CSEE5590/CS490: Big Data Programming

LAB ASSIGNMENT #2

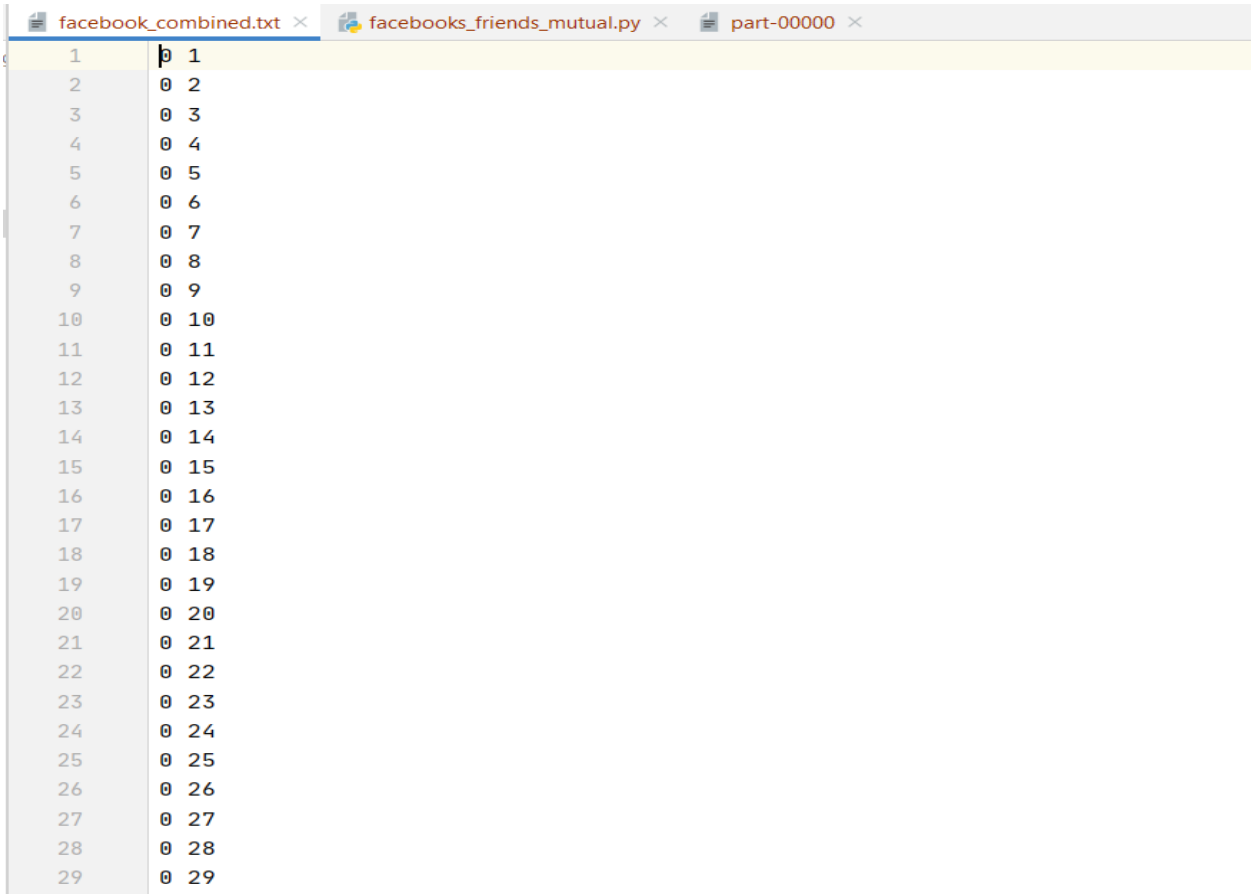
Team 4:

Aparna Manda(11)
Lohitha Yenugu(19)

Task 1

a. Implement MapReduce algorithm for finding Facebook common friends problem and run the MapReduce job on Apache Spark.

Input



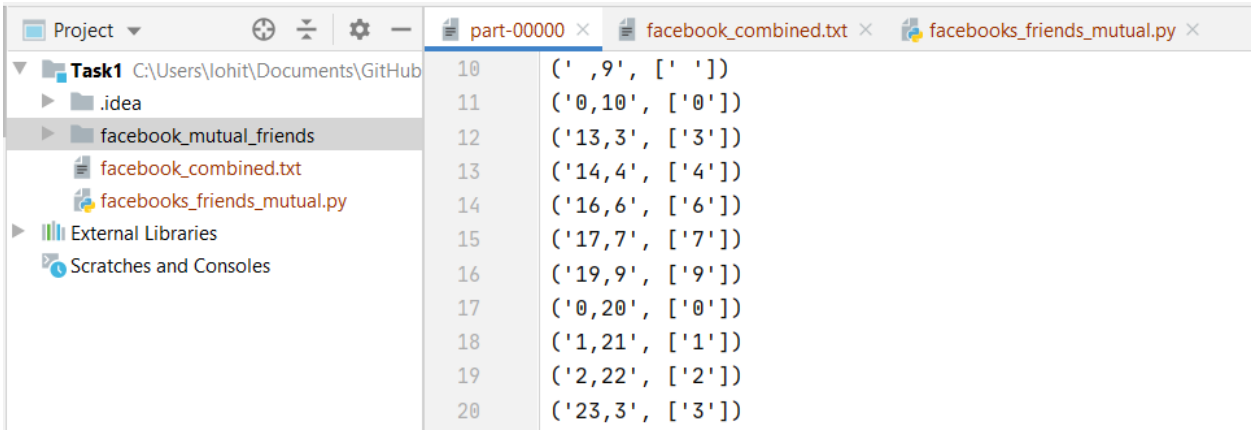
Code:

facebook_combined.txt | facebooks_friends_mutual.py | part-00000

```
from builtins import print, list, len, set
from pyspark import SparkContext


def map_friends(line):
    people = line.split(" ")
    person = people[0]
    friend = line[1]
    return person, friend


def map_togroup(tuple):
    person = tuple[0]
    final = []
    for friend in tuple[1]:
        if person < friend:
            key = person + "," + friend
        else:
            key = friend + "," + person
        value = key, list(tuple[1])
        final.append(value)
    return final

```

```
24
25      def reduce(key, value):
26          return list(set(key) & set(value))
27
28
29      def run(input, output):
30          sc = SparkContext.getOrCreate()
31          lines = sc.textFile(input, 1)
32          mapped_friends = lines.map(map_friends).groupByKey()
33          grouped_friends = mapped_friends.flatMap(map_togroup)
34          mutual_friends = grouped_friends.reduceByKey(reduce).filter(lambda x: len(x[1]) > 0)
35          print(mutual_friends.collect())
36          mutual_friends.coalesce(1).saveAsTextFile(output)
37
38
39      if __name__ == "__main__":
40          run("facebook_combined.txt", "facebook_mutual_friends")
41
```

Output:



Task 2

a. Create a Spark DataFrame

b. Perform 10 intuitive queries in Dataset

c. Using 5 queries compare between Spark RDD's and Spark Data Frames.

