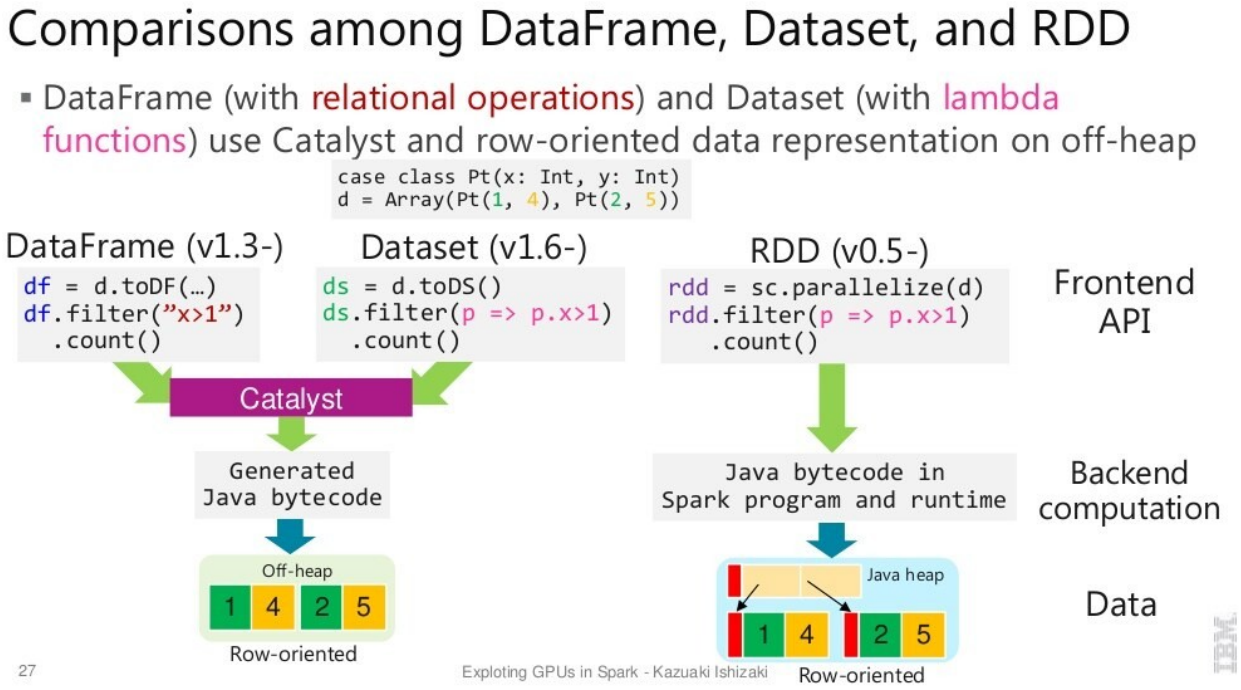
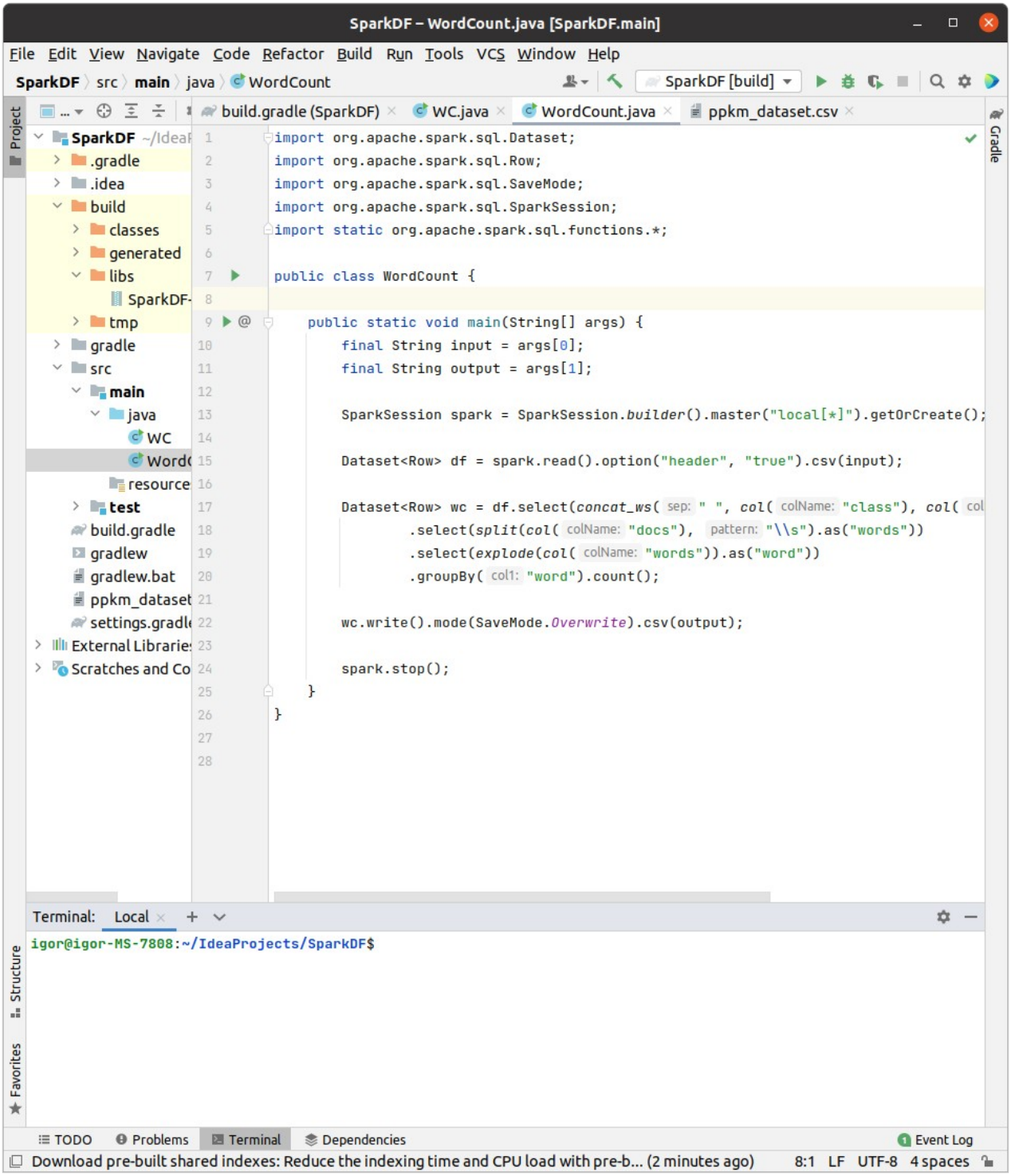
Домашнее задание к 3 уроку

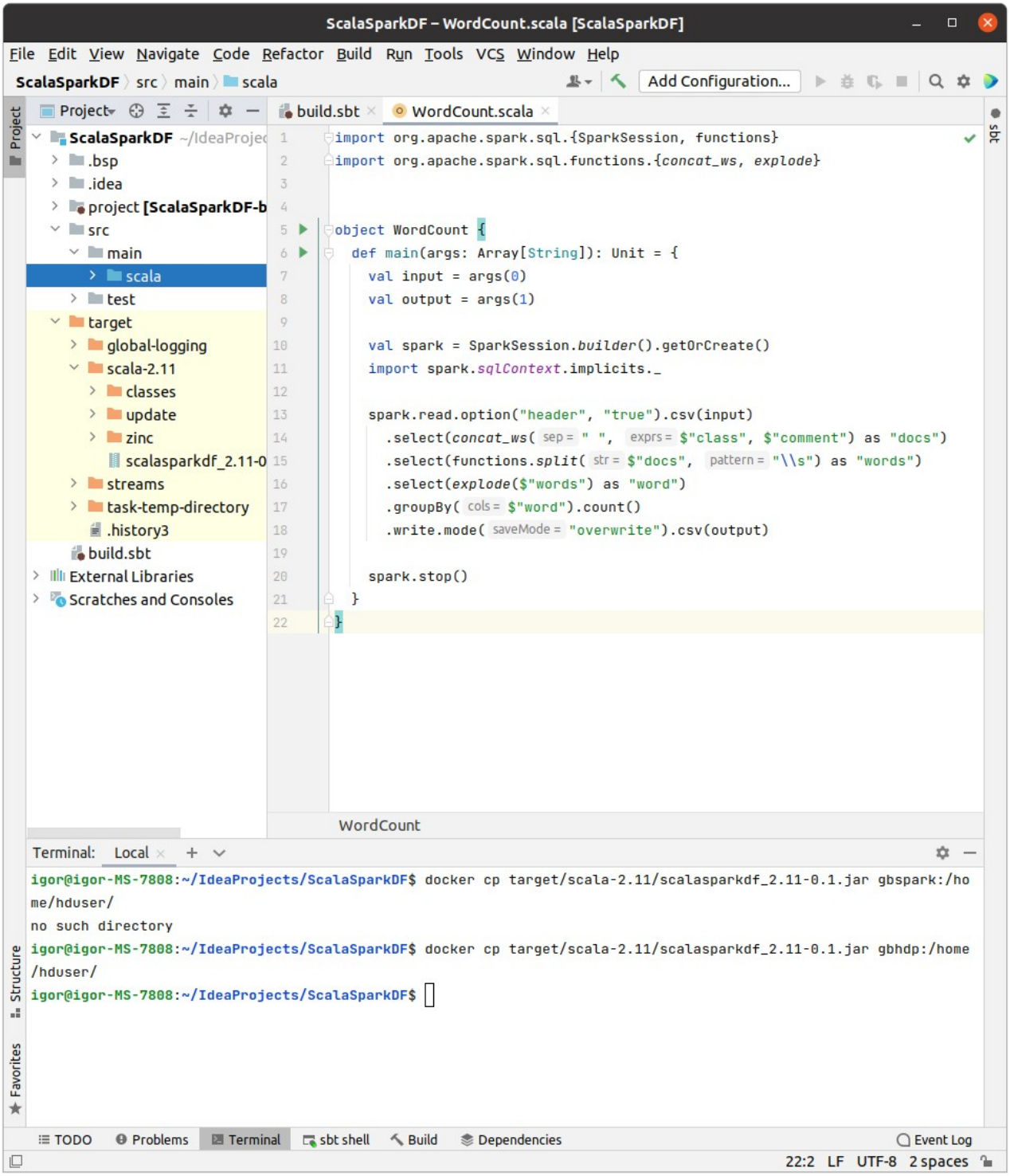
1. **Какая связь между DataFrame и Dataset?**

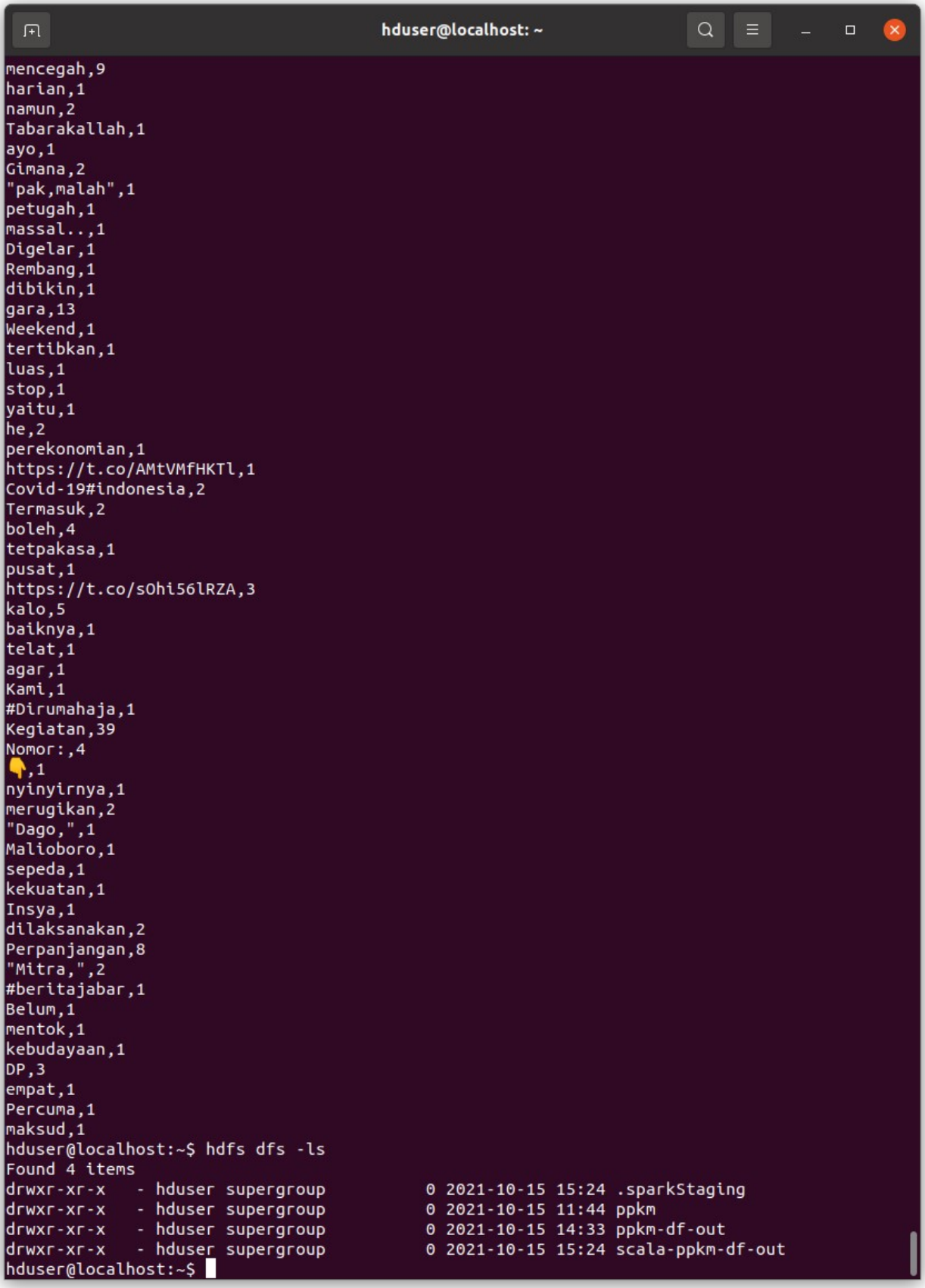
В Catalist генерируется байт код



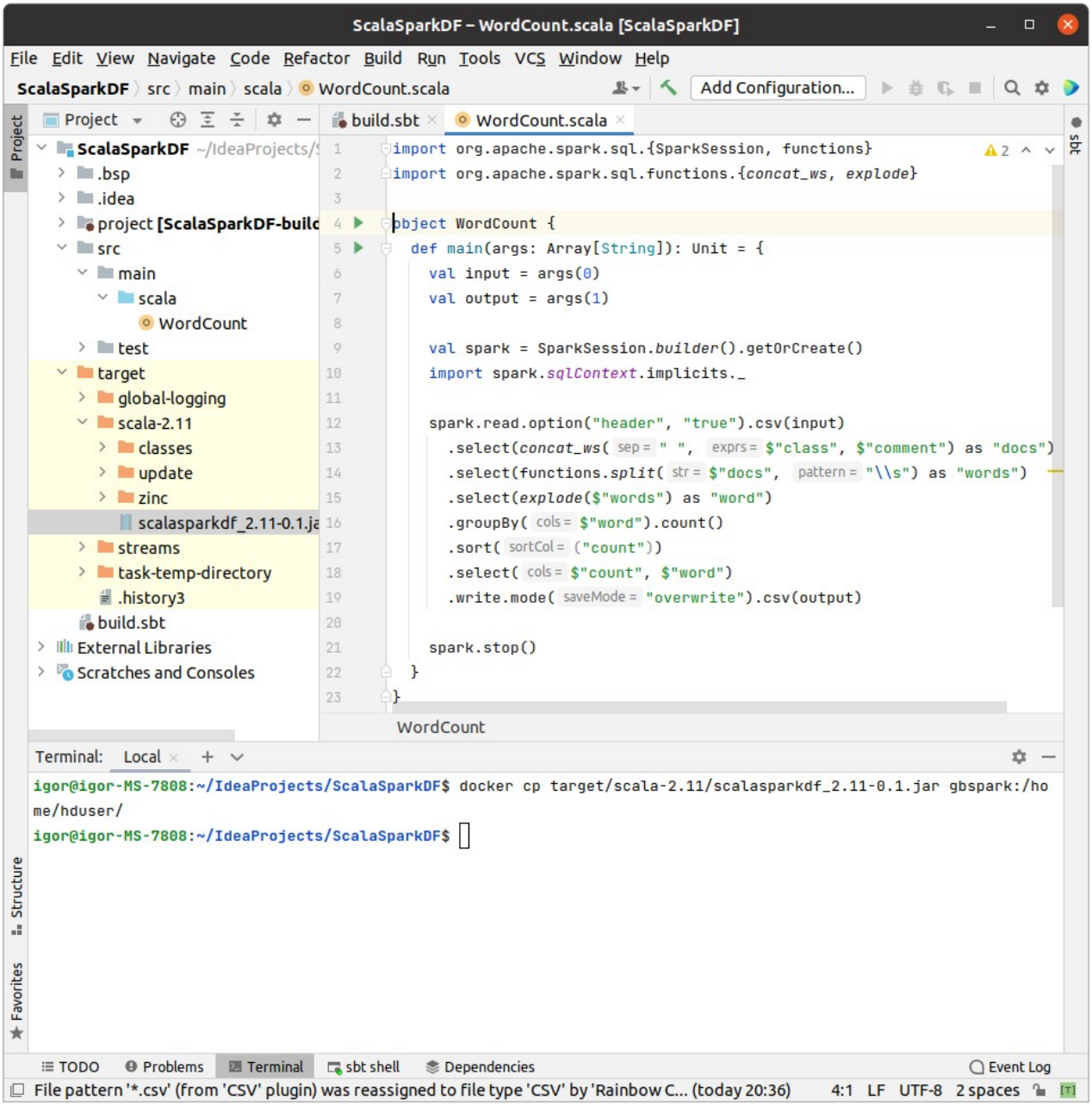
