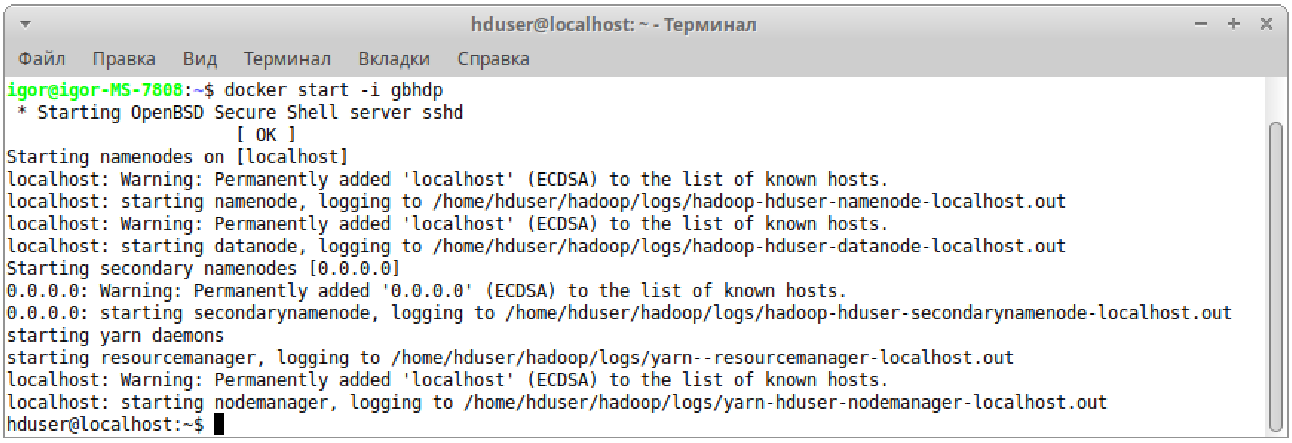
1. Проведите анализ тональности датасета *IMDB*: обучите модель на Train.csv, после чего проверьте её на Valid.csv

2. Посчитайте получившуюся точность (accuracy — количество правильных предсказаний от общего количества ответов) модели. Ваш код должен использовать ML Pipelines и использовать трансформеры для подготовки данных к обучению

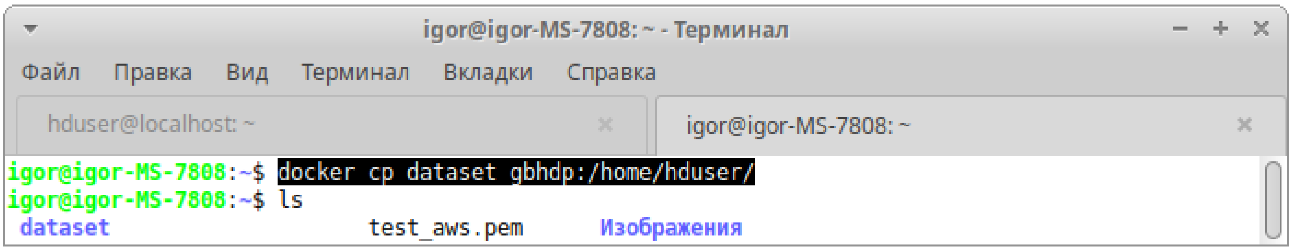
Стартуем остановленный контейнер с Hadoop и Spark:

$ docker start -i gbhdp


hduser@localhost: ~ - Терминал
Файл Правка Вид Терминал Вкладки Справка

```
igor@igor-MS-7808:~$ docker start -i gbhdp
 * Starting OpenBSD Secure Shell server sshd
                                               [ OK ]
Starting namenodes on [localhost]
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: starting namenode, logging to /home/hduser/hadoop/logs/hadoop-hduser-namenode-localhost.out
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: starting datanode, logging to /home/hduser/hadoop/logs/hadoop-hduser-datanode-localhost.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
0.0.0.0: starting secondarynamenode, logging to /home/hduser/hadoop/logs/hadoop-hduser-secondarynamenode-localhost.out
starting yarn daemons
starting resourcemanager, logging to /home/hduser/hadoop/logs/yarn--resourcemanager-localhost.out
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: starting nodemanager, logging to /home/hduser/hadoop/logs/yarn-hduser-nodemanager-localhost.out
hduser@localhost:~$ 
```

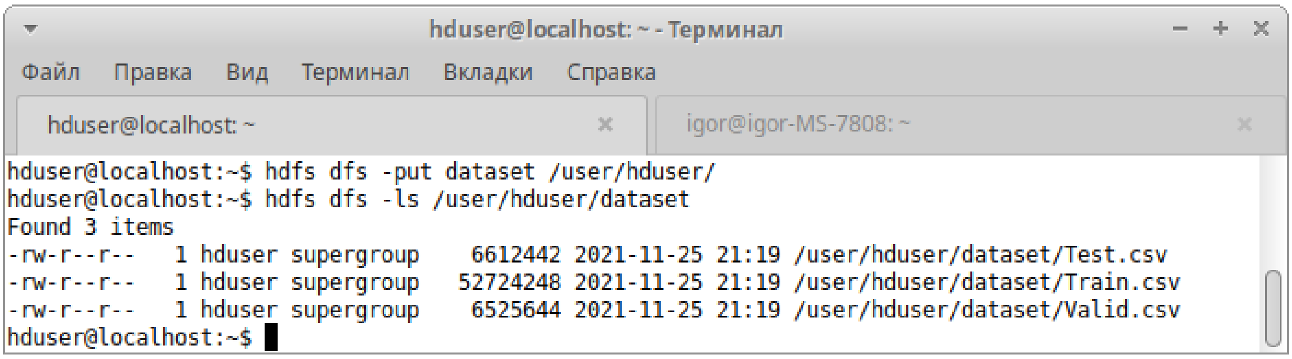
Копируем датасет в контейнер: docker cp dataset gbhdp:/home/hduser/


igor@igor-MS-7808: ~ - Терминал
Файл Правка Вид Терминал Вкладки Справка
hduser@localhost: ~
igor@igor-MS-7808: ~

```
igor@igor-MS-7808:~$ docker cp dataset gbhdp:/home/hduser/
igor@igor-MS-7808:~$ ls
 dataset                  test_aws.pem      Изображения
```