10 Queries

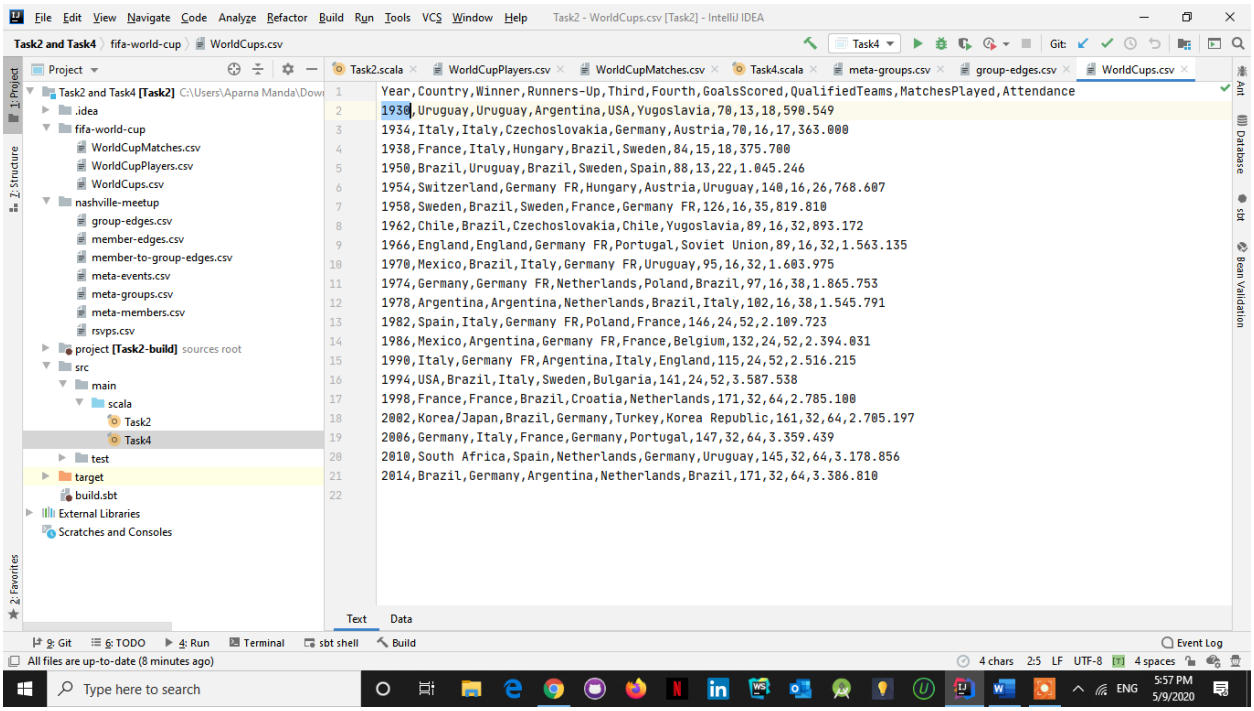
1. The country which was winner for highest number of times.
2. The year in which there was maximum craze and attendance for the games.
3. Home Team Goals and Away Team Goals for each stage
4. Referees with highest number of matches
5. Home Team Goals for Winners
6. Countries that hosted highest world cup matches
7. No of matches in year 1934 and in Estadio Centenario stadium
8. Captains of Teams of Final Matches for all Years
9. Number of matches of referee Macias Jose
10. Display all Goal-Keeper Names

5 Queries Comparing RDD's and DF's

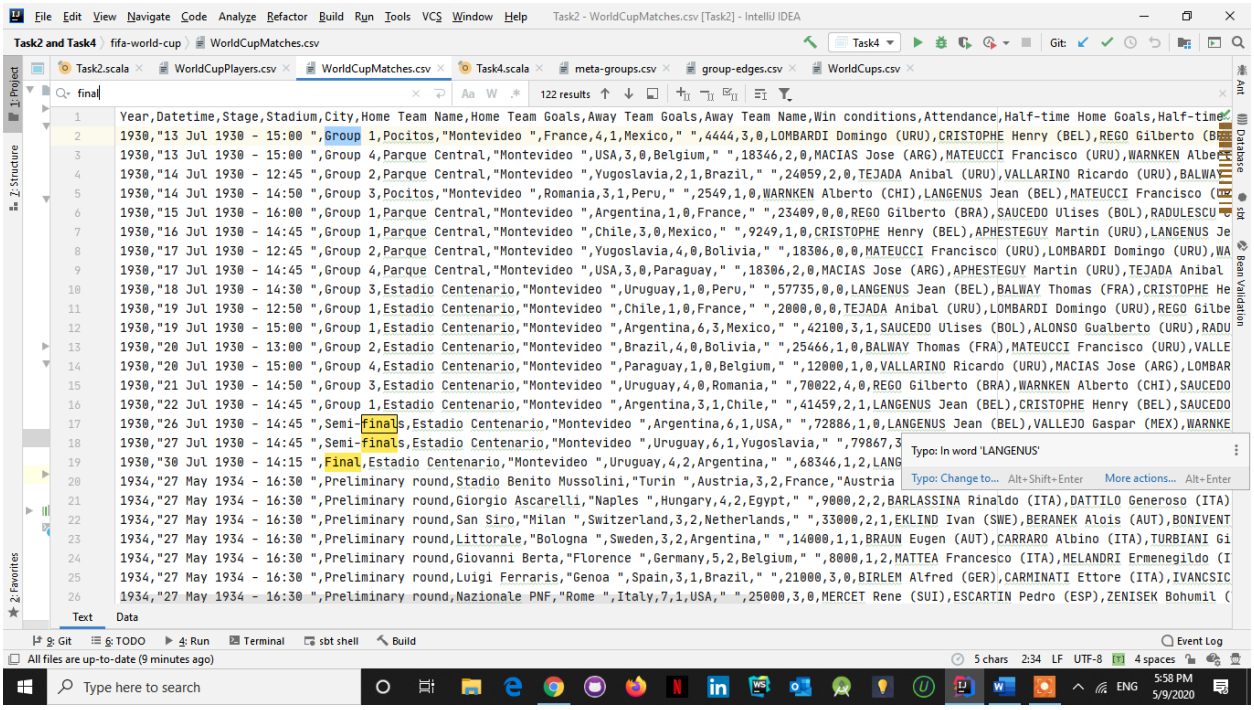
1. Winner of 1930
2. Years Uruguay was winner
3. Display Year, Country when Attendance was greater than 500.00
4. Display Year Country Winner where Hosted Country was the Winner
5. Display Year, Winner, Runners-Up, Third, Fourth where USA was in Winner, Runner-Up, Third, Fourth

