.HttpFileServer: HTTP File server directory is /tmp/spark-0ee6db58-79dd-40ce-b0ef-f498f402b5f9/httpd-19e27785-a61f-49ef-b1ba-a88503988ac1
16/02/26 14:38:31 INFO spark.HttpServer: Starting HTTP Server
16/02/26 14:38:32 INFO server.Server: jetty-8.y.z-SNAPSHOT
16/02/26 14:38:32 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0:34862
16/02/26 14:38:32 INFO util.Utils: Successfully started service 'HTTP file server' on port 34862.
16/02/26 14:38:32 INFO spark.SparkEnv: Registering OutputCommitCoordinator
16/02/26 14:38:32 INFO server.Server: jetty-8.y.z-SNAPSHOT
16/02/26 14:38:32 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
16/02/26 14:38:32 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
16/02/26 14:38:32 INFO ui.SparkUI: Started SparkUI at http://172.16.40.129:4040
16/02/26 14:38:33 INFO util.Utils: Copying /home/cloudera/mini-examples-python/my_script.py to /tmp/spark-0ee6db58-79dd-40ce-b0ef-f498f402b5f9/userFiles-416a5aea-31a9-4860-b763-b1598922250e/my_script.py
16/02/26 14:38:33 INFO spark.SparkContext: Added file file:/home/cloudera/mini-examples-python/my_script.py at file:/home/cloudera/mini-examples-python/my_script.py with timestamp 1456526313110
16/02/26 14:38:33 WARN metrics.MetricsSystem: Using default name DAGScheduler for source because spark.app.id is not set.
16/02/26 14:38:33 INFO executor.Executor: Starting executor ID driver on host localhost
16/02/26 14:38:33 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 59622.
16/02/26 14:38:33 INFO netty.NettyBlockTransferService: Server created on 59622
16/02/26 14:38:33 INFO storage.BlockManagerMaster: Trying to register BlockManager
16/02/26 14:38:33 INFO storage.BlockManagerMasterEndpoint: Registering block manager localhost:59622 with 534.5 MB RAM, BlockManagerId(driver, localhost, 59622)
16/02/26 14:38:33 INFO storage.BlockManagerMaster: Registered BlockManager
16/02/26 14:38:35 INFO storage.MemoryStore: ensureFreeSpace(235376) called with curMem=0, maxMem=560497950
16/02/26 14:38:35 INFO storage.MemoryStore: Block broadcast_0 stored as values in memory (estimated size 229.9 KB, free 534.3 MB)
16/02/26 14:38:35 INFO storage.MemoryStore: ensureFreeSpace(21339) called with curMem=235376, maxMem=560497950
16/02/26 14:38:35 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 20.8 KB, free 534.3 MB)
16/02/26 14:38:35 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:59622 (size: 20.8 KB, free: 534.5 MB)
16/02/26 14:38:35 INFO spark.SparkContext: Created broadcast 0 from textFile at NativeMethodAccessorImpl.java:-2

16/02/26 14:38:36 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
16/02/26 14:38:37 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/26 14:38:37 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/02/26 14:38:37 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/02/26 14:38:37 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
16/02/26 14:38:37 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
16/02/26 14:38:37 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/02/26 14:38:37 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/26 14:38:38 INFO spark.SparkContext: Starting job: saveAsTextFile at NativeMethodAccessorImpl.java:-2
16/02/26 14:38:38 INFO scheduler.DAGScheduler: Registering RDD 3 (reduceByKey at /home/cloudera/mini-examples-python/my_script.py:14)
16/02/26 14:38:38 INFO scheduler.DAGScheduler: Got job 0 (saveAsTextFile at NativeMethodAccessorImpl.java:-2) with 1 output partitions
16/02/26 14:38:38 INFO scheduler.DAGScheduler: Final stage: ResultStage 1(saveAsTextFile at NativeMethodAccessorImpl.java:-2)
16/02/26 14:38:38 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 0)
16/02/26 14:38:38 INFO scheduler.DAGScheduler: Missing parents: List(ShuffleMapStage 0)
16/02/26 14:38:38 INFO scheduler.DAGScheduler: Submitting ShuffleMapStage 0 (PairwiseRDD[3] at reduceByKey at /home/cloudera/mini-examples-python/my_script.py:14), which has no missing parents
16/02/26 14:38:38 INFO storage.MemoryStore: ensureFreeSpace(8704) called with curMem=256715, maxMem=560497950
16/02/26 14:38:38 INFO storage.MemoryStore: Block broadcast_1 stored as values in memory (estimated size 8.5 KB, free 534.3 MB)
16/02/26 14:38:38 INFO storage.MemoryStore: ensureFreeSpace(5355) called with curMem=265419, maxMem=560497950
16/02/26 14:38:38 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 5.2 KB, free 534.3 MB)
16/02/26 14:38:38 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:59622 (size: 5.2 KB, free: 534.5 MB)
16/02/26 14:38:38 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:861

16/02/26 14:38:38 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 0 (PairwiseRDD[3] at reduceByKey at /home/cloudera/mini-examples-python/my_script.py:14)
16/02/26 14:38:38 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
16/02/26 14:38:38 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, partition 0,PROCESS_LOCAL, 2213 bytes)
16/02/26 14:38:38 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
16/02/26 14:38:38 INFO executor.Executor: Fetching file:/home/cloudera/mini-examples-python/my_script.py with timestamp 1456526313110
16/02/26 14:38:38 INFO util.Utils: /home/cloudera/mini-examples-python/my_script.py has been previously copied to /tmp/spark-0ee6db58-79dd-40ce-b0ef-f498f402b5f9/userFiles-416a5aea-31a9-4860-b763-b1598922250e/my_script.py
16/02/26 14:38:38 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/user/cloudera/ulysses/4300.txt:0+1573079
16/02/26 14:38:42 INFO python.PythonRDD: Times: total = 4090, boot = 297, init = 386, finish = 3407
16/02/26 14:38:42 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 2317 bytes result sent to driver
16/02/26 14:38:42 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 4514 ms on localhost (1/1)
16/02/26 14:38:42 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
16/02/26 14:38:42 INFO scheduler.DAGScheduler: ShuffleMapStage 0 (reduceByKey at /home/cloudera/mini-examples-python/my_script.py:14) finished in 4.596 s
16/02/26 14:38:42 INFO scheduler.DAGScheduler: looking for newly runnable stages
16/02/26 14:38:42 INFO scheduler.DAGScheduler: running: Set()
16/02/26 14:38:42 INFO scheduler.DAGScheduler: waiting: Set(ResultStage 1)
16/02/26 14:38:42 INFO scheduler.DAGScheduler: failed: Set()
16/02/26 14:38:42 INFO scheduler.DAGScheduler: Missing parents for ResultStage 1: List()
16/02/26 14:38:42 INFO scheduler.DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[8] at saveAsTextFile at NativeMethodAccessorImpl.java:-2), which is now runnable
16/02/26 14:38:43 INFO storage.MemoryStore: ensureFreeSpace(139496) called with curMem=270774, maxMem=560497950
16/02/26 14:38:43 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory (estimated size 136.2 KB, free 534.1 MB)
16/02/26 14:38:43 INFO storage.MemoryStore: ensureFreeSpace(49073) called with curMem=410270, maxMem=560497950
16/02/26 14:38:43 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 47.9 KB, free 534.1 MB)
16/02/26 14:38:43 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:59622 (size: 47.9 KB, free: 534.5 MB)

16/02/26 14:38:43 INFO spark.SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:861
16/02/26 14:38:43 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (MapPartitionsRDD[8] at saveAsTextFile at NativeMethodAccessorImpl.java:-2)
16/02/26 14:38:43 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
16/02/26 14:38:43 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, partition 0,PROCESS_LOCAL, 1964 bytes)
16/02/26 14:38:43 INFO executor.Executor: Running task 0.0 in stage 1.0 (TID 1)
16/02/26 14:38:43 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
16/02/26 14:38:43 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 27 ms
16/02/26 14:38:43 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/26 14:38:45 INFO python.PythonRDD: Times: total = 1633, boot = -912, init = 987, finish = 1558
16/02/26 14:38:46 INFO output.FileOutputCommitter: Saved output of task 'attempt_201602261438_0001_m_000000_1' to hdfs://localhost:8020/user/cloudera/pycounts4/_temporary/0/task_201602261438_0001_m_000000
16/02/26 14:38:46 INFO mapred.SparkHadoopMapRedUtil: attempt_201602261438_0001_m_000000_1: Committed
16/02/26 14:38:46 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 1). 2144 bytes result sent to driver
16/02/26 14:38:46 INFO scheduler.DAGScheduler: ResultStage 1 (saveAsTextFile at NativeMethodAccessorImpl.java:-2) finished in 3.267 s
16/02/26 14:38:46 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 3265 ms on localhost (1/1)
16/02/26 14:38:46 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
16/02/26 14:38:46 INFO scheduler.DAGScheduler: Job 0 finished: saveAsTextFile at NativeMethodAccessorImpl.java:-2, took 8.546944 s
16/02/26 14:38:46 INFO spark.SparkContext: Invoking stop() from shutdown hook
16/02/26 14:38:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/metrics/json,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/kill,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/api,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/static,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/threadDump/json,null}