Копируем директорию в hdfs: hdfs dfs -put dataset /user/hduser/

Проверим, что файлы в hdfs: hdfs dfs -ls /user/hduser/dataset

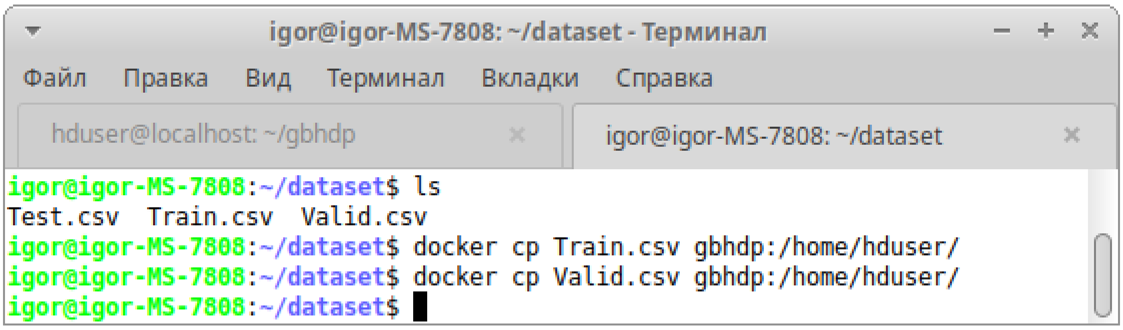

hduser@localhost: ~ - Терминал
Файл Правка Вид Терминал Вкладки Справка
hduser@localhost: ~
igor@igor-MS-7808: ~

```
hduser@localhost:~$ hdfs dfs -put dataset /user/hduser/
hduser@localhost:~$ hdfs dfs -ls /user/hduser/dataset
Found 3 items
-rw-r--r--   1 hduser supergroup    6612442 2021-11-25 21:19 /user/hduser/dataset/Test.csv
-rw-r--r--   1 hduser supergroup   52724248 2021-11-25 21:19 /user/hduser/dataset/Train.csv
-rw-r--r--   1 hduser supergroup    6525644 2021-11-25 21:19 /user/hduser/dataset/Valid.csv
hduser@localhost:~$ 
```

Копируем датасет в контейнер:

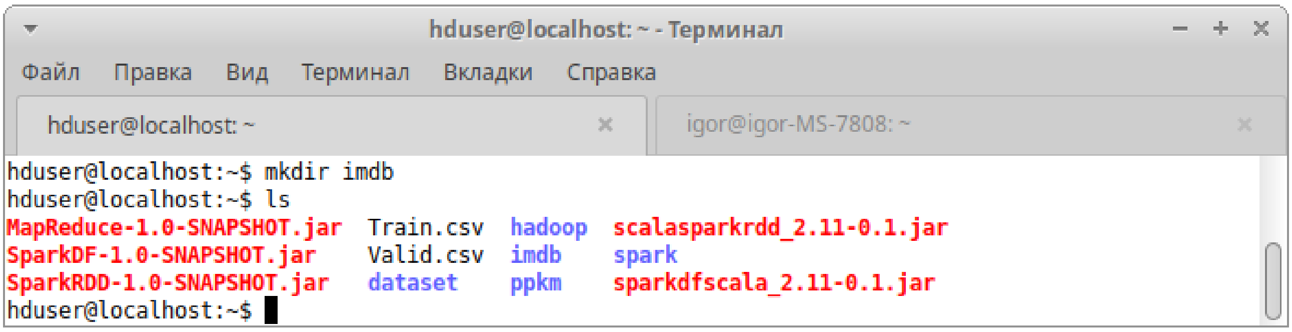
docker cp Train.csv gbhdp:/home/hduser/

docker cp Valid.csv gbhdp:/home/hduser/


igor@igor-MS-7808: ~/dataset - Терминал
Файл Правка Вид Терминал Вкладки Справка
hduser@localhost: ~/gbhdp
igor@igor-MS-7808: ~/dataset

```
igor@igor-MS-7808:~/dataset$ ls
Test.csv  Train.csv  Valid.csv
igor@igor-MS-7808:~/dataset$ docker cp Train.csv gbhdp:/home/hduser/
igor@igor-MS-7808:~/dataset$ docker cp Valid.csv gbhdp:/home/hduser/
igor@igor-MS-7808:~/dataset$ 
```