Input

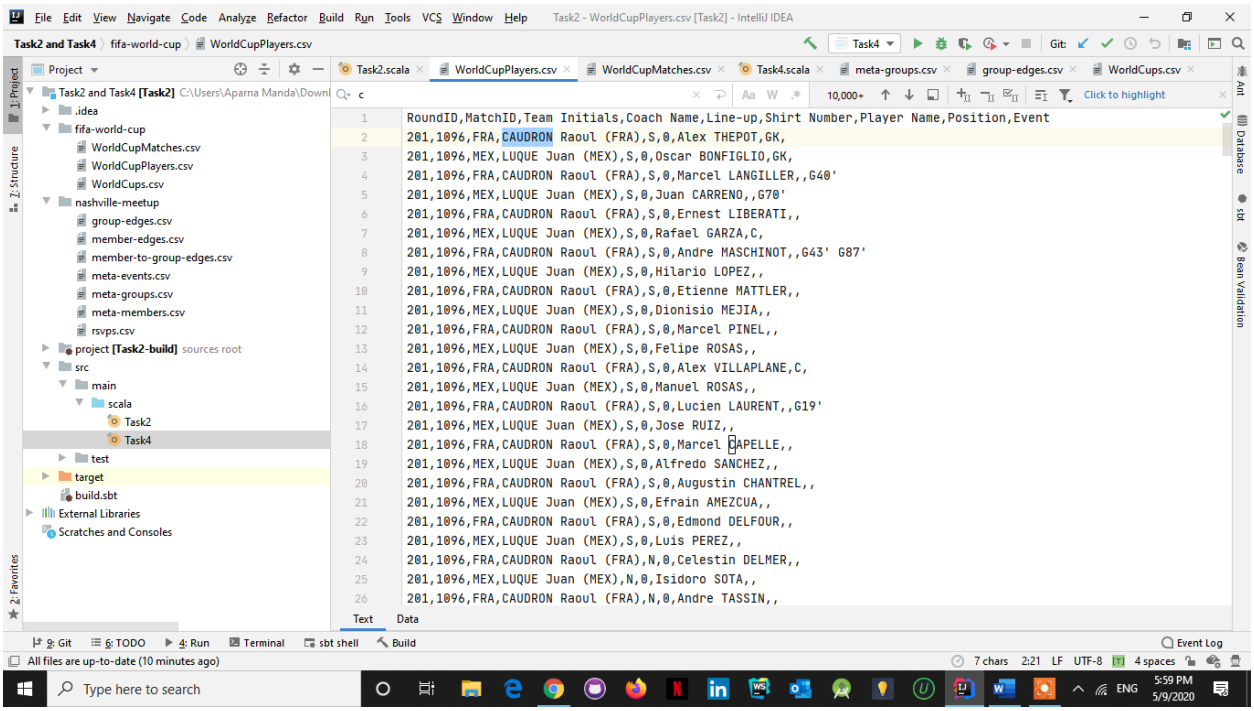
WorldCups.csv as WC View



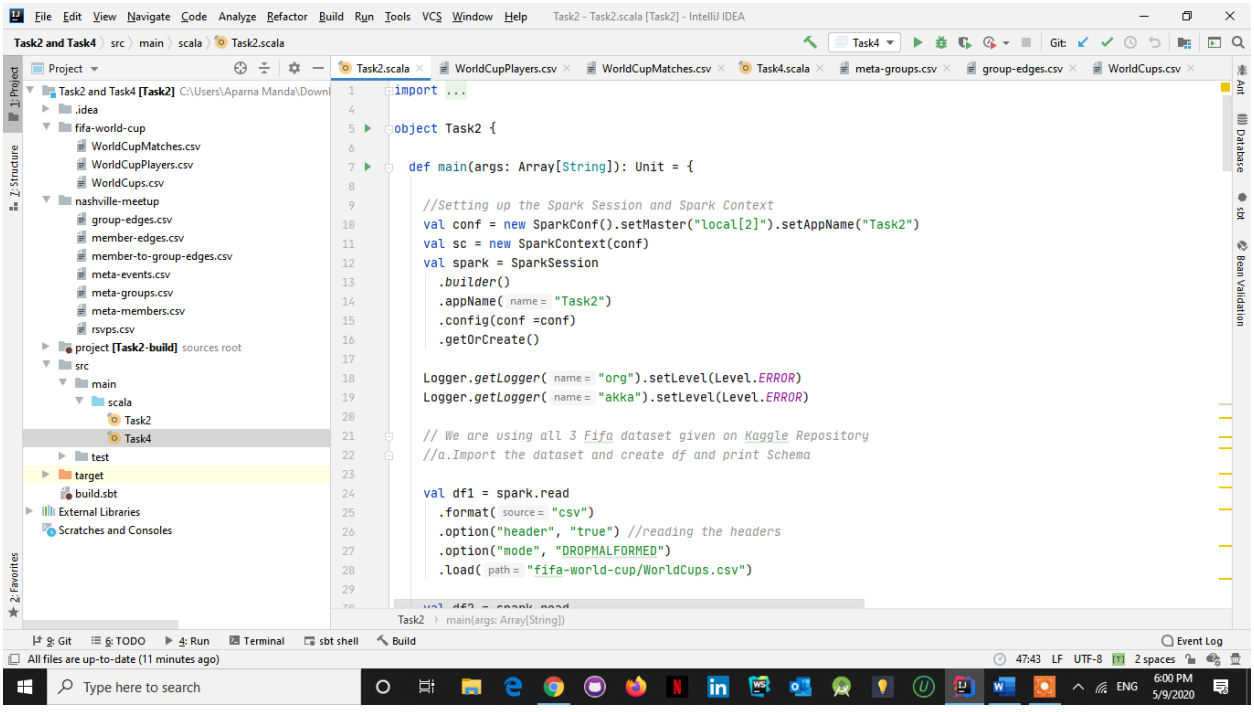
WorldCupMatches.csv as Matches View

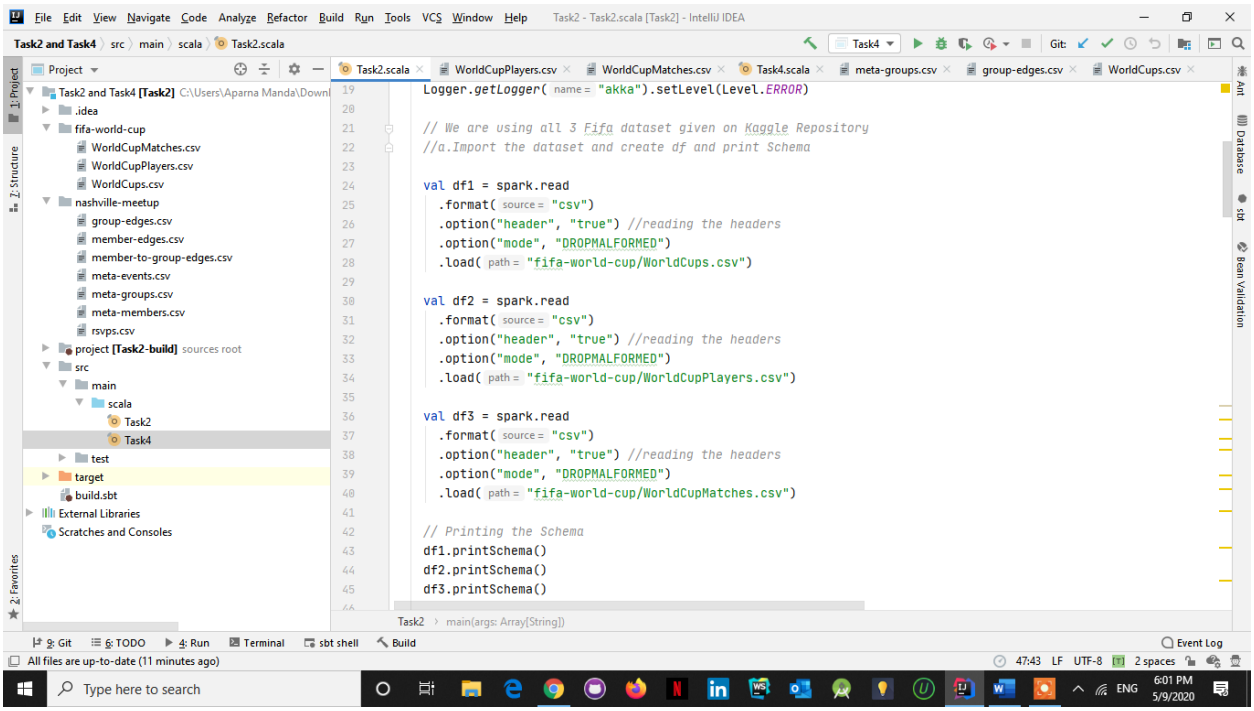


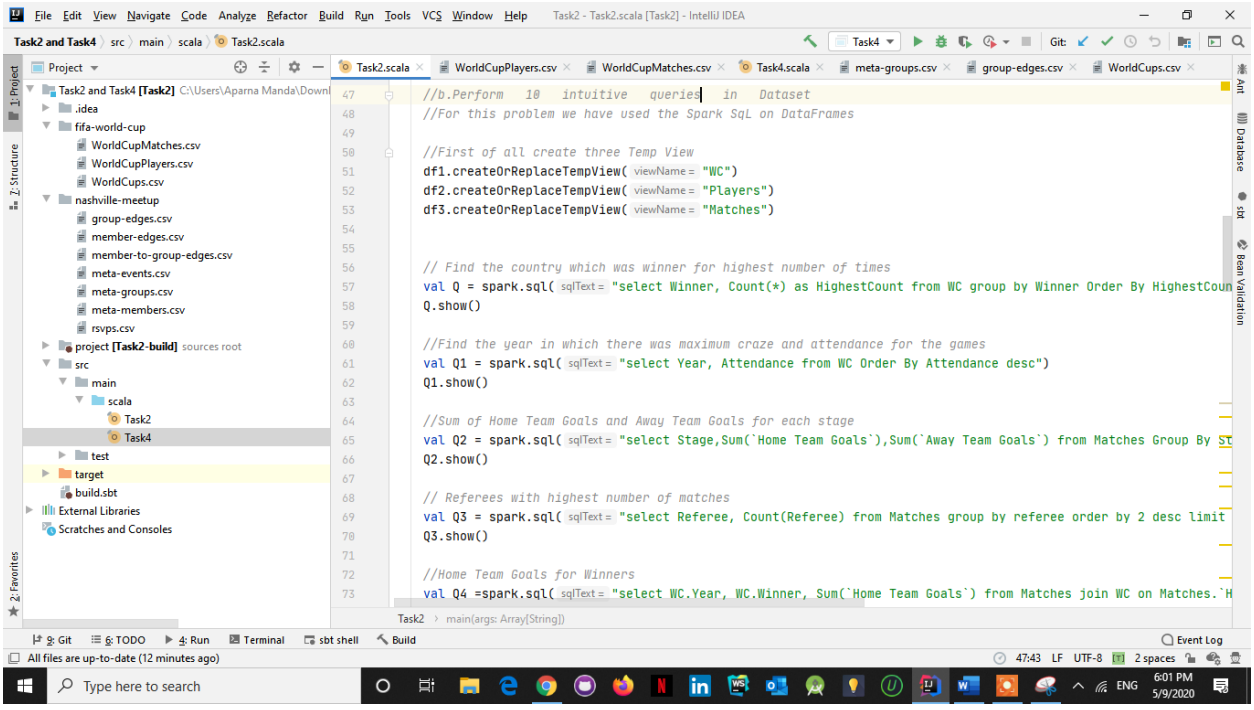
WorldCupPlayers.csv as Players View

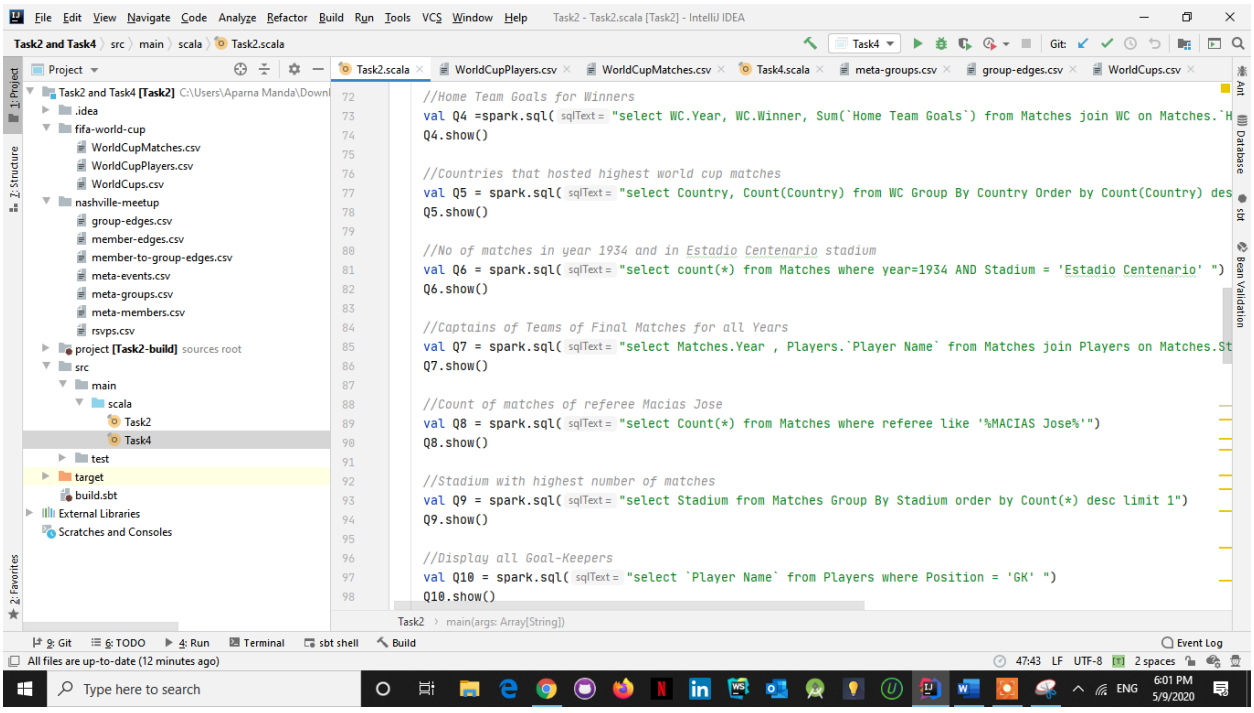

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - WorldCupPlayers.csv [Task2] - IntelliJ IDEA
Task2 and Task4 fifa-world-cup WorldCupPlayers.csv
Task2.scala WorldCupPlayers.csv WorldCupMatches.csv Task4.scala meta-groups.csv group-edges.csv WorldCups.csv
RoundID,MatchID,Team Initials,Coach Name,Line-up,Shirt Number,Player Name,Position,Event
201,1096,FRA,CAUDRON Raoul (FRA),S,0,Alex THEPOT,GK,
201,1096,MEX,LUQUE Juan (MEX),S,0,Oscar BONFIGLIO,GK,
201,1096,FRA,CAUDRON Raoul (FRA),S,0,Marcel LANGILLER,,G40'
201,1096,MEX,LUQUE Juan (MEX),S,0,Juan CARRENO,,G70'
201,1096,FRA,CAUDRON Raoul (FRA),S,0,Ernest LIBERATI,,
201,1096,MEX,LUQUE Juan (MEX),S,0,Rafael GARZA,C,
201,1096,FRA,CAUDRON Raoul (FRA),S,0,Andre MASCHINOT,,G43' G87'