2. **Соберите WordCount приложение, запустите на датасете ppkm_sentiment**

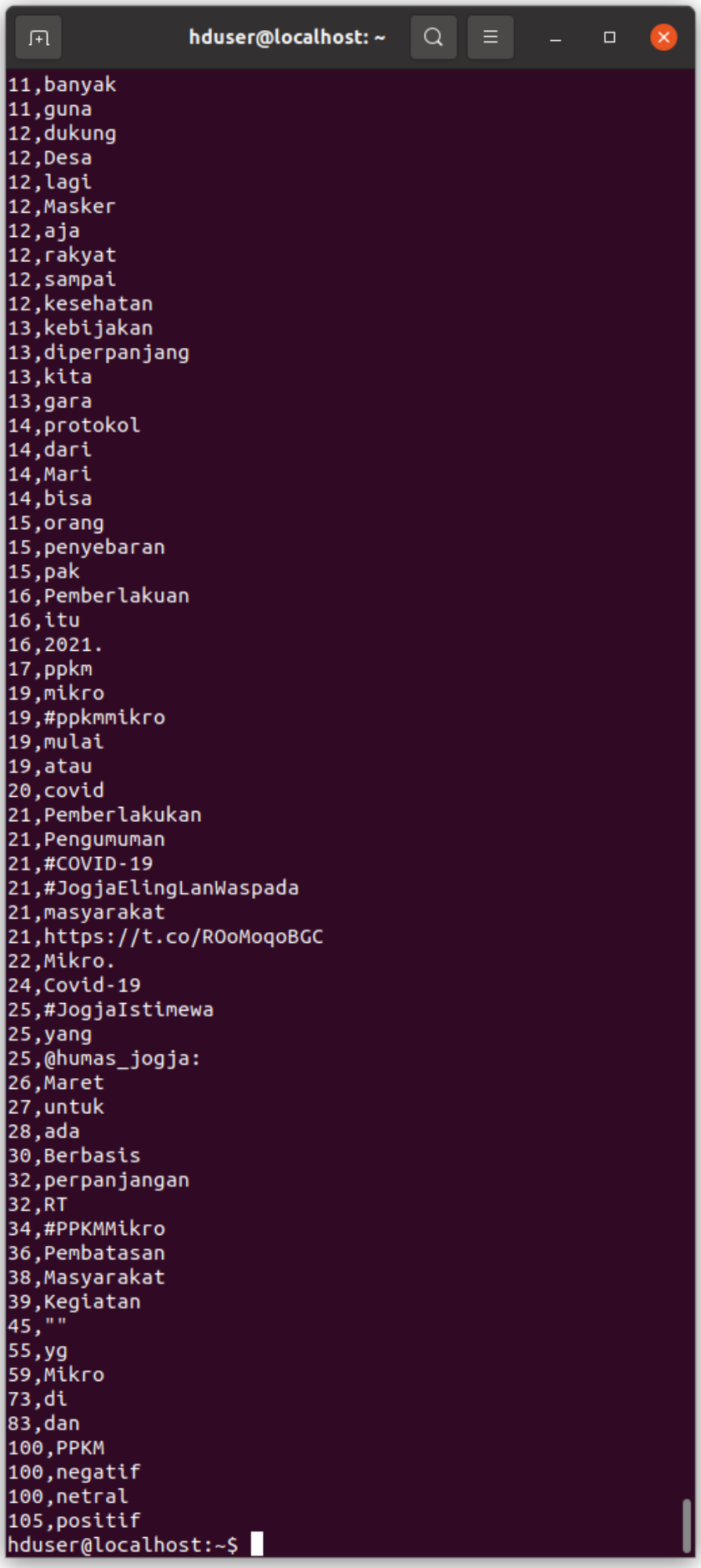

SparkDF – WordCount.java [SparkDF.main]
File Edit View Navigate Code Refactor Build Run Tools VCS Window Help
SparkDF 〉 src 〉 main 〉 java 〉 WordCount
SparkDF [build]
build.gradle (SparkDF)
WC.java
WordCount.java
ppkm_dataset.csv
Project
Gradle
SparkDF ~/IdeaP