создаём директорию imdb: mkdir imdb

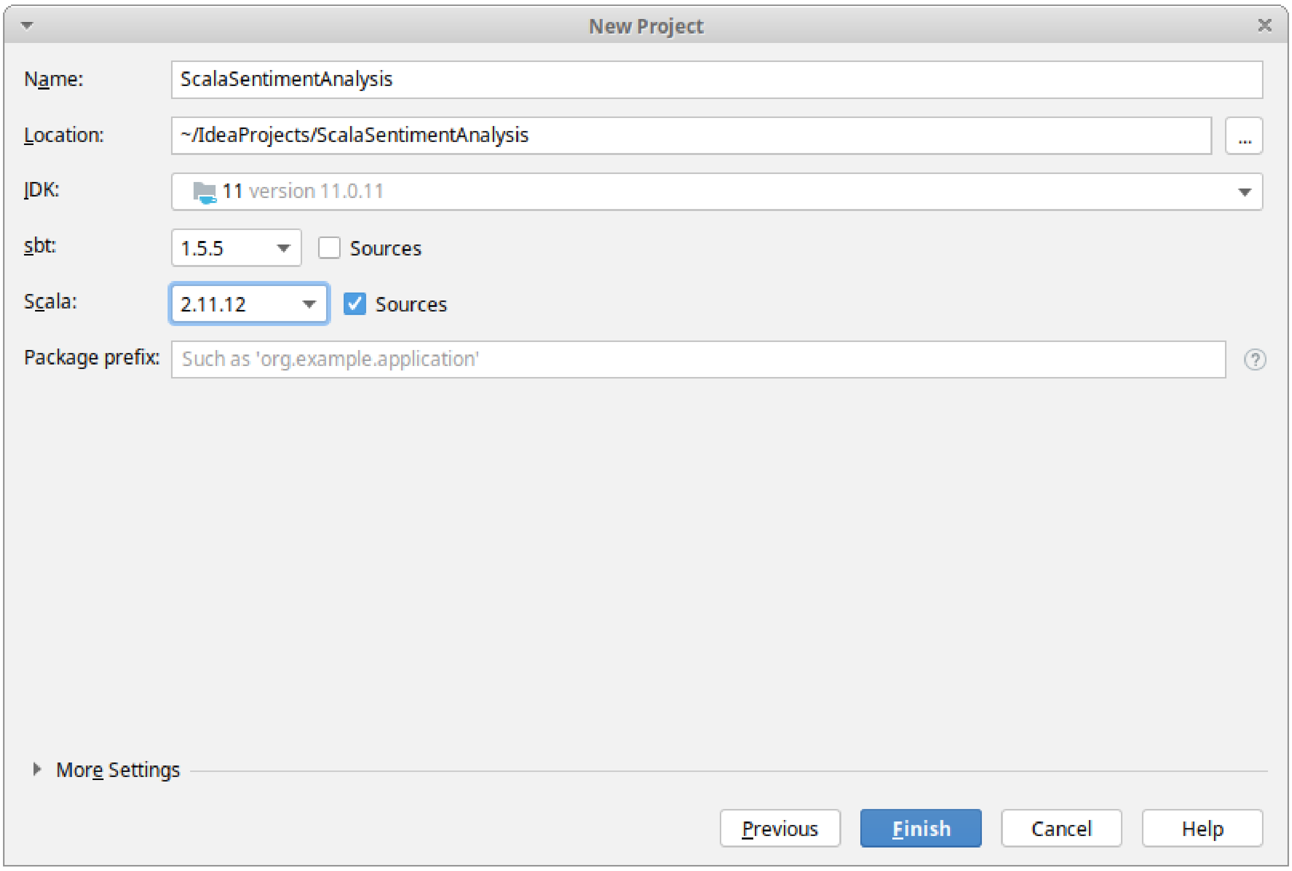

hduser@localhost: ~ - Терминал
Файл Правка Вид Терминал Вкладки Справка
hduser@localhost: ~
igor@igor-MS-7808: ~

```
hduser@localhost:~$ mkdir imdb
hduser@localhost:~$ ls
MapReduce-1.0-SNAPSHOT.jar  Train.csv  hadoop  scalasparkrdd_2.11-0.1.jar
SparkDF-1.0-SNAPSHOT.jar    Valid.csv  imdb    spark
SparkRDD-1.0-SNAPSHOT.jar   dataset    ppkm    sparkdfscala_2.11-0.1.jar
hduser@localhost:~$ 
```

hdfs dfs -put Train.csv /user/hduser/imdb/

hdfs dfs -put Valid.csv /user/hduser/imdb/

SCALA датасета IMDB: ПРОЕКТ В IDEA



Name: ScalaSentimentAnalysis

Location: ~/IdeaProjects/ScalaSentimentAnalysis

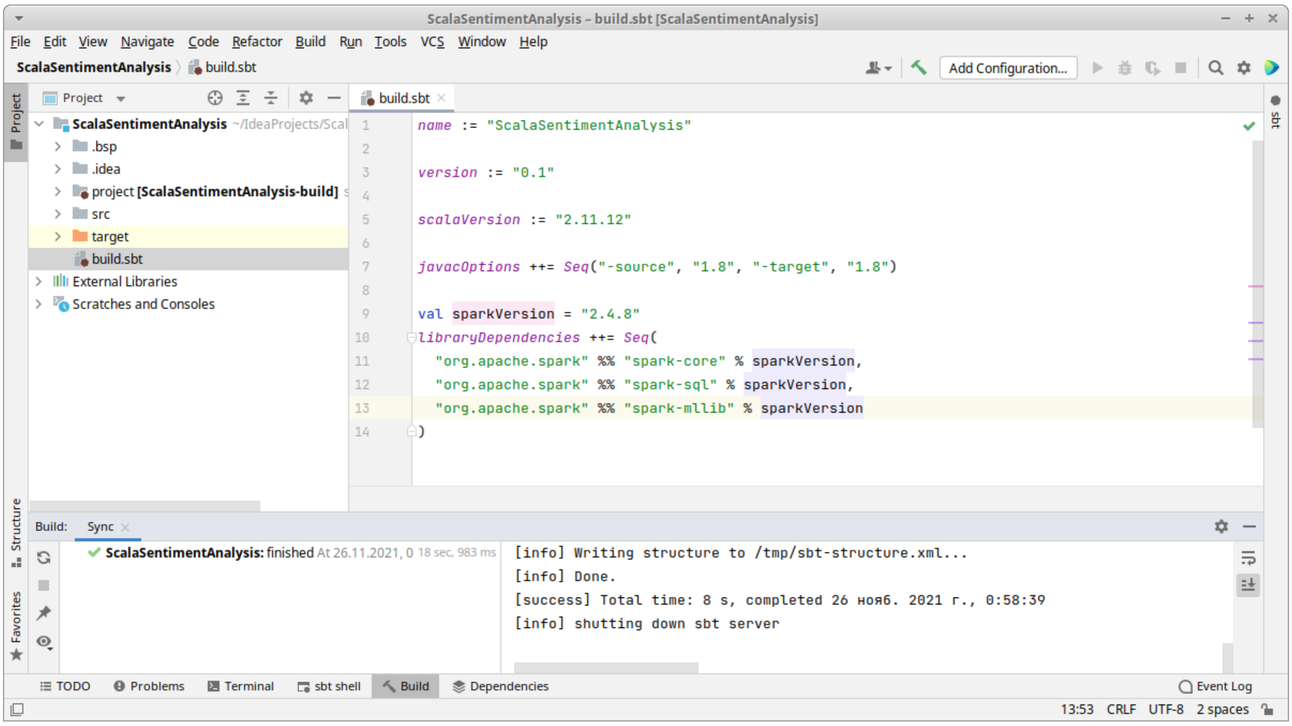
JDK: 11 version 11.0.11

sbt: 1.5.5 □ Sources

Scala: 2.11.12 ☑ Sources

Package prefix: Such as 'org.example.application'

Добавим в build.sbt зависимость на Apache Spark:

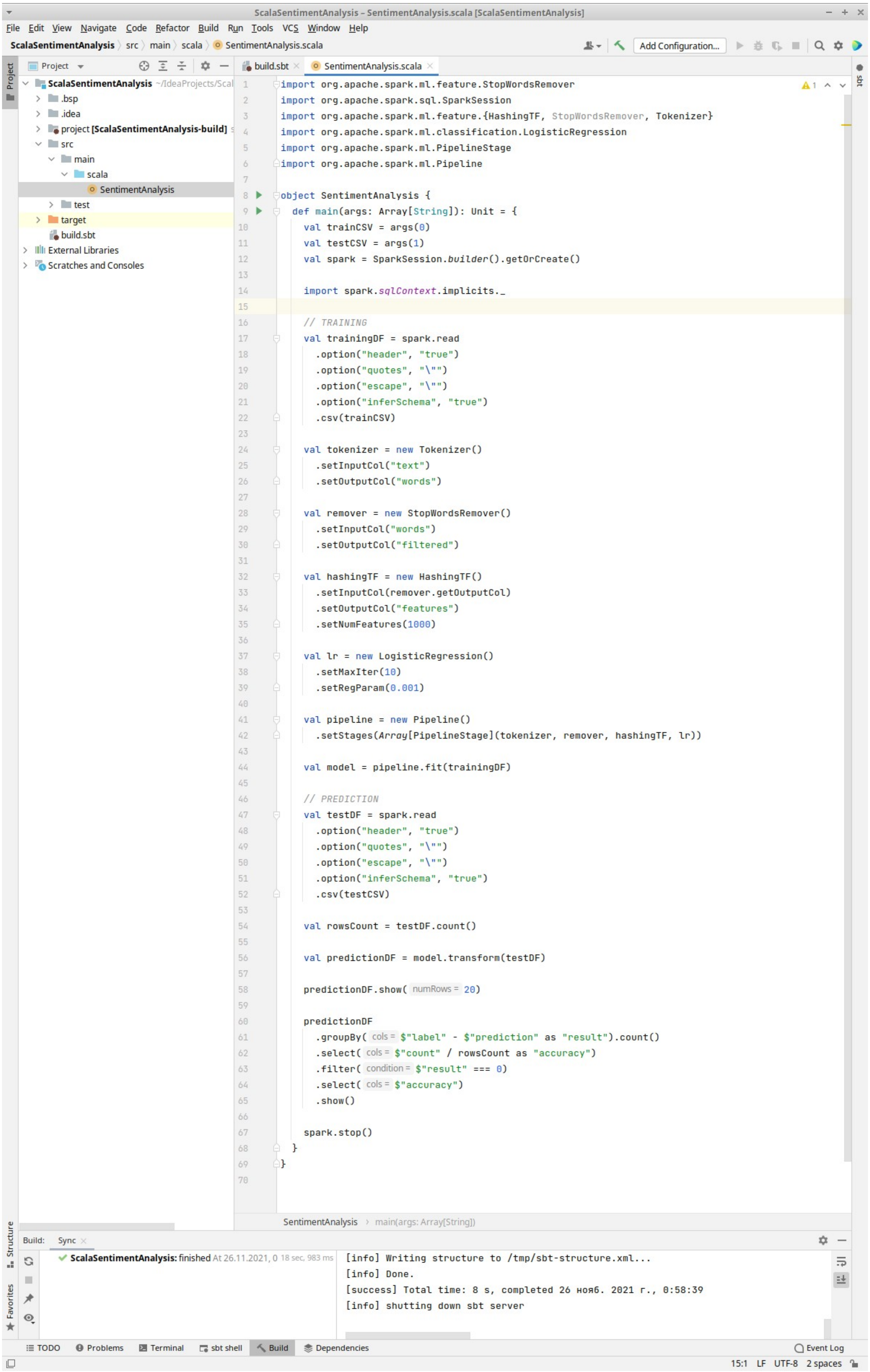


```
name := "ScalaSentimentAnalysis"

version := "0.1"

scalaVersion := "2.11.12"

javacOptions ++= Seq("-source", "1.8", "-target", "1.8")

val sparkVersion = "2.4.8"
libraryDependencies ++= Seq(
  "org.apache.spark" %% "spark-core" % sparkVersion,
  "org.apache.spark" %% "spark-sql" % sparkVersion,
  "org.apache.spark" %% "spark-mllib" % sparkVersion
)
```

ScalaSentimentAnalysis: finished At 26.11.2021, 0 18 sec, 983 ms

```
[info] Writing structure to /tmp/sbt-structure.xml...
[info] Done.
[success] Total time: 8 s, completed 26 нояб. 2021 г., 0:58:39
[info] shutting down sbt server
```

Создадим ScalaSentimentAnalysis.scala:


ScalaSentimentAnalysis – SentimentAnalysis.scala [ScalaSentimentAnalysis]
File Edit View Navigate Code Refactor Build Run Tools VCS Window Help
ScalaSentimentAnalysis 〉 src 〉 main 〉 scala 〉 SentimentAnalysis.scala
Add Configuration...
Project
ScalaSentimentAnalysis ~/IdeaProjects/Scal
.bsp
.idea
project [ScalaSentimentAnalysis-build]
src
main
scala
SentimentAnalysis
test
target
build.sbt
External Libraries
Scratches and Consoles
build.sbt
SentimentAnalysis.scala
SentimentAnalysis 〉 main(args: Array[String])
Build: Sync
TODO Problems Terminal sbt shell Build Dependencies
Event Log
15:1 LF UTF-8 2 spaces