201,1096,MEX,LUQUE Juan (MEX),S,0,Hilario LOPEZ,,
201,1096,FRA,CAUDRON Raoul (FRA),S,0,Etienne MATTLER,,
201,1096,MEX,LUQUE Juan (MEX),S,0,Dionisio MEJIA,,
201,1096,FRA,CAUDRON Raoul (FRA),S,0,Marcel PINEL,,
201,1096,MEX,LUQUE Juan (MEX),S,0,Felipe ROSAS,,
201,1096,FRA,CAUDRON Raoul (FRA),S,0,Alex VILLAPLANE,C,
201,1096,MEX,LUQUE Juan (MEX),S,0,Manuel ROSAS,,
201,1096,FRA,CAUDRON Raoul (FRA),S,0,Lucien LAURENT,,G19'
201,1096,MEX,LUQUE Juan (MEX),S,0,Jose RUIZ,,
201,1096,FRA,CAUDRON Raoul (FRA),S,0,Marcel CAPELLE,,
201,1096,MEX,LUQUE Juan (MEX),S,0,Alfredo SANCHEZ,,
201,1096,FRA,CAUDRON Raoul (FRA),S,0,Augustin CHANTREL,,
201,1096,MEX,LUQUE Juan (MEX),S,0,Efrain AMEZCUA,,
201,1096,FRA,CAUDRON Raoul (FRA),S,0,Edmond DELFOUR,,
201,1096,MEX,LUQUE Juan (MEX),S,0,Luis PEREZ,,
201,1096,FRA,CAUDRON Raoul (FRA),N,0,Celestin DELMER,,
201,1096,MEX,LUQUE Juan (MEX),N,0,Isidoro SOTA,,
201,1096,FRA,CAUDRON Raoul (FRA),N,0,Andre TASSIN,,
Text Data
All files are up-to-date (10 minutes ago)
Type here to search
5:59 PM 5/9/2020