16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/executors/threadDump,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/executors/json,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/executors,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/environment/json,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/environment,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage/rdd/json,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage/rdd,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage/json,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/pool/json,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/pool,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/stage/json,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/stage,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/json,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs/job/json,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs/job,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs/json,null}
16/02/26 14:38:46 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs,null}
16/02/26 14:38:47 INFO ui.SparkUI: Stopped Spark web UI at http://172.16.40.129:4040
16/02/26 14:38:47 INFO scheduler.DAGScheduler: Stopping DAGScheduler
16/02/26 14:38:47 INFO spark.MapOutputTrackerMasterEndpoint:
MapOutputTrackerMasterEndpoint stopped!
16/02/26 14:38:47 INFO storage.MemoryStore: MemoryStore cleared

16/02/26 14:38:47 INFO storage.BlockManager: BlockManager stopped
16/02/26 14:38:47 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
16/02/26 14:38:47 INFO
scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:
OutputCommitCoordinator stopped!
16/02/26 14:38:47 INFO spark.SparkContext: Successfully stopped SparkContext
16/02/26 14:38:47 INFO util.ShutdownHookManager: Shutdown hook called
16/02/26 14:38:47 INFO util.ShutdownHookManager: Deleting directory
/tmp/spark-0ee6db58-79dd-40ce-b0ef-f498f402b5f9
16/02/26 14:38:47 INFO remote.RemoteActorRefProvider$RemotingTerminator: Shutting down
remote daemon.

Spark Python API - First 20 Results

(u'fawn', 3)
(u'highspliced', 1)
(u'piffpaff', 1)
(u'askew', 4)
(u'woods', 6)
(u'clotted', 4)
(u'spiders', 1)
(u'phenomenologist', 1)
(u'hanging', 29)
(u'noctambules', 1)
(u'comically', 1)
(u'houyhnhnm', 1)
(u'sevens', 1)
(u'canes', 1)
(u'sprague', 1)
(u'scutter', 1)
(u'originality', 2)
(u'alphabetic', 1)
(u'stipulate', 1)
(u'pigment', 1)

Description of WordCount.py

I used the framework for python presented in lecture. A Spark Configuration is initiated, a Spark Context is created, and the file is loaded into the context.

Next the flatmap function is used to split the lines by spaces. In the following map function for the RDD, I add code to convert all letters to lowercase and eliminate punctuation.

The .lower() string function is used and re.sub() converts all non-alphanumerics to the empty string.

As suggested, I then use the reduceByKey() function on the mapped data. This sums all the instances of different words. Because it's possible to have the empty string as a word when punctuation is removed, I then filter the data removing instances where a == "". Lastly, I save the results as a text file.

WordCount.py

```
import re
from pyspark import SparkConf, SparkContext
conf = SparkConf().setMaster("local").setAppName("MyApp")
sc = SparkContext(conf = conf)
theFile = sc.textFile("/user/cloudera/ulysses/4300.txt")

# split the line by space and map to (word, 1)
# also convert to lowercase and eliminate punctuation
counts = theFile.flatMap(lambda line: line.split(" ")).map(lambda word: (re.sub("[^A-Za-z0-9]", "",
word.lower()), 1))

# reduce the mapped data and sum instances of words
counts = counts.reduceByKey(lambda a, b: a +b)

# filter out the empty string
counts = counts.filter(lambda (a,b): a!="")


# save to file
counts.saveAsTextFile("pycounts4")
```

Problem 4. In a Spark API of your choice, write a working BigramCount program which would count occurrences of every pair of consecutive words. You should clean your words just as you did in the previous problem by removing punctuations and cases. However, do not count two words separated by a point at the end of a sentence as a bigram. If you are an experienced programmer add to the bigram count word pairs in which the first word is the last word on the line and the second word is the first word on the subsequent line. If you are not an experienced programmer, than do not do it. Test your program on a small text file, where for comparison, you could identify bigrams manually. Run your program on Ulysis(4300.txt) file and demonstrate that it works. Provide us with the total count of your bigrams, first 20 bigrams and all bigrams containing word "heaven". Read your file from the local operating system and write results to the local operating system. Include working code in the MS Word Document. Submit the file with the complete working code separately. Describe steps in your program in the MS Word document.

Bigram Terminal Commands

```
// I clean, compile, and package the program with Maven. This creates the executable
// application from the Java files.
[cloudera@localhost bigram]$ mvn clean && mvn compile && mvn package
[INFO] Scanning for projects...
[INFO]
[INFO] ------------------------------------------------------------------------
[INFO] Building example 0.0.1
[INFO] ------------------------------------------------------------------------
[INFO]
[INFO] --- maven-clean-plugin:2.5:clean (default-clean) @ spark-example ---
[INFO] Deleting /home/cloudera/bigram/target
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time: 3.830 s
[INFO] Finished at: 2016-02-26T15:05:54-08:00
[INFO] Final Memory: 5M/29M
[INFO] ------------------------------------------------------------------------
[INFO] Scanning for projects...
[INFO]
[INFO] ------------------------------------------------------------------------
[INFO] Building example 0.0.1
[INFO] ------------------------------------------------------------------------
[INFO]
[INFO] --- maven-resources-plugin:2.6:resources (default-resources) @ spark-example ---
[WARNING] Using platform encoding (UTF-8 actually) to copy filtered resources, i.e. build is
platform dependent!
[INFO] skip non existing resourceDirectory /home/cloudera/bigram/src/main/resources
[INFO]
[INFO] --- maven-compiler-plugin:3.1:compile (default-compile) @ spark-example ---
[INFO] Changes detected - recompiling the module!
[WARNING] File encoding has not been set, using platform encoding UTF-8, i.e. build is
platform dependent!
[INFO] Compiling 2 source files to /home/cloudera/bigram/target/classes
[WARNING] /home/cloudera/bigram/src/main/java/edu/hu/examples/WordCount.java:
/home/cloudera/bigram/src/main/java/edu/hu/examples/WordCount.java uses unchecked or
unsafe operations.
[WARNING] /home/cloudera/bigram/src/main/java/edu/hu/examples/WordCount.java:
Recompile with -Xlint:unchecked for details.
```

```
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time: 19.604 s
[INFO] Finished at: 2016-02-26T15:06:19-08:00
[INFO] Final Memory: 22M/54M
[INFO] ------------------------------------------------------------------------
[INFO] Scanning for projects...
[INFO]
[INFO] ------------------------------------------------------------------------
[INFO] Building example 0.0.1
[INFO] ------------------------------------------------------------------------
[INFO]
[INFO] --- maven-resources-plugin:2.6:resources (default-resources) @ spark-example ---
[WARNING] Using platform encoding (UTF-8 actually) to copy filtered resources, i.e. build is
platform dependent!
[INFO] skip non existing resourceDirectory /home/cloudera/bigram/src/main/resources
[INFO]
[INFO] --- maven-compiler-plugin:3.1:compile (default-compile) @ spark-example ---
[INFO] Changes detected - recompiling the module!
[WARNING] File encoding has not been set, using platform encoding UTF-8, i.e. build is
platform dependent!
[INFO] Compiling 2 source files to /home/cloudera/bigram/target/classes
[WARNING] /home/cloudera/bigram/src/main/java/edu/hu/examples/WordCount.java:
/home/cloudera/bigram/src/main/java/edu/hu/examples/WordCount.java uses unchecked or
unsafe operations.
[WARNING] /home/cloudera/bigram/src/main/java/edu/hu/examples/WordCount.java:
Recompile with -Xlint:unchecked for details.
[INFO]
[INFO] --- maven-resources-plugin:2.6:testResources (default-testResources) @ spark-example
---
[WARNING] Using platform encoding (UTF-8 actually) to copy filtered resources, i.e. build is
platform dependent!
[INFO] skip non existing resourceDirectory /home/cloudera/bigram/src/test/resources
[INFO]
[INFO] --- maven-compiler-plugin:3.1:testCompile (default-testCompile) @ spark-example ---
[INFO] No sources to compile
[INFO]
[INFO] --- maven-surefire-plugin:2.12.4:test (default-test) @ spark-example ---
[INFO] No tests to run.
[INFO]
[INFO] --- maven-jar-plugin:2.4:jar (default-jar) @ spark-example ---
[INFO] Building jar: /home/cloudera/bigram/target/spark-example-0.0.1.jar
```

[INFO] --
[INFO] BUILD SUCCESS
[INFO] --
[INFO] Total time: 8.216 s
[INFO] Finished at: 2016-02-26T15:06:30-08:00
[INFO] Final Memory: 24M/57M
[INFO] --

// Next I submit the Java application to Spark. It uses the compiled .jar file to run the bigram
// program on the local version of Ulysses' text 4300.txt and outputs to the directed local file

[cloudera@localhost bigram]$ spark-submit --class edu.hu.examples.WordCount
./target/spark-example-0.0.1.jar file:///mnt/hgfs/VM_shared/ulysses/4300.txt
file:///mnt/hgfs/VM_shared/bcounts11

SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in
[jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in
[jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/02/26 16:26:11 INFO spark.SparkContext: Running Spark version 1.5.0-cdh5.5.2
16/02/26 16:26:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
16/02/26 16:26:12 WARN util.Utils: Your hostname, localhost.localdomain resolves to a
loopback address: 127.0.0.1; using 172.16.40.129 instead (on interface eth0)
16/02/26 16:26:12 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
16/02/26 16:26:12 INFO spark.SecurityManager: Changing view acls to: cloudera
16/02/26 16:26:12 INFO spark.SecurityManager: Changing modify acls to: cloudera
16/02/26 16:26:12 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui
acls disabled; users with view permissions: Set(cloudera); users with modify permissions:
Set(cloudera)
16/02/26 16:26:14 INFO slf4j.Slf4jLogger: Slf4jLogger started
16/02/26 16:26:14 INFO Remoting: Starting remoting
16/02/26 16:26:14 INFO Remoting: Remoting started; listening on addresses
:[akka.tcp://sparkDriver@172.16.40.129:40226]
16/02/26 16:26:14 INFO Remoting: Remoting now listens on addresses:
[akka.tcp://sparkDriver@172.16.40.129:40226]
16/02/26 16:26:14 INFO util.Utils: Successfully started service 'sparkDriver' on port 40226.
16/02/26 16:26:14 INFO spark.SparkEnv: Registering MapOutputTracker
16/02/26 16:26:14 INFO spark.SparkEnv: Registering BlockManagerMaster

16/02/26 16:26:15 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-8997492b-c0e6-40ee-bdb6-7f9d047bd7c7
16/02/26 16:26:15 INFO storage.MemoryStore: MemoryStore started with capacity 534.5 MB
16/02/26 16:26:15 INFO spark.HttpFileServer: HTTP File server directory is /tmp/spark-740bfdae-570b-4e17-abfd-3a508e98c868/httpd-47771a21-43d3-4631-870d-f1321cb7bd67
16/02/26 16:26:15 INFO spark.HttpServer: Starting HTTP Server
16/02/26 16:26:15 INFO server.Server: jetty-8.y.z-SNAPSHOT
16/02/26 16:26:15 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0:37517
16/02/26 16:26:15 INFO util.Utils: Successfully started service 'HTTP file server' on port 37517.
16/02/26 16:26:15 INFO spark.SparkEnv: Registering OutputCommitCoordinator
16/02/26 16:26:15 INFO server.Server: jetty-8.y.z-SNAPSHOT
16/02/26 16:26:15 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
16/02/26 16:26:15 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
16/02/26 16:26:15 INFO ui.SparkUI: Started SparkUI at http://172.16.40.129:4040
16/02/26 16:26:16 INFO spark.SparkContext: Added JAR file:/home/cloudera/bigram/./target/spark-example-0.0.1.jar at http://172.16.40.129:37517/jars/spark-example-0.0.1.jar with timestamp 1456532776109
16/02/26 16:26:16 WARN metrics.MetricsSystem: Using default name DAGScheduler for source because spark.app.id is not set.
16/02/26 16:26:16 INFO executor.Executor: Starting executor ID driver on host localhost
16/02/26 16:26:16 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 41251.
16/02/26 16:26:16 INFO netty.NettyBlockTransferService: Server created on 41251
16/02/26 16:26:16 INFO storage.BlockManagerMaster: Trying to register BlockManager
16/02/26 16:26:16 INFO storage.BlockManagerMasterEndpoint: Registering block manager localhost:41251 with 534.5 MB RAM, BlockManagerId(driver, localhost, 41251)
16/02/26 16:26:16 INFO storage.BlockManagerMaster: Registered BlockManager
16/02/26 16:26:18 INFO storage.MemoryStore: ensureFreeSpace(136720) called with curMem=0, maxMem=560497950
16/02/26 16:26:18 INFO storage.MemoryStore: Block broadcast_0 stored as values in memory (estimated size 133.5 KB, free 534.4 MB)
16/02/26 16:26:19 INFO storage.MemoryStore: ensureFreeSpace(15386) called with curMem=136720, maxMem=560497950
16/02/26 16:26:19 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 15.0 KB, free 534.4 MB)
16/02/26 16:26:19 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:41251 (size: 15.0 KB, free: 534.5 MB)
16/02/26 16:26:19 INFO spark.SparkContext: Created broadcast 0 from textFile at WordCount.java:32
16/02/26 16:26:20 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.

16/02/26 16:26:20 INFO mapred.FileInputFormat: Total input paths to process : 1
16/02/26 16:26:21 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/02/26 16:26:21 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/02/26 16:26:21 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
16/02/26 16:26:21 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
16/02/26 16:26:21 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/02/26 16:26:21 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/26 16:26:21 INFO spark.SparkContext: Starting job: saveAsTextFile at WordCount.java:91
16/02/26 16:26:21 INFO scheduler.DAGScheduler: Registering RDD 3 (mapToPair at WordCount.java:78)
16/02/26 16:26:21 INFO scheduler.DAGScheduler: Got job 0 (saveAsTextFile at WordCount.java:91) with 1 output partitions
16/02/26 16:26:21 INFO scheduler.DAGScheduler: Final stage: ResultStage 1(saveAsTextFile at WordCount.java:91)
16/02/26 16:26:21 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 0)
16/02/26 16:26:21 INFO scheduler.DAGScheduler: Missing parents: List(ShuffleMapStage 0)
16/02/26 16:26:21 INFO scheduler.DAGScheduler: Submitting ShuffleMapStage 0 (MapPartitionsRDD[3] at mapToPair at WordCount.java:78), which has no missing parents
16/02/26 16:26:21 INFO storage.MemoryStore: ensureFreeSpace(4776) called with curMem=152106, maxMem=560497950
16/02/26 16:26:21 INFO storage.MemoryStore: Block broadcast_1 stored as values in memory (estimated size 4.7 KB, free 534.4 MB)
16/02/26 16:26:21 INFO storage.MemoryStore: ensureFreeSpace(2700) called with curMem=156882, maxMem=560497950
16/02/26 16:26:21 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 2.6 KB, free 534.4 MB)
16/02/26 16:26:21 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:41251 (size: 2.6 KB, free: 534.5 MB)
16/02/26 16:26:21 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:861
16/02/26 16:26:21 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 0 (MapPartitionsRDD[3] at mapToPair at WordCount.java:78)
16/02/26 16:26:21 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
16/02/26 16:26:21 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, partition 0,PROCESS_LOCAL, 2204 bytes)

16/02/26 16:26:21 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
16/02/26 16:26:21 INFO executor.Executor: Fetching
http://172.16.40.129:37517/jars/spark-example-0.0.1.jar with timestamp 1456532776109
16/02/26 16:26:22 INFO util.Utils: Fetching
http://172.16.40.129:37517/jars/spark-example-0.0.1.jar to
/tmp/spark-740bfdae-570b-4e17-abfd-3a508e98c868/userFiles-5abaca43-6ccc-409f-9b8f-ffa67
6672e0f/fetchFileTemp5913333008822433379.tmp
16/02/26 16:26:22 INFO executor.Executor: Adding
file:/tmp/spark-740bfdae-570b-4e17-abfd-3a508e98c868/userFiles-5abaca43-6ccc-409f-9b8f-ffa
676672e0f/spark-example-0.0.1.jar to class loader
16/02/26 16:26:22 INFO rdd.HadoopRDD: Input split:
file:/mnt/hgfs/VM_shared/ulysses/4300.txt:0+1573079
16/02/26 16:26:28 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 2253 bytes
result sent to driver
16/02/26 16:26:28 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in
6597 ms on localhost (1/1)
16/02/26 16:26:28 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks
have all completed, from pool
16/02/26 16:26:28 INFO scheduler.DAGScheduler: ShuffleMapStage 0 (mapToPair at
WordCount.java:78) finished in 6.675 s
16/02/26 16:26:28 INFO scheduler.DAGScheduler: looking for newly runnable stages
16/02/26 16:26:28 INFO scheduler.DAGScheduler: running: Set()
16/02/26 16:26:28 INFO scheduler.DAGScheduler: waiting: Set(ResultStage 1)
16/02/26 16:26:28 INFO scheduler.DAGScheduler: failed: Set()
16/02/26 16:26:28 INFO scheduler.DAGScheduler: Missing parents for ResultStage 1: List()
16/02/26 16:26:28 INFO scheduler.DAGScheduler: Submitting ResultStage 1
(MapPartitionsRDD[5] at saveAsTextFile at WordCount.java:91), which is now runnable
16/02/26 16:26:29 INFO storage.MemoryStore: ensureFreeSpace(135480) called with
curMem=159582, maxMem=560497950
16/02/26 16:26:29 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory
(estimated size 132.3 KB, free 534.3 MB)
16/02/26 16:26:29 INFO storage.MemoryStore: ensureFreeSpace(46397) called with
curMem=295062, maxMem=560497950
16/02/26 16:26:29 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in
memory (estimated size 45.3 KB, free 534.2 MB)
16/02/26 16:26:29 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on
localhost:41251 (size: 45.3 KB, free: 534.5 MB)
16/02/26 16:26:29 INFO spark.SparkContext: Created broadcast 2 from broadcast at
DAGScheduler.scala:861
16/02/26 16:26:29 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from
ResultStage 1 (MapPartitionsRDD[5] at saveAsTextFile at WordCount.java:91)
16/02/26 16:26:29 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 1 tasks

16/02/26 16:26:29 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, partition 0,PROCESS_LOCAL, 1966 bytes)
16/02/26 16:26:29 INFO executor.Executor: Running task 0.0 in stage 1.0 (TID 1)
16/02/26 16:26:29 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
16/02/26 16:26:29 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 19 ms
16/02/26 16:26:30 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/02/26 16:26:31 INFO storage.BlockManagerInfo: Removed broadcast_1_piece0 on localhost:41251 in memory (size: 2.6 KB, free: 534.5 MB)
16/02/26 16:26:32 INFO output.FileOutputCommitter: Saved output of task 'attempt_201602261626_0001_m_000000_1' to file:/mnt/hgfs/VM_shared/bcounts11/_temporary/0/task_201602261626_0001_m_000000
16/02/26 16:26:32 INFO mapred.SparkHadoopMapRedUtil: attempt_201602261626_0001_m_000000_1: Committed
16/02/26 16:26:32 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 1). 2080 bytes result sent to driver
16/02/26 16:26:32 INFO scheduler.DAGScheduler: ResultStage 1 (saveAsTextFile at WordCount.java:91) finished in 2.899 s
16/02/26 16:26:32 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 2902 ms on localhost (1/1)
16/02/26 16:26:32 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
16/02/26 16:26:32 INFO scheduler.DAGScheduler: Job 0 finished: saveAsTextFile at WordCount.java:91, took 10.520455 s
16/02/26 16:26:32 INFO spark.SparkContext: Invoking stop() from shutdown hook
16/02/26 16:26:32 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/metrics/json,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/kill,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/api,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/static,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/threadDump/json,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/threadDump,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/json,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors,null}

16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/environment/json,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/environment,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage/rdd/json,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage/rdd,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage/json,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/storage,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/pool/json,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/pool,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/stage/json,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/stage,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages/json,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/stages,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs/job/json,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs/job,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs/json,null}
16/02/26 16:26:32 INFO handler.ContextHandler: stopped
o.s.j.s.ServletContextHandler{/jobs,null}
16/02/26 16:26:32 INFO ui.SparkUI: Stopped Spark web UI at http://172.16.40.129:4040
16/02/26 16:26:32 INFO scheduler.DAGScheduler: Stopping DAGScheduler
16/02/26 16:26:33 INFO spark.MapOutputTrackerMasterEndpoint:
MapOutputTrackerMasterEndpoint stopped!
16/02/26 16:26:33 INFO storage.MemoryStore: MemoryStore cleared
16/02/26 16:26:33 INFO storage.BlockManager: BlockManager stopped
16/02/26 16:26:33 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
16/02/26 16:26:33 INFO
scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:
OutputCommitCoordinator stopped!
16/02/26 16:26:33 INFO spark.SparkContext: Successfully stopped SparkContext

16/02/26 16:26:33 INFO util.ShutdownHookManager: Shutdown hook called
16/02/26 16:26:33 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-740bfdae-570b-4e17-abfd-3a508e98c868
16/02/26 16:26:33 INFO remote.RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
16/02/26 16:26:33 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.

Bigram Results

4300.txt File

Total Count: 128,672

First 20 Results:
(or, modes,1)
(shillings, offered,1)
(brawnyhanded, hairylegged,1)
(some, law,2)
(but, want,1)
(bloom, dittoed,1)
(one, shilling,1)
(from, berkeley,1)
(i, put,20)
(a, capacious,2)
(gunwale, of,1)
(the, womancity,1)
(received, from,2)
(afterwits, miss,1)
(lift, it,1)
(jinglejaunty, blazes,1)
(kerry, cows,1)
(he, pops,1)
(good, clip,1)
(word, a,1)

Bigrams with “heaven”:
(heaven, was,2)
(visit, heaven,1)
(to, heaven,8)

(in, heaven,8)
(nearer, heaven,1)
(heaven, theodore,1)
(and, heaven,1)
(like, heaven,1)
(hearts, heaven,1)
(heaven, and,3)
(heaven, murmuring,1)
(heaven, ned,1)
(heaven, spilt,1)
(heaven, when,1)
(into, heaven,1)
(by, heaven,2)
(gracious, heaven,1)
(heaven, a,1)
(a, heaven,1)
(of, heaven,7)
(heaven, had,1)
(heaven, i,1)
(heaven, 4,1)
(heaven, if,1)
(heaven, theres,1)
(heaven, becomes,1)
(seventh, heaven,1)
(heaven, by,2)
(heaven, hight,1)
(heaven, ever,1)
(heaven, calling,1)
(heaven, foretold,1)
(thank, heaven,1)

Small Test File:
This is a “test” program. It verifies the “test” program works.
Any questions?

Small Test File Results:
(test, program,2)
(the, test,1)
(this, is,1)
(is, a,1)
(program, works,1)
(a, test,1)
(any, questions,1)

(it, verifies,1)
(verifies, the,1)

Description of the bigram code

The bigram program is similar to the WordCount.java file from problem 1. It loads the input and output file locations. It creates a Spark Configuration, a Java Spark Context, and loads the text file to a JavaRDD. It then calls the .flatmap function on the input. A new FlatMapFunction is created that contains a function call() to create an Iterable<String>.

Here I add my code that's different than the template provided. The string is split on the spaces. My program iterates through the strings, converting to lowercase and and replacing all punctuation with the empty string except end-of-sentence punctuation. If the word is not empty it's added to the new cleaned list.

I then have a loop that iterates through the cleaned words. If the previous word was not null, which designates the end of a sentence or the first word, a bigram is made. That bigram is added to an array list. Otherwise, the current word is set to the previous word for the next iteration. If an end-of-sentence punctuation is detected, the previous word is set to null to prevent a bigram from occurring during the next iteration.

After the bigrams are totaled, they are transformed into bigrams and counts. Then the reduceToKey() function is used to total the occurrences of the bigrams. Lastly, this data is saved to a text file at the output location.

Bigram Code

Note: The class is named WordCount.java because it used that class's format.

```
/**
 * Illustrates a wordcount in Java
 */
package edu.hu.examples;

import java.util.*;
import java.util.Arrays;
import java.util.List;
import java.lang.Iterable;

import scala.Tuple2;

import org.apache.commons.lang.StringUtils;

import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaRDD;
import org.apache.spark.api.java.JavaPairRDD;
import org.apache.spark.api.java.JavaSparkContext;
```

```
import org.apache.spark.api.java.function.FlatMapFunction;
import org.apache.spark.api.java.function.Function2;
import org.apache.spark.api.java.function.PairFunction;


public class WordCount {
  public static void main(String[] args) throws Exception {
    String inputFile = args[0];
    String outputFile = args[1];
    // Create a Java Spark Context.
    SparkConf conf = new SparkConf().setAppName("wordCount");
                    JavaSparkContext sc = new JavaSparkContext(conf);
    // Load our input data.
    JavaRDD<String> input = sc.textFile(inputFile);
    // Split up into words.
    JavaRDD<String> words = input.flatMap(
      new FlatMapFunction<String, String>() {
        public Iterable<String> call(String x) {
          List<String> list = Arrays.asList(x.split(" "));
          List<String> cleanedList = new ArrayList<String>();


          for(int i=0; i<list.size(); i++){
            String word = list.get(i);

            // remove everything except alphanumerics and end-of-sentence punctuation
            word = word.toLowerCase().replaceAll("[^A-Za-z0-9.?!]", "");

            // if word contains no alphanumerics or end-of-sentence punctuation, ignore it
            if(!word.equals("")){
              cleanedList.add(word);
            }
          }

          List<String> bigrams = new ArrayList<String>();
          String previous = null;


          for(int i=0; i<cleanedList.size(); i++){
            String word = cleanedList.get(i);
            String bigramString = null;

            if(previous != null){
```

```
            bigramString = previous + ", " + word.replaceAll("[.!?]","");
            bigrams.add(bigramString);
          }
          previous = word;

          // if words contains end-of-sentence punctuation, set previous to null
          if(!word.replaceAll("[.!?]","").equals(word)){
            previous = null;
          }
        }

        return bigrams;
      }});


    // Transform into word and count.
    JavaPairRDD<String, Integer> counts = words.mapToPair(
      new PairFunction<String, String, Integer>(){
        public Tuple2<String, Integer> call(String x){
          return new Tuple2(x, 1);
        }
      }

    ).reduceByKey(
      new Function2<Integer, Integer, Integer>(){
        public Integer call(Integer x, Integer y){ return x + y;}
      }
    );
    // Save the word count back out to a text file, causing evaluation.
    counts.saveAsTextFile(outputFile);
          }
}
```