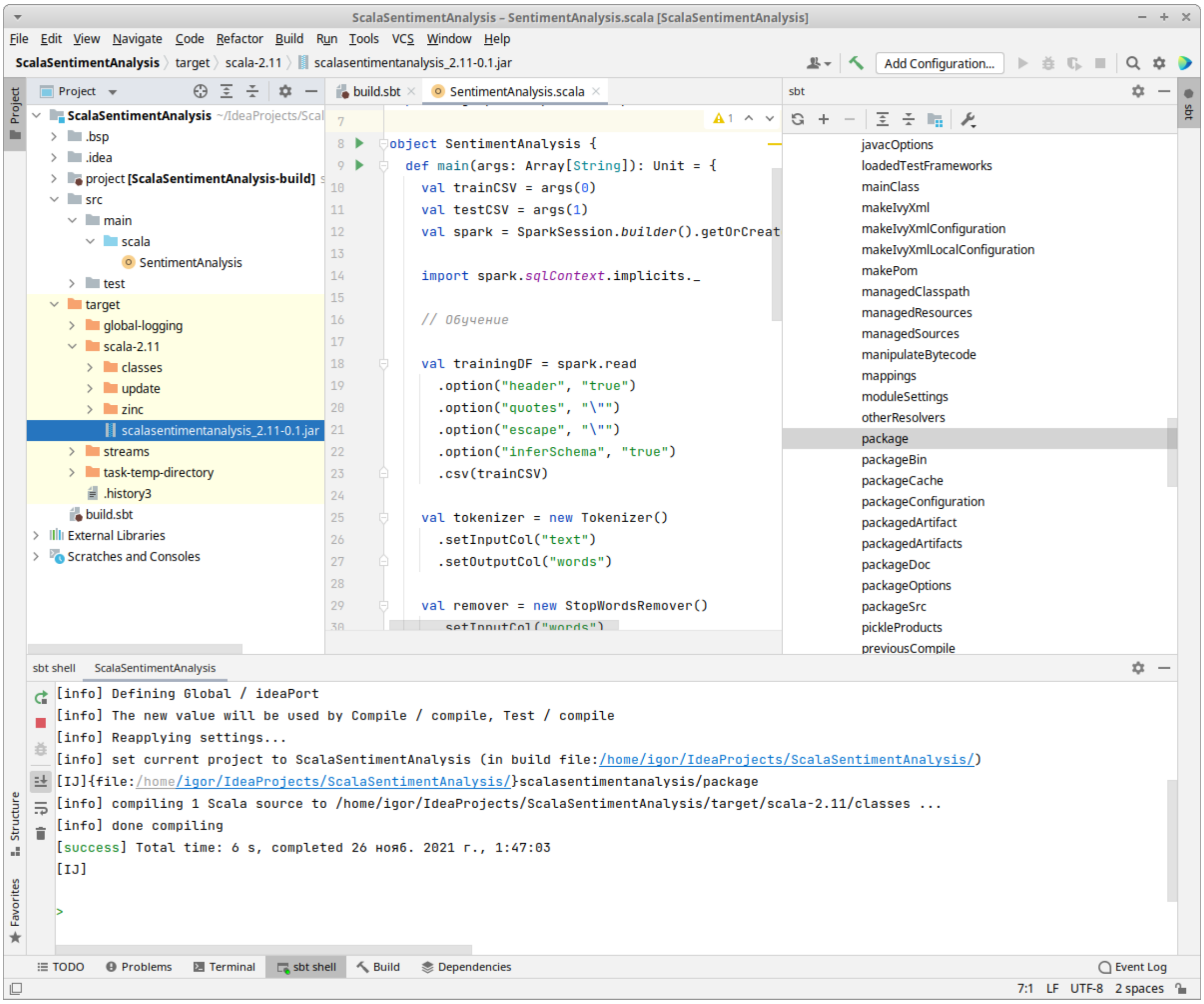
```scala
import org.apache.spark.ml.feature.StopWordsRemover
import org.apache.spark.sql.SparkSession
import org.apache.spark.ml.feature.{HashingTF, StopWordsRemover, Tokenizer}
import org.apache.spark.ml.classification.LogisticRegression
import org.apache.spark.ml.PipelineStage
import org.apache.spark.ml.Pipeline

object SentimentAnalysis {
  def main(args: Array[String]): Unit = {
    val trainCSV = args(0)
    val testCSV = args(1)
    val spark = SparkSession.builder().getOrCreate()

    import spark.sqlContext.implicits._

    // TRAINING
    val trainingDF = spark.read
      .option("header", "true")
      .option("quotes", "\"")
      .option("escape", "\"")
      .option("inferSchema", "true")
      .csv(trainCSV)

    val tokenizer = new Tokenizer()
      .setInputCol("text")
      .setOutputCol("words")

    val remover = new StopWordsRemover()
      .setInputCol("words")
      .setOutputCol("filtered")

    val hashingTF = new HashingTF()
      .setInputCol(remover.getOutputCol)
      .setOutputCol("features")
      .setNumFeatures(1000)

    val lr = new LogisticRegression()
      .setMaxIter(10)
      .setRegParam(0.001)

    val pipeline = new Pipeline()
      .setStages(Array[PipelineStage](tokenizer, remover, hashingTF, lr))

    val model = pipeline.fit(trainingDF)

    // PREDICTION
    val testDF = spark.read
      .option("header", "true")
      .option("quotes", "\"")
      .option("escape", "\"")
      .option("inferSchema", "true")
      .csv(testCSV)

    val rowsCount = testDF.count()

    val predictionDF = model.transform(testDF)

    predictionDF.show( numRows = 20)

    predictionDF
      .groupBy( cols = $"label" - $"prediction" as "result").count()
      .select( cols = $"count" / rowsCount as "accuracy")
      .filter( condition = $"result" === 0)
      .select( cols = $"accuracy")
      .show()

    spark.stop()
  }
}
```

ScalaSentimentAnalysis: finished At 26.11.2021, 0 18 sec, 983 ms

```
[info] Writing structure to /tmp/sbt-structure.xml...
[info] Done.
[success] Total time: 8 s, completed 26 нояб. 2021 г., 0:58:39
[info] shutting down sbt server
```

SCALA `SentimentAnalysis`: СБОРКА И ДОСТАВКА: sbt package

Соберем jar: sbt package

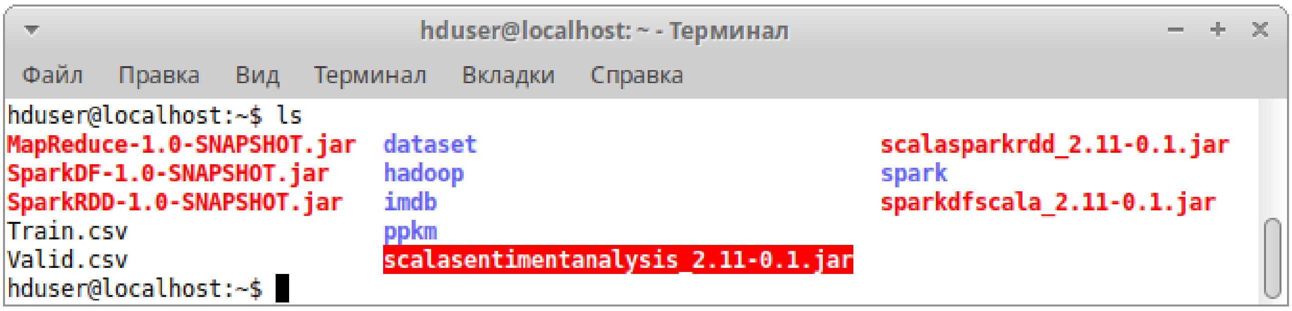


```scala

object SentimentAnalysis {
  def main(args: Array[String]): Unit = {
    val trainCSV = args(0)
    val testCSV = args(1)
    val spark = SparkSession.builder().getOrCreat

    import spark.sqlContext.implicits._

    // Обучение

    val trainingDF = spark.read
      .option("header", "true")
      .option("quotes", "\"")
      .option("escape", "\"")
      .option("inferSchema", "true")
      .csv(trainCSV)

    val tokenizer = new Tokenizer()
      .setInputCol("text")
      .setOutputCol("words")

    val remover = new StopWordsRemover()
      .setInputCol("words")
```