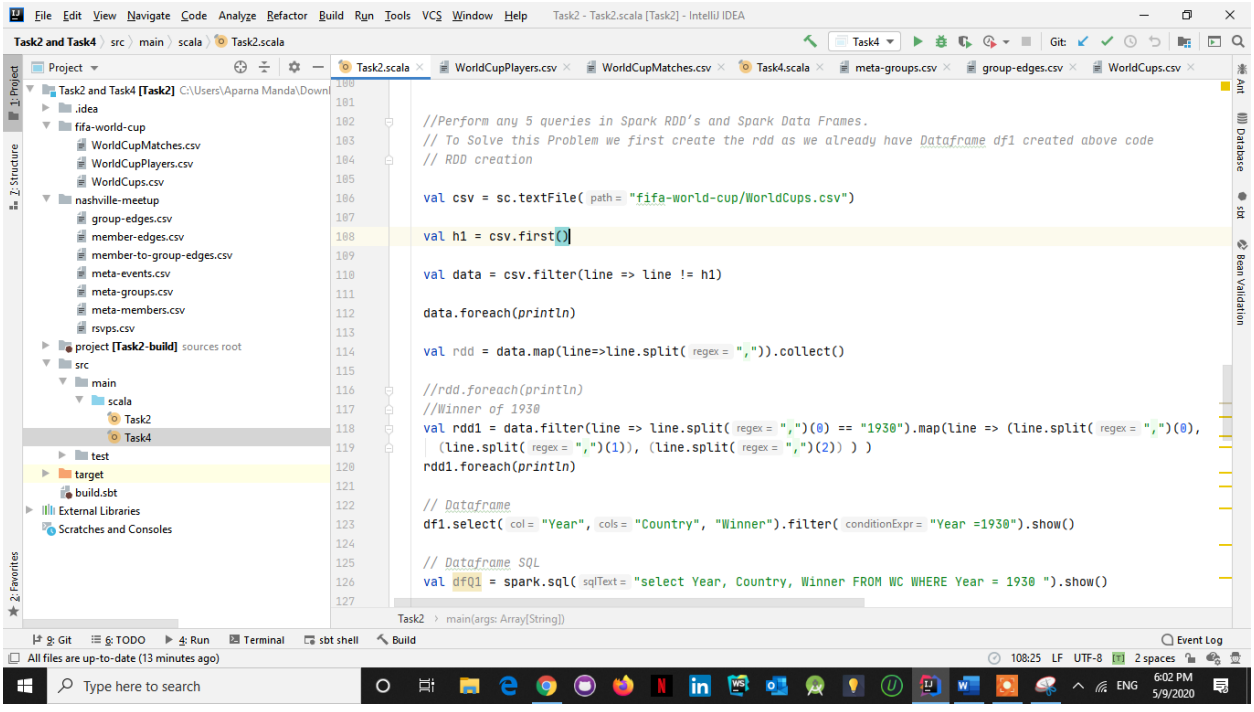
Code

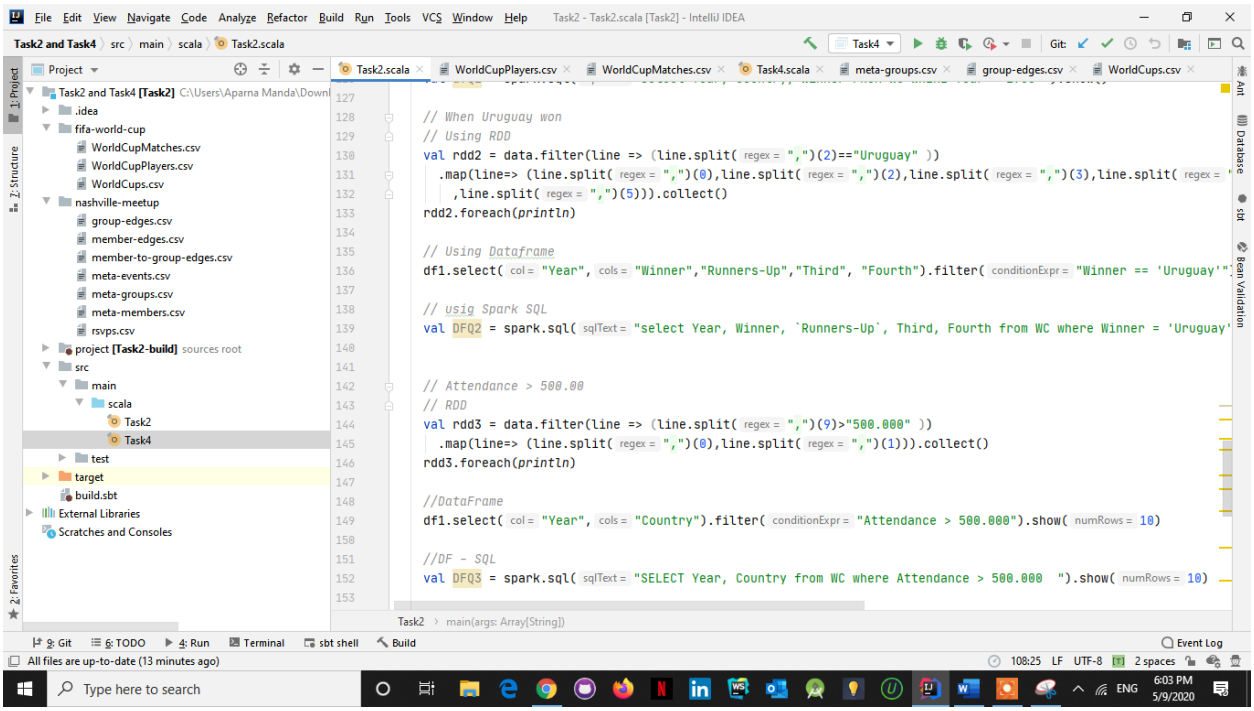

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Task2.scala WorldCupPlayers.csv WorldCupMatches.csv Task4.scala meta-groups.csv group-edges.csv WorldCups.csv
import ...
object Task2 {
def main(args: Array[String]): Unit = {
//Setting up the Spark Session and Spark Context
val conf = new SparkConf().setMaster("local[2]").setAppName("Task2")
val sc = new SparkContext(conf)
val spark = SparkSession
.builder()
.appName(name = "Task2")
.config(conf =conf)
.getOrCreate()
Logger.getLogger(name = "org").setLevel(Level.ERROR)
Logger.getLogger(name = "akka").setLevel(Level.ERROR)
// We are using all 3 Fifa dataset given on Kaggle Repository
//a.Import the dataset and create df and print Schema
val df1 = spark.read
.format(source = "csv")
.option("header", "true") //reading the headers
.option("mode", "DROPMALFORMED")
.load(path = "fifa-world-cup/WorldCups.csv")
Task2 main(args: Array[String])
All files are up-to-date (11 minutes ago)
Type here to search
6:00 PM 5/9/2020

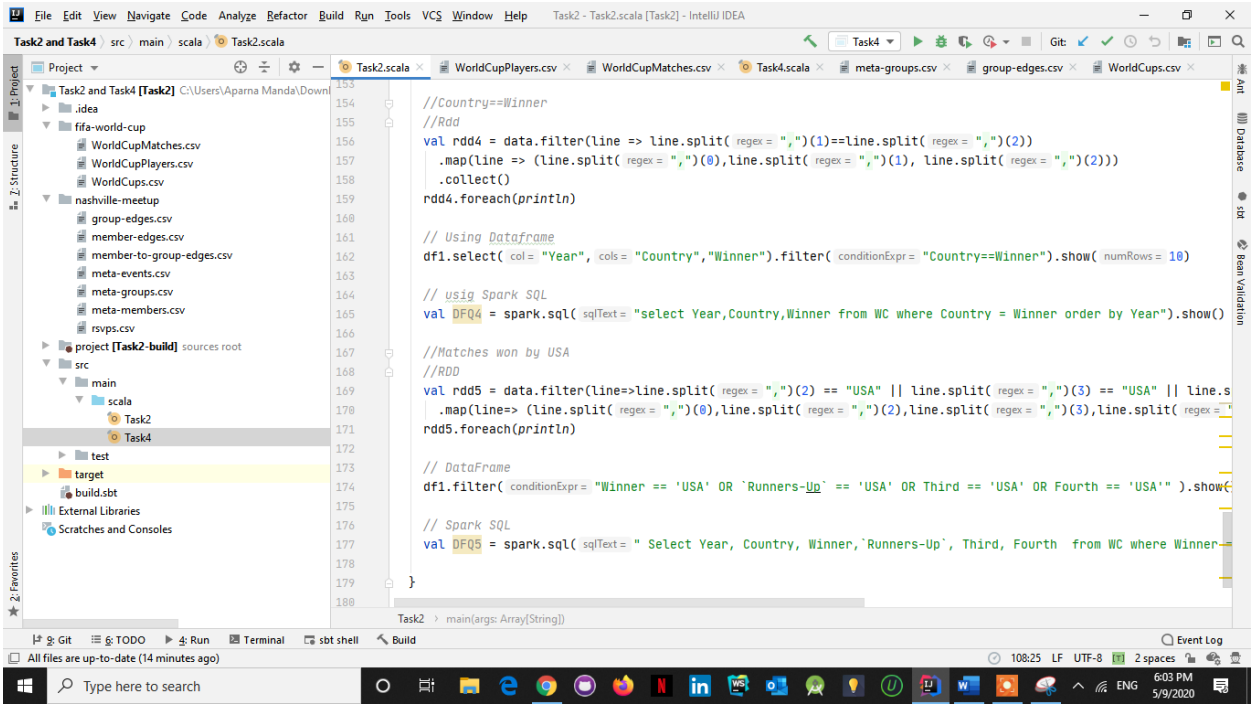

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Task4
Git:
Project
Task2 and Task4 [Task2] C:\Users\Aparna Manda\Downl
.idea
fifa-world-cup
WorldCupMatches.csv
WorldCupPlayers.csv
WorldCups.csv
nashville-meetup
group-edges.csv
member-edges.csv
member-to-group-edges.csv
meta-events.csv
meta-groups.csv
meta-members.csv
rsvps.csv
project [Task2-build] sources root
src
main
scala
Task2
Task4
test
target
build.sbt
External Libraries
Scratches and Consoles
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
Logger.getLogger(name = "akka").setLevel(Level.ERROR)
// We are using all 3 Fifa dataset given on Kaggle Repository
//a.Import the dataset and create df and print Schema
val df1 = spark.read
.format(source = "csv")
.option("header", "true") //reading the headers
.option("mode", "DROPMALFORMED")
.load(path = "fifa-world-cup/WorldCups.csv")
val df2 = spark.read
.format(source = "csv")
.option("header", "true") //reading the headers
.option("mode", "DROPMALFORMED")
.load(path = "fifa-world-cup/WorldCupPlayers.csv")
val df3 = spark.read
.format(source = "csv")
.option("header", "true") //reading the headers
.option("mode", "DROPMALFORMED")
.load(path = "fifa-world-cup/WorldCupMatches.csv")
// Printing the Schema
df1.printSchema()
df2.printSchema()
df3.printSchema()
Task2 main(args: Array[String])
9: Git
6: TODO
4: Run
Terminal
sbt shell
Build
Event Log
All files are up-to-date (11 minutes ago)
47:43 LF UTF-8 2 spaces
Type here to search
ENG
6:01 PM
5/9/2020