.gradle
.idea
build
classes
generated
libs
SparkDF-
tmp
gradle
src
main
java
WC
WordC
resource
test
build.gradle
gradlew
gradlew.bat
ppkm_dataset
settings.gradl
External Librarie
Scratches and Co
import org.apache.spark.sql.Dataset;
import org.apache.spark.sql.Row;
import org.apache.spark.sql.SaveMode;
import org.apache.spark.sql.SparkSession;
import static org.apache.spark.sql.functions.*;
public class WordCount {
public static void main(String[] args) {
final String input = args[0];
final String output = args[1];
SparkSession spark = SparkSession.builder().master("local[*]").getOrCreate();
Dataset<Row> df = spark.read().option("header", "true").csv(input);
Dataset<Row> wc = df.select(concat_ws(sep: " ", col(colName: "class"), col(col
.select(split(col(colName: "docs"), pattern: "\\s").as("words"))
.select(explode(col(colName: "words")).as("word"))
.groupBy(col1: "word").count();
wc.write().mode(SaveMode.Overwrite).csv(output);
spark.stop();
}
}
Terminal: Local
igor@igor-MS-7808:~/IdeaProjects/SparkDF$
Structure
Favorites
TODO
Problems
Terminal
Dependencies
Event Log
Download pre-built shared indexes: Reduce the indexing time and CPU load with pre-b... (2 minutes ago)
8:1 LF UTF-8 4 spaces


ScalaSparkDF – WordCount.scala [ScalaSparkDF]
File Edit View Navigate Code Refactor Build Run Tools VCS Window Help
ScalaSparkDF 〉 src 〉 main 〉 scala
Add Configuration...
Project
build.sbt
WordCount.scala
ScalaSparkDF ~/IdeaProjec
.bsp
.idea
project [ScalaSparkDF-b
src
main
scala
test
target
global-logging
scala-2.11
classes
update
zinc
scalasparkdf_2.11-0
streams
task-temp-directory
.history3
build.sbt
External Libraries
Scratches and Consoles
import org.apache.spark.sql.{SparkSession, functions}
import org.apache.spark.sql.functions.{concat_ws, explode}
object WordCount {
def main(args: Array[String]): Unit = {
val input = args(0)
val output = args(1)
val spark = SparkSession.builder().getOrCreate()
import spark.sqlContext.implicits._
spark.read.option("header", "true").csv(input)
.select(concat_ws(sep = " ", exprs = $"class", $"comment") as "docs")
.select(functions.split(str = $"docs", pattern = "\\s") as "words")
.select(explode($"words") as "word")
.groupBy(cols = $"word").count()
.write.mode(saveMode = "overwrite").csv(output)
spark.stop()
}
}
WordCount
Terminal: Local
igor@igor-MS-7808:~/IdeaProjects/ScalaSparkDF$ docker cp target/scala-2.11/scalasparkdf_2.11-0.1.jar gbspark:/home/hduser/
no such directory
igor@igor-MS-7808:~/IdeaProjects/ScalaSparkDF$ docker cp target/scala-2.11/scalasparkdf_2.11-0.1.jar gbhdp:/home/hduser/
igor@igor-MS-7808:~/IdeaProjects/ScalaSparkDF$
Structure
Favorites
TODO
Problems
Terminal
sbt shell
Build
Dependencies
Event Log
22:2 LF UTF-8 2 spaces


hduser@localhost: ~
mencegah,9
harian,1
namun,2
Tabarakallah,1
ayo,1
Gimana,2
"pak,malah",1
petugah,1
massal..,1
Digelar,1
Rembang,1
dibikin,1
gara,13
Weekend,1
tertibkan,1
luas,1
stop,1
yaitu,1
he,2
perekonomian,1
https://t.co/AMtVMfHKTl,1
Covid-19#indonesia,2
Termasuk,2
boleh,4
tetpakasa,1
pusat,1
https://t.co/sOhi56lRZA,3
kalo,5
baiknya,1
telat,1
agar,1
Kami,1
#Dirumahaja,1
Kegiatan,39
Nomor:,4
👇,1
nyinyirnya,1
merugikan,2
"Dago,",1
Malioboro,1
sepeda,1
kekuatan,1
Insya,1
dilaksanakan,2
Perpanjangan,8
"Mitra,",2
#beritajabar,1
Belum,1
mentok,1
kebudayaan,1
DP,3
empat,1
Percuma,1
maksud,1
hduser@localhost:~$ hdfs dfs -ls
Found 4 items
drwxr-xr-x - hduser supergroup 0 2021-10-15 15:24 .sparkStaging
drwxr-xr-x - hduser supergroup 0 2021-10-15 11:44 ppkm
drwxr-xr-x - hduser supergroup 0 2021-10-15 14:33 ppkm-df-out
drwxr-xr-x - hduser supergroup 0 2021-10-15 15:24 scala-ppkm-df-out
hduser@localhost:~$

3. *** Измените WordCount так, чтобы он удалял знаки препинания и приводил все слова к единому регистру**
4. *** Добавьте в WordCount возможность через конфигурацию задать список стоп-слов, которые будут отфильтрованы во время работы приложения**
5. **Измените выход WordCount так, чтобы сортировка была по количеству повторений, а список слов был во втором столбце, а не в первом**




hduser@localhost: ~
11,banyak
11,guna
12,dukung
12,Desa
12,lagi
12,Masker
12,aja
12,rakyat
12,sampai
12,kesehatan
13,kebijakan
13,diperpanjang
13,kita
13,gara
14,protokol
14,dari
14,Mari
14,bisa
15,orang
15,penyebaran
15,pak
16,Pemberlakuan
16,itu
16,2021.
17,ppkm
19,mikro
19,#ppkmmikro
19,mulai
19,atau
20,covid
21,Pemberlakukan
21,Pengumuman
21,#COVID-19
21,#JogjaElingLanWaspada
21,masyarakat
21,https://t.co/ROoMoqoBGC
22,Mikro.
24,Covid-19
25,#JogjaIstimewa
25,yang
25,@humas_jogja:
26,Maret
27,untuk
28,ada
30,Berbasis
32,perpanjangan
32,RT
34,#PPKMMikro
36,Pembatasan
38,Masyarakat
39,Kegiatan
45,""
55,yg
59,Mikro
73,di
83,dan
100,PPKM
100,negatif
100,netral
105,positif
hduser@localhost:~$

6. * **Почему в примере в выходном файле получилось 200 партиций?**

В конфигураторе SparkSession по умолчанию установленно - 200 партиций