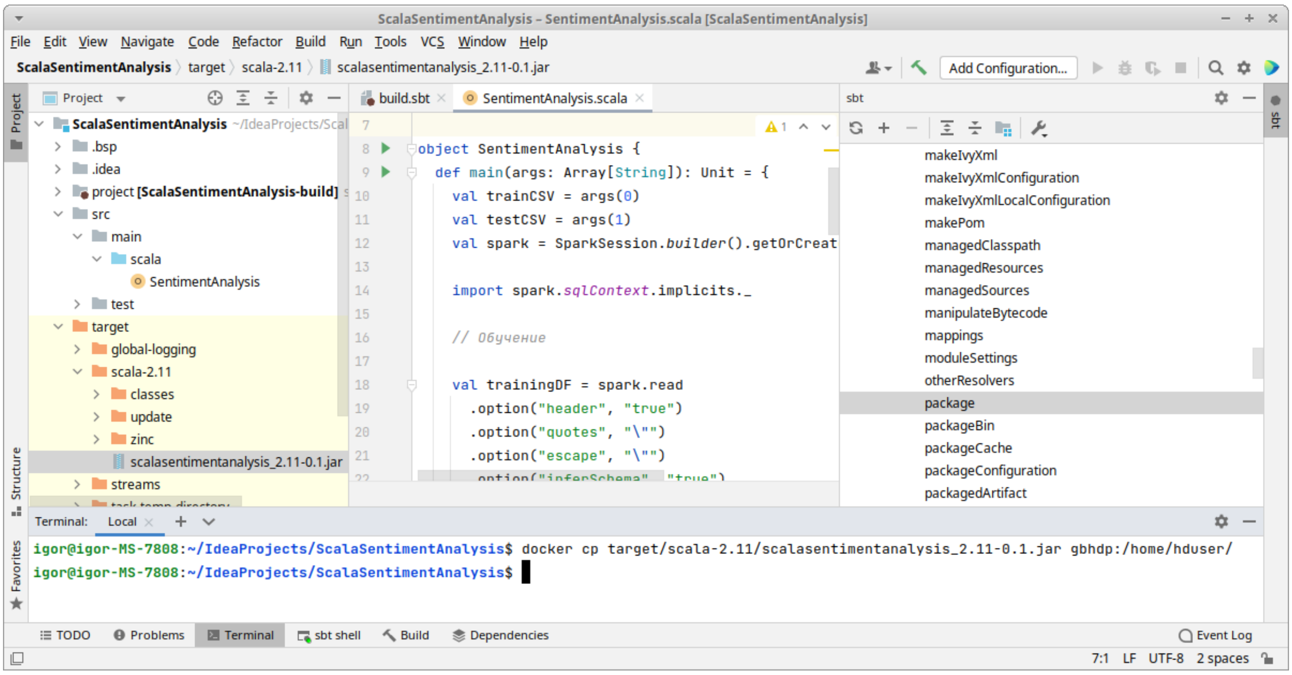
```
[info] Defining Global / ideaPort
[info] The new value will be used by Compile / compile, Test / compile
[info] Reapplying settings...
[info] set current project to ScalaSentimentAnalysis (in build file:/home/igor/IdeaProjects/ScalaSentimentAnalysis/)
[IJ]{file:/home/igor/IdeaProjects/ScalaSentimentAnalysis/}scalasentimentanalysis/package
[info] compiling 1 Scala source to /home/igor/IdeaProjects/ScalaSentimentAnalysis/target/scala-2.11/classes ...
[info] done compiling
[success] Total time: 6 s, completed 26 нояб. 2021 г., 1:47:03
[IJ]

>
```

Перенесем его на кластер:

docker cp target/scala-2.11/scalasentimentanalysis_2.11-0.1.jar gbhdp:/home/hduser/


hduser@localhost: ~ - Терминал
Файл Правка Вид Терминал Вкладки Справка

```
hduser@localhost:~$ ls
MapReduce-1.0-SNAPSHOT.jar  dataset                              scalasparkrdd_2.11-0.1.jar
SparkDF-1.0-SNAPSHOT.jar    hadoop                               spark
SparkRDD-1.0-SNAPSHOT.jar   imdb                                 sparkdfscala_2.11-0.1.jar
Train.csv                   ppkm
Valid.csv                   scalasentimentanalysis_2.11-0.1.jar
hduser@localhost:~$
```

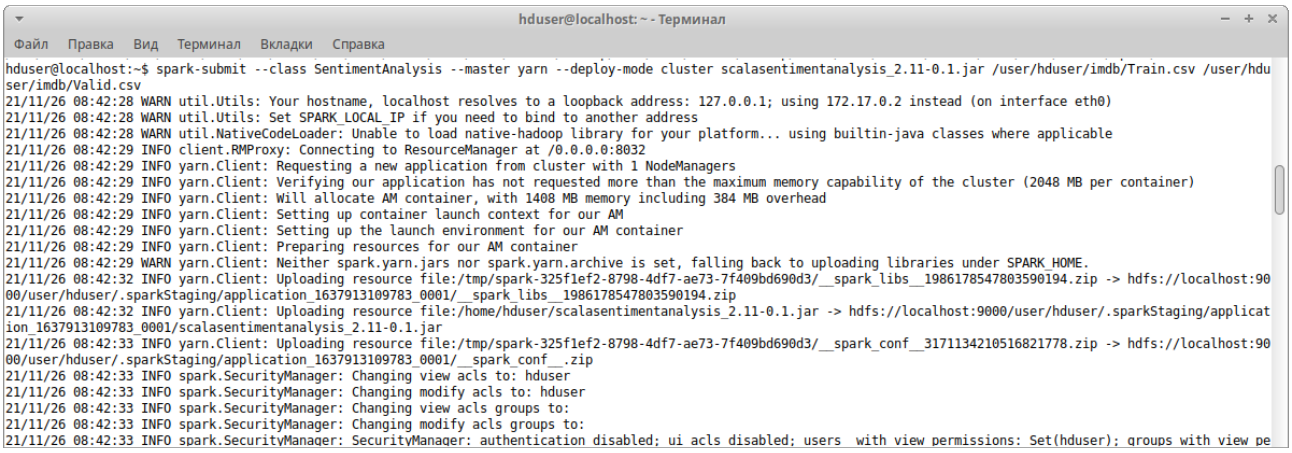


```
object SentimentAnalysis {
  def main(args: Array[String]): Unit = {
    val trainCSV = args(0)
    val testCSV = args(1)
    val spark = SparkSession.builder().getOrCreat

    import spark.sqlContext.implicits._

    // Обучение

    val trainingDF = spark.read
      .option("header", "true")
      .option("quotes", "\"")
      .option("escape", "\"")
      .option("inferSchema", "true")
```

```
igor@igor-MS-7808:~/IdeaProjects/ScalaSentimentAnalysis$ docker cp target/scala-2.11/scalasentimentanalysis_2.11-0.1.jar gbhdp:/home/hduser/
igor@igor-MS-7808:~/IdeaProjects/ScalaSentimentAnalysis$
```

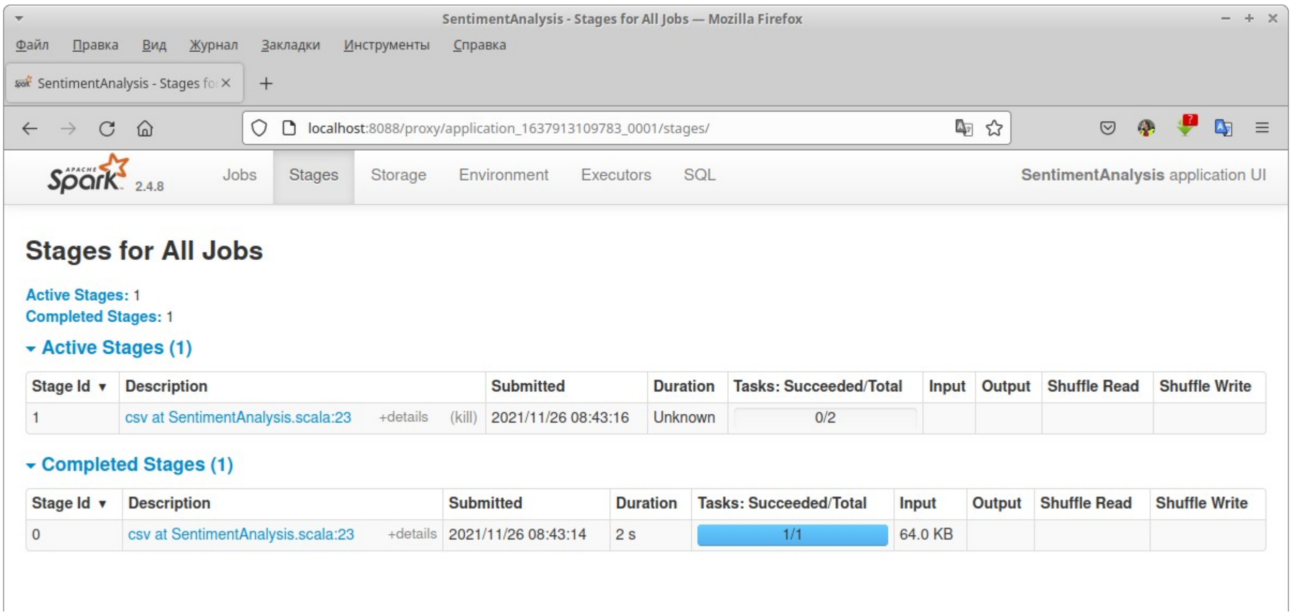
Запускаем приложение:

spark-submit --class SentimentAnalysis --master yarn --deploy-mode cluster scalasentimentanalysis_2.11-0.1.jar /user/hduser/imdb/Train.csv /user/hduser/imdb/Valid.csv


hduser@localhost: ~ - Терминал
Файл Правка Вид Терминал Вкладки Справка

```
hduser@localhost:~$ spark-submit --class SentimentAnalysis --master yarn --deploy-mode cluster scalasentimentanalysis_2.11-0.1.jar /user/hduser/imdb/Train.csv /user/hdu
ser/imdb/Valid.csv
21/11/26 08:42:28 WARN util.Utils: Your hostname, localhost resolves to a loopback address: 127.0.0.1; using 172.17.0.2 instead (on interface eth0)
21/11/26 08:42:28 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
21/11/26 08:42:28 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
21/11/26 08:42:29 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
21/11/26 08:42:29 INFO yarn.Client: Requesting a new application from cluster with 1 NodeManagers
21/11/26 08:42:29 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster (2048 MB per container)
21/11/26 08:42:29 INFO yarn.Client: Will allocate AM container, with 1408 MB memory including 384 MB overhead
21/11/26 08:42:29 INFO yarn.Client: Setting up container launch context for our AM
21/11/26 08:42:29 INFO yarn.Client: Setting up the launch environment for our AM container
21/11/26 08:42:29 INFO yarn.Client: Preparing resources for our AM container
21/11/26 08:42:29 WARN yarn.Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
21/11/26 08:42:32 INFO yarn.Client: Uploading resource file:/tmp/spark-325f1ef2-8798-4df7-ae73-7f409bd690d3/__spark_libs__1986178547803590194.zip -> hdfs://localhost:90
00/user/hduser/.sparkStaging/application_1637913109783_0001/__spark_libs__1986178547803590194.zip
21/11/26 08:42:32 INFO yarn.Client: Uploading resource file:/home/hduser/scalasentimentanalysis_2.11-0.1.jar -> hdfs://localhost:9000/user/hduser/.sparkStaging/applicat
ion_1637913109783_0001/scalasentimentanalysis_2.11-0.1.jar
21/11/26 08:42:33 INFO yarn.Client: Uploading resource file:/tmp/spark-325f1ef2-8798-4df7-ae73-7f409bd690d3/__spark_conf__3171134210516821778.zip -> hdfs://localhost:90
00/user/hduser/.sparkStaging/application_1637913109783_0001/__spark_conf__.zip
21/11/26 08:42:33 INFO spark.SecurityManager: Changing view acls to: hduser
21/11/26 08:42:33 INFO spark.SecurityManager: Changing modify acls to: hduser
21/11/26 08:42:33 INFO spark.SecurityManager: Changing view acls groups to:
21/11/26 08:42:33 INFO spark.SecurityManager: Changing modify acls groups to:
21/11/26 08:42:33 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(hduser); groups with view pe
```

tracking URL: http://localhost:8088/proxy/application_1637913109783_0001/


SentimentAnalysis - Stages for All Jobs — Mozilla Firefox
Файл Правка Вид Журнал Закладки Инструменты Справка
SentimentAnalysis - Stages fo
localhost:8088/proxy/application_1637913109783_0001/stages/
Spark 2.4.8
Jobs
Stages
Storage
Environment
Executors
SQL
SentimentAnalysis application UI

Stages for All Jobs

Active Stages: 1
Completed Stages: 1

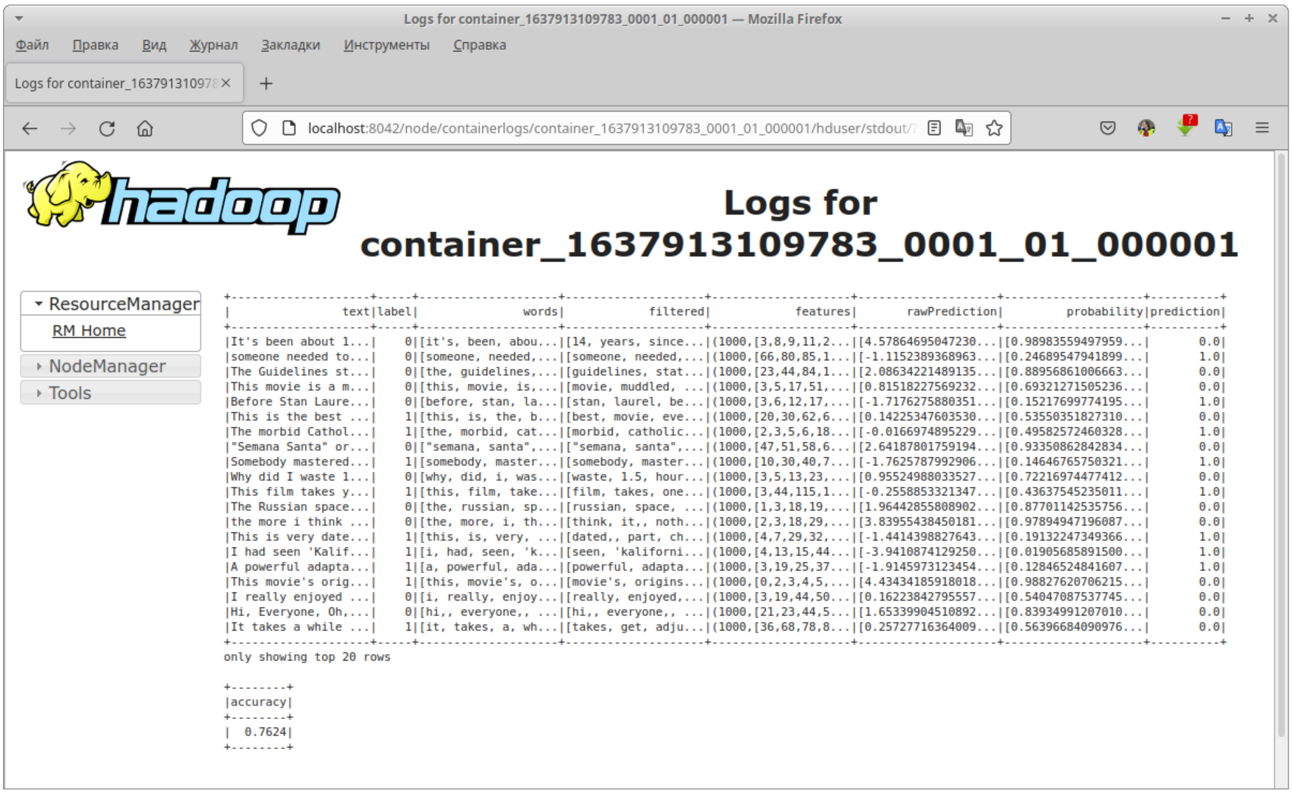
Active Stages (1)

Stage Id ▾	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
1	csv at SentimentAnalysis.scala:23 +details (kill)	2021/11/26 08:43:16	Unknown	0/2				

Completed Stages (1)

Stage Id ▾	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
0	csv at SentimentAnalysis.scala:23 +details	2021/11/26 08:43:14	2 s	1/1	64.0 KB			

Logs for container_1637913109783_0001_01_000001


Logs for container_1637913109783_0001_01_000001 — Mozilla Firefox
Файл Правка Вид Журнал Закладки Инструменты Справка
Logs for container_16379131097
localhost:8042/node/containerlogs/container_1637913109783_0001_01_000001/hduser/stdout/
hadoop
ResourceManager
RM Home
NodeManager
Tools

Logs for container_1637913109783_0001_01_000001

```
+--------------------+-----+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
|                text|label|               words|            filtered|            features|       rawPrediction|         probability|prediction|
+--------------------+-----+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
|It's been about 1...|    0|[it's, been, abou...|[14, years, since...|(1000,[3,8,9,11,2...|[4.57864695047230...|[0.98983559497959...|       0.0|
|someone needed to...|    0|[someone, needed,...|[someone, needed,...|(1000,[66,80,85,1...|[-1.1152389368963...|[0.24689547941899...|       1.0|
|The Guidelines st...|    0|[the, guidelines,...|[guidelines, stat...|(1000,[23,44,84,1...|[2.08634221489135...|[0.88956861006663...|       0.0|
|This movie is a m...|    0|[this, movie, is,...|[movie, muddled, ...|(1000,[3,5,17,51,...|[0.81518227569232...|[0.69321271505236...|       0.0|
|Before Stan Laure...|    0|[before, stan, la...|[stan, laurel, be...|(1000,[3,6,12,17,...|[-1.7176275880351...|[0.15217699774195...|       1.0|
|This is the best ...|    1|[this, is, the, b...|[best, movie, eve...|(1000,[20,30,62,6...|[0.14225347603530...|[0.53550351827310...|       0.0|
|The morbid Cathol...|    1|[the, morbid, cat...|[morbid, catholic...|(1000,[2,3,5,6,18...|[-0.0166974895229...|[0.49582572460328...|       1.0|
|"Semana Santa" or...|    0|["semana, santa",...|["semana, santa",...|(1000,[47,51,58,6...|[2.64187801759194...|[0.93350862842834...|       0.0|
|Somebody mastered...|    1|[somebody, master...|[somebody, master...|(1000,[10,30,40,7...|[-1.7625787992906...|[0.14646765750321...|       1.0|
|Why did I waste 1...|    0|[why, did, i, was...|[waste, 1.5, hour...|(1000,[3,5,13,23,...|[0.95524988033527...|[0.72216974477412...|       0.0|
|This film takes y...|    1|[this, film, take...|[film, takes, one...|(1000,[3,44,115,1...|[-0.2558853321347...|[0.43637545235011...|       1.0|
|The Russian space...|    0|[the, russian, sp...|[russian, space, ...|(1000,[1,3,18,19,...|[1.96442855808902...|[0.87701142535756...|       0.0|
|the more i think ...|    0|[the, more, i, th...|[think, it,, noth...|(1000,[2,3,18,29,...|[3.83955438450181...|[0.97894947196087...|       0.0|
|This is very date...|    1|[this, is, very, ...|[dated,, part, ch...|(1000,[4,7,29,32,...|[-1.4414398827643...|[0.19132247349366...|       1.0|
|I had seen 'Kalif...|    1|[i, had, seen, 'k...|[seen, 'kaliforni...|(1000,[4,13,15,44...|[-3.9410874129250...|[0.01905685891500...|       1.0|
|A powerful adapta...|    1|[a, powerful, ada...|[powerful, adapta...|(1000,[3,19,25,37...|[-1.9145973123454...|[0.12846524841607...|       1.0|
|This movie's orig...|    1|[this, movie's, o...|[movie's, origins...|(1000,[0,2,3,4,5,...|[4.43434185918018...|[0.98827620706215...|       0.0|
|I really enjoyed ...|    0|[i, really, enjoy...|[really, enjoyed,...|(1000,[3,19,44,50...|[0.16223842795557...|[0.54047087537745...|       0.0|
|Hi, Everyone, Oh,...|    0|[hi,, everyone,, ...|[hi,, everyone,, ...|(1000,[21,23,44,5...|[1.65339904510892...|[0.83934991207010...|       0.0|
|It takes a while ...|    1|[it, takes, a, wh...|[takes, get, adju...|(1000,[36,68,78,8...|[0.25727716364009...|[0.56396684090976...|       0.0|
+--------------------+-----+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
only showing top 20 rows

+--------+
|accuracy|
+--------+
|  0.7624|
+--------+
```

accuracy|

```
+--------+
|  0.7624|
```