File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Task4
Git:
Project
Task2 and Task4 [Task2] C:\Users\Aparna Manda\Downl
.idea
fifa-world-cup
WorldCupMatches.csv
WorldCupPlayers.csv
WorldCups.csv
nashville-meetup
group-edges.csv
member-edges.csv
member-to-group-edges.csv
meta-events.csv
meta-groups.csv
meta-members.csv
rsvps.csv
project [Task2-build] sources root
src
main
scala
Task2
Task4
test
target
build.sbt
External Libraries
Scratches and Consoles
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
//b.Perform 10 intuitive queries in Dataset
//For this problem we have used the Spark SqL on DataFrames
//First of all create three Temp View
df1.createOrReplaceTempView(viewName = "WC")
df2.createOrReplaceTempView(viewName = "Players")
df3.createOrReplaceTempView(viewName = "Matches")
// Find the country which was winner for highest number of times
val Q = spark.sql(sqlText = "select Winner, Count(*) as HighestCount from WC group by Winner Order By HighestCoun
Q.show()
//Find the year in which there was maximum craze and attendance for the games
val Q1 = spark.sql(sqlText = "select Year, Attendance from WC Order By Attendance desc")
Q1.show()
//Sum of Home Team Goals and Away Team Goals for each stage
val Q2 = spark.sql(sqlText = "select Stage,Sum(`Home Team Goals`),Sum(`Away Team Goals`) from Matches Group By St
Q2.show()
// Referees with highest number of matches
val Q3 = spark.sql(sqlText = "select Referee, Count(Referee) from Matches group by referee order by 2 desc limit
Q3.show()
//Home Team Goals for Winners
val Q4 =spark.sql(sqlText = "select WC.Year, WC.Winner, Sum(`Home Team Goals`) from Matches join WC on Matches.`H
Task2 main(args: Array[String])
9: Git
6: TODO
4: Run
Terminal
sbt shell
Build
Event Log
All files are up-to-date (12 minutes ago)
47:43 LF UTF-8 2 spaces
Type here to search
ENG
6:01 PM
5/9/2020

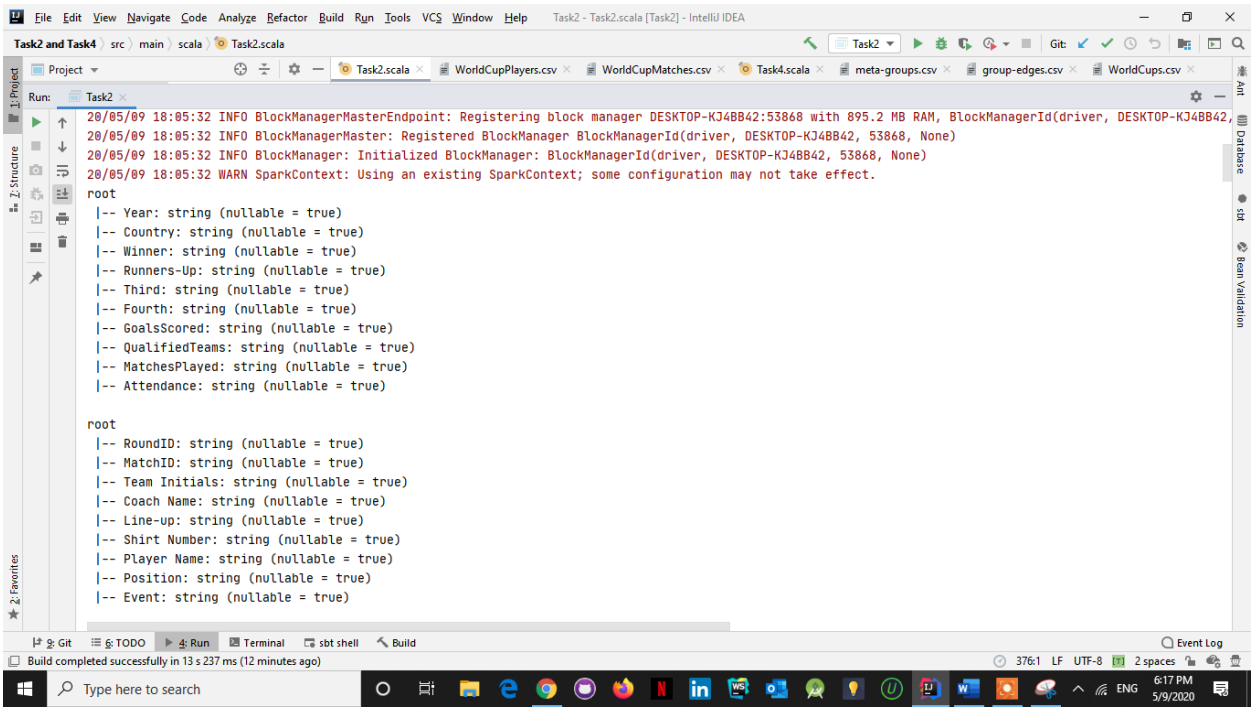

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Task4
Git:
Project
Task2 and Task4 [Task2] C:\Users\Aparna Manda\Downl
.idea
fifa-world-cup
WorldCupMatches.csv
WorldCupPlayers.csv
WorldCups.csv
nashville-meetup
group-edges.csv
member-edges.csv
member-to-group-edges.csv
meta-events.csv
meta-groups.csv
meta-members.csv
rsvps.csv
project [Task2-build] sources root
src
main
scala
Task2
Task4
test
target
build.sbt
External Libraries
Scratches and Consoles
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
//Home Team Goals for Winners
val Q4 =spark.sql(sqlText = "select WC.Year, WC.Winner, Sum(`Home Team Goals`) from Matches join WC on Matches.`H
Q4.show()
//Countries that hosted highest world cup matches
val Q5 = spark.sql(sqlText = "select Country, Count(Country) from WC Group By Country Order by Count(Country) des
Q5.show()
//No of matches in year 1934 and in Estadio Centenario stadium
val Q6 = spark.sql(sqlText = "select count(*) from Matches where year=1934 AND Stadium = 'Estadio Centenario' ")
Q6.show()
//Captains of Teams of Final Matches for all Years
val Q7 = spark.sql(sqlText = "select Matches.Year , Players.`Player Name` from Matches join Players on Matches.St
Q7.show()
//Count of matches of referee Macias Jose
val Q8 = spark.sql(sqlText = "select Count(*) from Matches where referee like '%MACIAS Jose%'")
Q8.show()
//Stadium with highest number of matches
val Q9 = spark.sql(sqlText = "select Stadium from Matches Group By Stadium order by Count(*) desc limit 1")
Q9.show()
//Display all Goal-Keepers
val Q10 = spark.sql(sqlText = "select `Player Name` from Players where Position = 'GK' ")
Q10.show()
Task2 main(args: Array[String])
9: Git
6: TODO
4: Run
Terminal
sbt shell
Build
Event Log
All files are up-to-date (12 minutes ago)
47:43 LF UTF-8 2 spaces
Type here to search
ENG
6:01 PM
5/9/2020

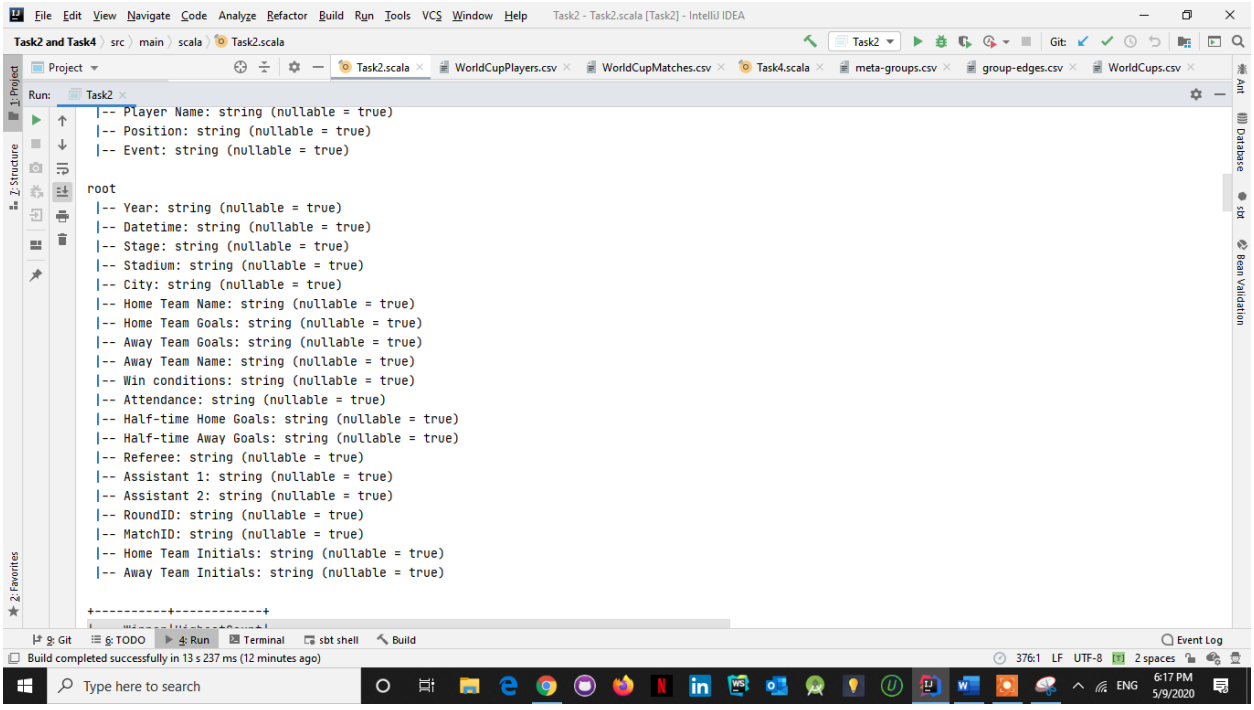

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Task4
Git:
Project
Task2 and Task4 [Task2] C:\Users\Aparna Manda\Downl
.idea
fifa-world-cup
WorldCupMatches.csv
WorldCupPlayers.csv
WorldCups.csv
nashville-meetup
group-edges.csv
member-edges.csv
member-to-group-edges.csv
meta-events.csv
meta-groups.csv
meta-members.csv
rsvps.csv
project [Task2-build] sources root
src
main
scala
Task2
Task4
test
target
build.sbt
External Libraries
Scratches and Consoles
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
//Perform any 5 queries in Spark RDD's and Spark Data Frames.
// To Solve this Problem we first create the rdd as we already have Dataframe df1 created above code
// RDD creation
val csv = sc.textFile(path = "fifa-world-cup/WorldCups.csv")
val h1 = csv.first()
val data = csv.filter(line => line != h1)
data.foreach(println)
val rdd = data.map(line=>line.split(regex = ",")).collect()
//rdd.foreach(println)
//Winner of 1930
val rdd1 = data.filter(line => line.split(regex = ",")(0) == "1930").map(line => (line.split(regex = ",")(0),
(line.split(regex = ",")(1)), (line.split(regex = ",")(2))))
rdd1.foreach(println)
// Dataframe
df1.select(col = "Year", cols = "Country", "Winner").filter(conditionExpr = "Year =1930").show()
// Dataframe SQL
val dfQ1 = spark.sql(sqlText = "select Year, Country, Winner FROM WC WHERE Year = 1930 ").show()
Task2 main(args: Array[String])
9: Git
6: TODO
4: Run
Terminal
sbt shell
Build
Event Log
All files are up-to-date (13 minutes ago)
108:25 LF UTF-8 2 spaces
Type here to search
ENG
6:02 PM
5/9/2020

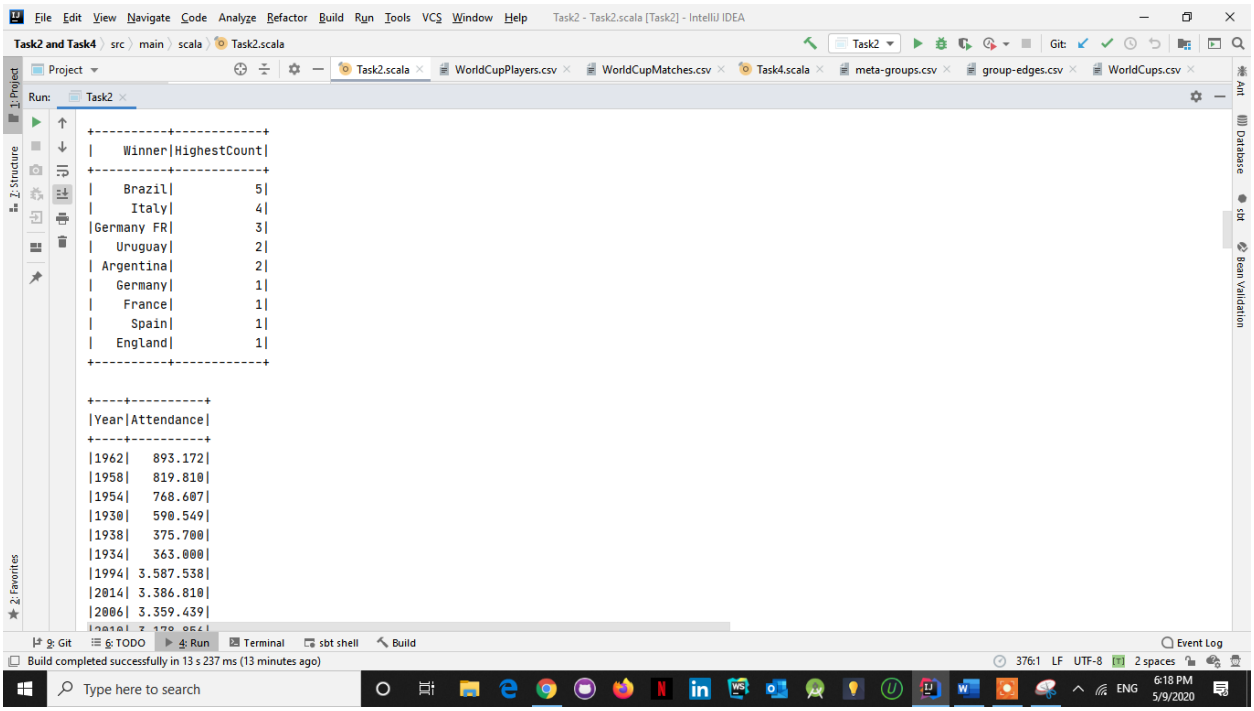

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Task4
Git:
Project
Task2 and Task4 [Task2] C:\Users\Aparna Manda\Down
.idea
fifa-world-cup
WorldCupMatches.csv
WorldCupPlayers.csv
WorldCups.csv
nashville-meetup
group-edges.csv
member-edges.csv
member-to-group-edges.csv
meta-events.csv
meta-groups.csv
meta-members.csv
rsvps.csv
project [Task2-build] sources root
src
main
scala
Task2
Task4
test
target
build.sbt
External Libraries
Scratches and Consoles
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
// When Uruguay won
// Using RDD
val rdd2 = data.filter(line => (line.split(regex = ",")(2)=="Uruguay"))
.map(line=> (line.split(regex = ",")(0),line.split(regex = ",")(2),line.split(regex = ",")(3),line.split(regex =
,line.split(regex = ",")(5))).collect()
rdd2.foreach(println)
// Using Dataframe
df1.select(col = "Year", cols = "Winner","Runners-Up","Third", "Fourth").filter(conditionExpr = "Winner == 'Uruguay'"
// usig Spark SQL
val DFQ2 = spark.sql(sqlText = "select Year, Winner, `Runners-Up`, Third, Fourth from WC where Winner = 'Uruguay'
// Attendance > 500.00
// RDD
val rdd3 = data.filter(line => (line.split(regex = ",")(9)>"500.000"))
.map(line=> (line.split(regex = ",")(0),line.split(regex = ",")(1))).collect()
rdd3.foreach(println)
//DataFrame
df1.select(col = "Year", cols = "Country").filter(conditionExpr = "Attendance > 500.000").show(numRows = 10)
//DF - SQL
val DFQ3 = spark.sql(sqlText = "SELECT Year, Country from WC where Attendance > 500.000 ").show(numRows = 10)
Task2 main(args: Array[String])
9: Git 6: TODO 4: Run Terminal sbt shell Build
Event Log
All files are up-to-date (13 minutes ago)
108:25 LF UTF-8 2 spaces
Type here to search
ENG
6:03 PM
5/9/2020


File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Task4
Git:
Project
Task2 and Task4 [Task2] C:\Users\Aparna Manda\Down
.idea
fifa-world-cup
WorldCupMatches.csv
WorldCupPlayers.csv
WorldCups.csv
nashville-meetup
group-edges.csv
member-edges.csv
member-to-group-edges.csv
meta-events.csv
meta-groups.csv
meta-members.csv
rsvps.csv
project [Task2-build] sources root
src
main
scala
Task2
Task4
test
target
build.sbt
External Libraries
Scratches and Consoles
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
//Country==Winner
//Rdd
val rdd4 = data.filter(line => line.split(regex = ",")(1)==line.split(regex = ",")(2))
.map(line => (line.split(regex = ",")(0),line.split(regex = ",")(1), line.split(regex = ",")(2)))
.collect()
rdd4.foreach(println)
// Using Dataframe
df1.select(col = "Year", cols = "Country","Winner").filter(conditionExpr = "Country==Winner").show(numRows = 10)
// usig Spark SQL
val DFQ4 = spark.sql(sqlText = "select Year,Country,Winner from WC where Country = Winner order by Year").show()
//Matches won by USA
//RDD
val rdd5 = data.filter(line=>line.split(regex = ",")(2) == "USA" || line.split(regex = ",")(3) == "USA" || line.s
.map(line=> (line.split(regex = ",")(0),line.split(regex = ",")(2),line.split(regex = ",")(3),line.split(regex =
rdd5.foreach(println)
// DataFrame
df1.filter(conditionExpr = "Winner == 'USA' OR `Runners-Up` == 'USA' OR Third == 'USA' OR Fourth == 'USA'").show(
// Spark SQL
val DFQ5 = spark.sql(sqlText = " Select Year, Country, Winner,`Runners-Up`, Third, Fourth from WC where Winner
}
Task2 main(args: Array[String])
9: Git 6: TODO 4: Run Terminal sbt shell Build
Event Log
All files are up-to-date (14 minutes ago)
108:25 LF UTF-8 2 spaces
Type here to search
ENG
6:03 PM
5/9/2020

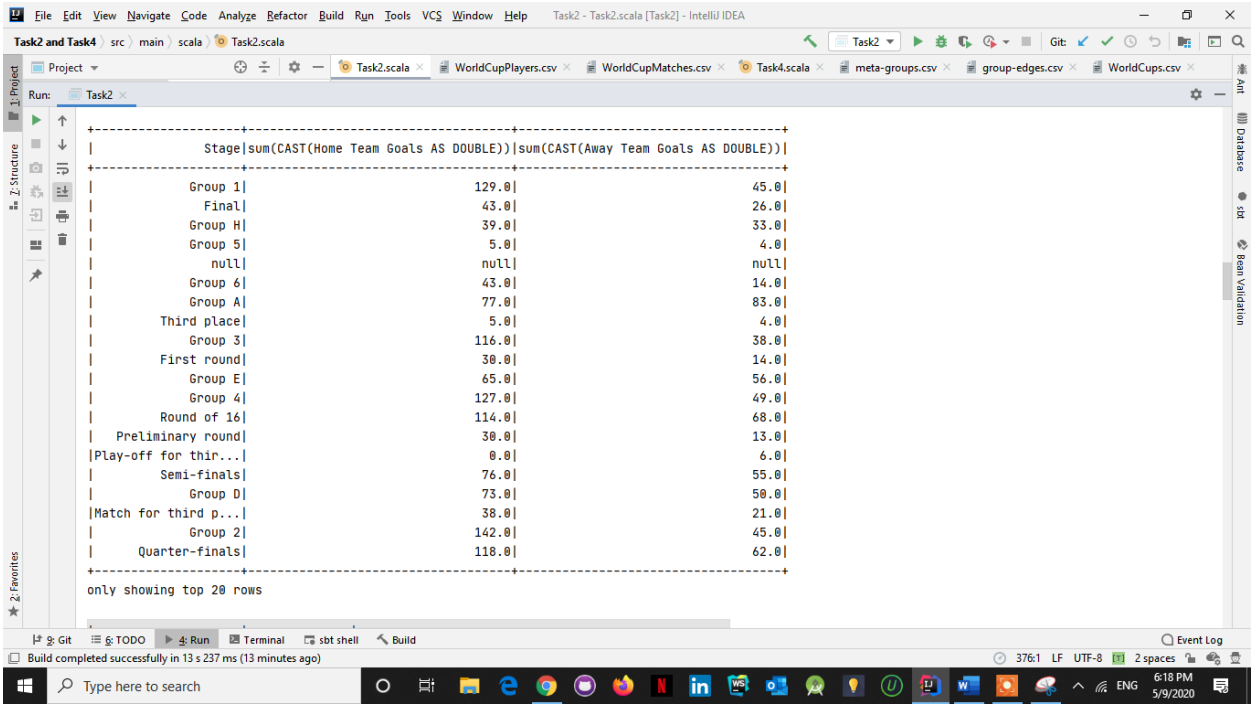
Output


File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Task2
Git:
Project
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
Run:
Task2
20/05/09 18:05:32 INFO BlockManagerMasterEndpoint: Registering block manager DESKTOP-KJ4BB42:53868 with 895.2 MB RAM, BlockManagerId(driver, DESKTOP-KJ4BB42,
20/05/09 18:05:32 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, DESKTOP-KJ4BB42, 53868, None)
20/05/09 18:05:32 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, DESKTOP-KJ4BB42, 53868, None)
20/05/09 18:05:32 WARN SparkContext: Using an existing SparkContext; some configuration may not take effect.
root
|-- Year: string (nullable = true)
|-- Country: string (nullable = true)
|-- Winner: string (nullable = true)
|-- Runners-Up: string (nullable = true)
|-- Third: string (nullable = true)
|-- Fourth: string (nullable = true)
|-- GoalsScored: string (nullable = true)
|-- QualifiedTeams: string (nullable = true)
|-- MatchesPlayed: string (nullable = true)
|-- Attendance: string (nullable = true)
root
|-- RoundID: string (nullable = true)
|-- MatchID: string (nullable = true)
|-- Team Initials: string (nullable = true)
|-- Coach Name: string (nullable = true)
|-- Line-up: string (nullable = true)
|-- Shirt Number: string (nullable = true)
|-- Player Name: string (nullable = true)
|-- Position: string (nullable = true)
|-- Event: string (nullable = true)
9: Git
6: TODO
4: Run
Terminal
sbt shell
Build
Event Log
Build completed successfully in 13 s 237 ms (12 minutes ago)
376:1 LF UTF-8 2 spaces
Type here to search
ENG
6:17 PM
5/9/2020

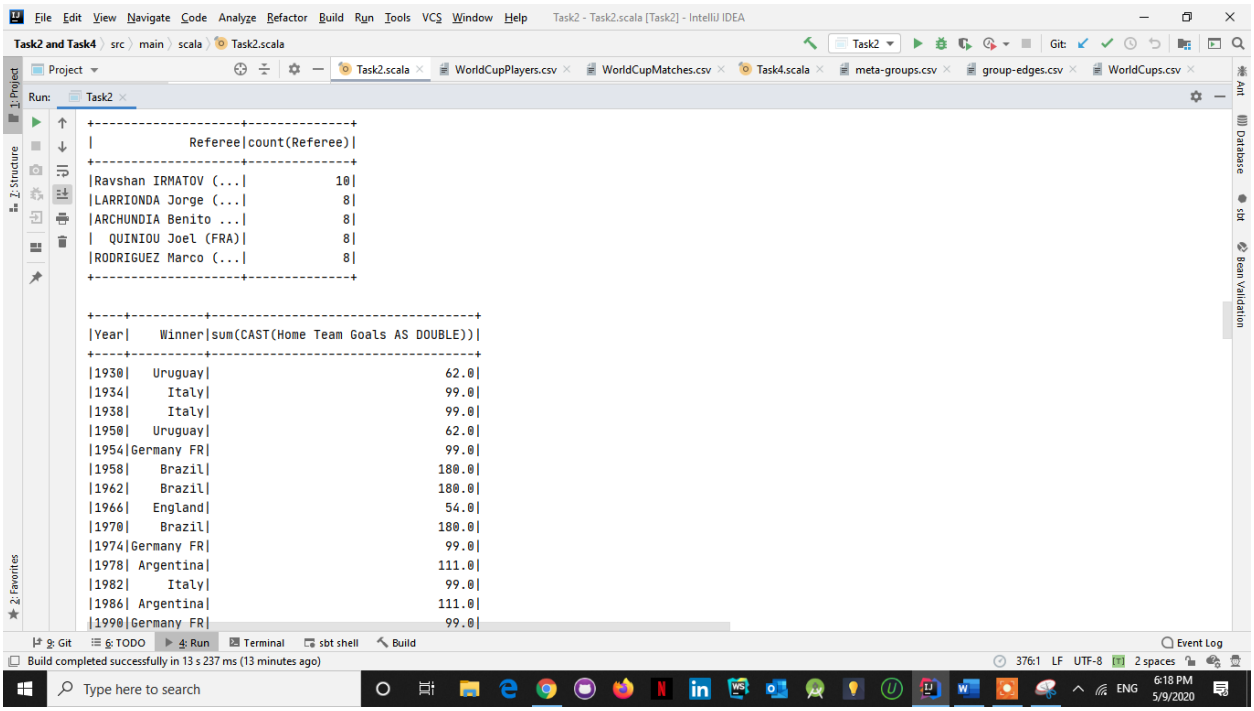

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Task2
Git:
Project
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
Run:
Task2
|-- Player Name: string (nullable = true)
|-- Position: string (nullable = true)
|-- Event: string (nullable = true)
root
|-- Year: string (nullable = true)
|-- Datetime: string (nullable = true)
|-- