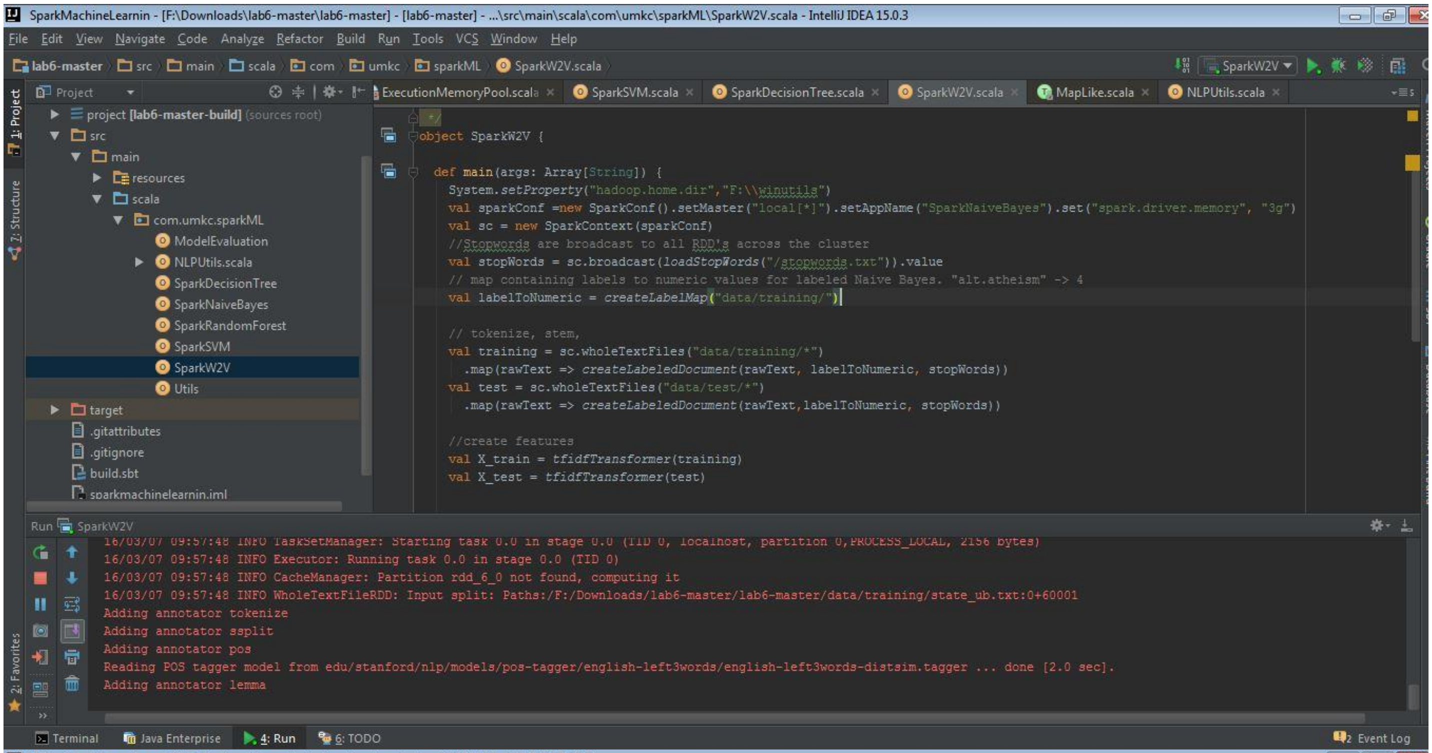
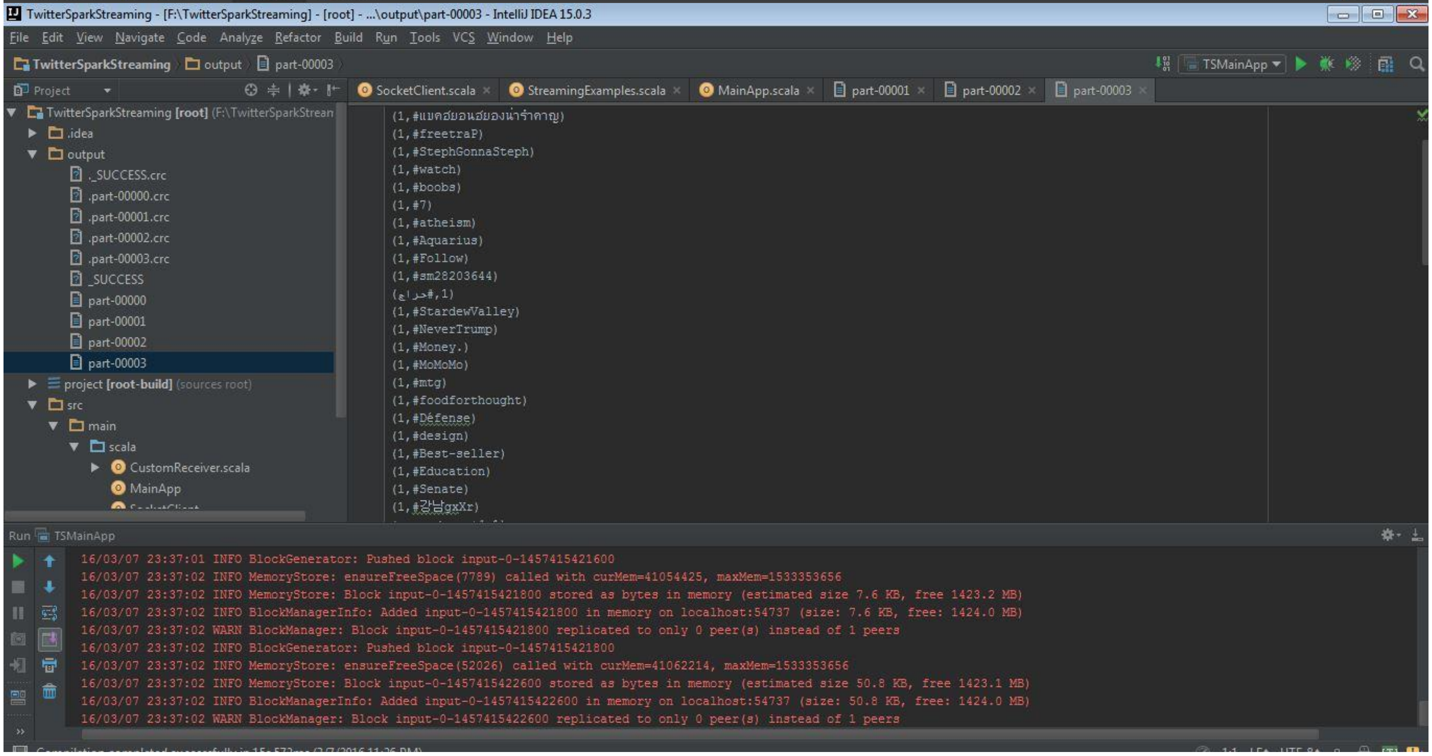
REPORT LAB 5 and 6: CS 5542


SparkMachineLearnin - [F:\Downloads\lab6-master\lab6-master] - [lab6-master] - ...\src\main\scala\com\umkc\sparkML\SparkW2V.scala - IntelliJ IDEA 15.0.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
lab6-master src main scala com umkc sparkML SparkW2V.scala
SparkW2V
Project
ExecutionMemoryPool.scala
SparkSVM.scala
SparkDecisionTree.scala
SparkW2V.scala
MapLike.scala
NLPUtils.scala
project [lab6-master-build] (sources root)
src
main
resources
scala
com.umkc.sparkML
ModelEvaluation
NLPUtils.scala
SparkDecisionTree
SparkNaiveBayes
SparkRandomForest
SparkSVM
SparkW2V
Utils
target
.gitattributes
.gitignore
build.sbt
sparkmachinelearnin.iml
Run SparkW2V
Terminal
Java Enterprise
4: Run
6: TODO
Event Log

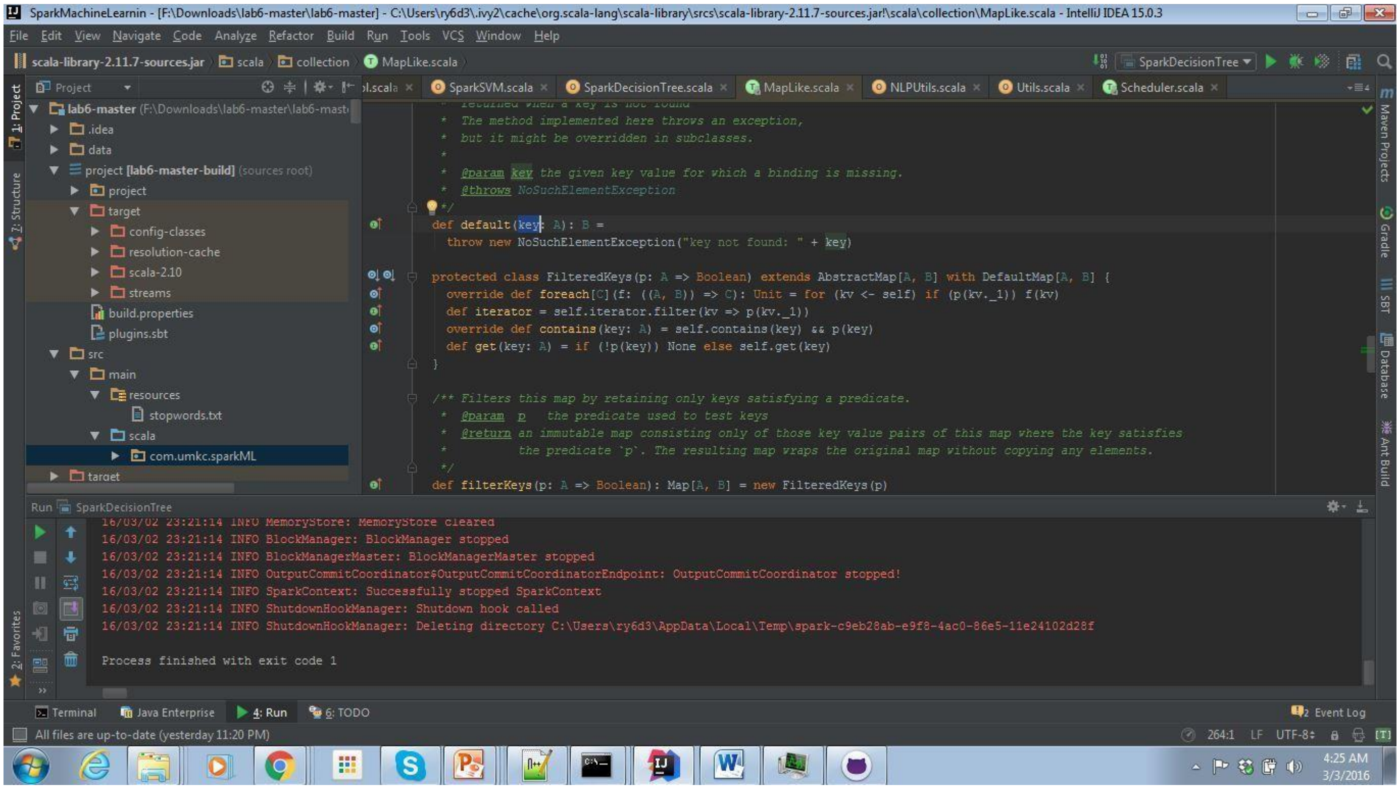
```
*/
object SparkW2V {

  def main(args: Array[String]) {
    System.setProperty("hadoop.home.dir","F:\\winutils")
    val sparkConf =new SparkConf().setMaster("local[*]").setAppName("SparkNaiveBayes").set("spark.driver.memory", "3g")
    val sc = new SparkContext(sparkConf)
    //Stopwords are broadcast to all RDD's across the cluster
    val stopWords = sc.broadcast(loadStopWords("/stopwords.txt")).value
    // map containing labels to numeric values for labeled Naive Bayes. "alt.atheism" -> 4
    val labelToNumeric = createLabelMap("data/training/")

    // tokenize, stem,
    val training = sc.wholeTextFiles("data/training/*")
      .map(rawText => createLabeledDocument(rawText, labelToNumeric, stopWords))
    val test = sc.wholeTextFiles("data/test/*")
      .map(rawText => createLabeledDocument(rawText,labelToNumeric, stopWords))

    //create features
    val X_train = tfidfTransformer(training)
    val X_test = tfidfTransformer(test)
```

```
16/03/07 09:57:48 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, partition 0,PROCESS_LOCAL, 2156 bytes)
16/03/07 09:57:48 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
16/03/07 09:57:48 INFO CacheManager: Partition rdd_6_0 not found, computing it
16/03/07 09:57:48 INFO WholeTextFileRDD: Input split: Paths:/F:/Downloads/lab6-master/lab6-master/data/training/state_ub.txt:0+60001
Adding annotator tokenize
Adding annotator ssplit
Adding annotator pos
Reading POS tagger model from edu/stanford/nlp/models/pos-tagger/english-left3words/english-left3words-distsim.tagger ... done [2.0 sec].
Adding annotator lemma
```


TwitterSparkStreaming - [F:\TwitterSparkStreaming] - [root] - ...\output\part-00003 - IntelliJ IDEA 15.0.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
TwitterSparkStreaming output part-00003
TSMainApp
Project
SocketClient.scala
StreamingExamples.scala
MainApp.scala
part-00001
part-00002
part-00003
TwitterSparkStreaming [root] (F:\TwitterSparkStrean
.idea
output
._SUCCESS.crc
.part-00000.crc
.part-00001.crc
.part-00002.crc
.part-00003.crc
_SUCCESS
part-00000
part-00001
part-00002
part-00003
project [root-build] (sources root)
src
main
scala
CustomReceiver.scala
MainApp
Run TSMainApp

```
(1,#แมดฮ่อนฮ่อนน่ารำคาญ)
(1,#freetraP)
(1,#StephGonnaSteph)
(1,#watch)
(1,#boobs)
(1,#7)
(1,#atheism)
(1,#Aquarius)
(1,#Follow)
(1,#sm28203644)
(1,#حراع)
(1,#StardewValley)
(1,#NeverTrump)
(1,#Money.)
(1,#MoMoMo)
(1,#mtg)
(1,#foodforthought)
(1,#Défense)
(1,#design)
(1,#Best-seller)
(1,#Education)
(1,#Senate)
(1,#강남gxXr)
```

```
16/03/07 23:37:01 INFO BlockGenerator: Pushed block input-0-1457415421600
16/03/07 23:37:02 INFO MemoryStore: ensureFreeSpace(7789) called with curMem=41054425, maxMem=1533353656
16/03/07 23:37:02 INFO MemoryStore: Block input-0-1457415421800 stored as bytes in memory (estimated size 7.6 KB, free 1423.2 MB)
16/03/07 23:37:02 INFO BlockManagerInfo: Added input-0-1457415421800 in memory on localhost:54737 (size: 7.6 KB, free: 1424.0 MB)
16/03/07 23:37:02 WARN BlockManager: Block input-0-1457415421800 replicated to only 0 peer(s) instead of 1 peers
16/03/07 23:37:02 INFO BlockGenerator: Pushed block input-0-1457415421800
16/03/07 23:37:02 INFO MemoryStore: ensureFreeSpace(52026) called with curMem=41062214, maxMem=1533353656
16/03/07 23:37:02 INFO MemoryStore: Block input-0-1457415422600 stored as bytes in memory (estimated size 50.8 KB, free 1423.1 MB)
16/03/07 23:37:02 INFO BlockManagerInfo: Added input-0-1457415422600 in memory on localhost:54737 (size: 50.8 KB, free: 1424.0 MB)
16/03/07 23:37:02 WARN BlockManager: Block input-0-1457415422600 replicated to only 0 peer(s) instead of 1 peers
```



```
   *  returned when a key is not found
   *  The method implemented here throws an exception,
   *  but it might be overridden in subclasses.
   *
   *  @param key the given key value for which a binding is missing.
   *  @throws NoSuchElementException
   */
  def default(key: A): B =
    throw new NoSuchElementException("key not found: " + key)

  protected class FilteredKeys(p: A => Boolean) extends AbstractMap[A, B] with DefaultMap[A, B] {
    override def foreach[C](f: ((A, B)) => C): Unit = for (kv <- self) if (p(kv._1)) f(kv)
    def iterator = self.iterator.filter(kv => p(kv._1))
    override def contains(key: A) = self.contains(key) && p(key)
    def get(key: A) = if (!p(key)) None else self.get(key)
  }

  /** Filters this map by retaining only keys satisfying a predicate.
   *  @param  p   the predicate used to test keys
   *  @return an immutable map consisting only of those key value pairs of this map where the key satisfies
   *          the predicate `p`. The resulting map wraps the original map without copying any elements.
   */
  def filterKeys(p: A => Boolean): Map[A, B] = new FilteredKeys(p)
```

```
16/03/02 23:21:14 INFO MemoryStore: MemoryStore cleared
16/03/02 23:21:14 INFO BlockManager: BlockManager stopped
16/03/02 23:21:14 INFO BlockManagerMaster: BlockManagerMaster stopped
16/03/02 23:21:14 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
16/03/02 23:21:14 INFO SparkContext: Successfully stopped SparkContext
16/03/02 23:21:14 INFO ShutdownHookManager: Shutdown hook called
16/03/02 23:21:14 INFO ShutdownHookManager: Deleting directory C:\Users\ry6d3\AppData\Local\Temp\spark-c9eb28ab-e9f8-4ac0-86e5-11e24102d28f

Process finished with exit code 1
```

All files are up-to-date (yesterday 11:20 PM)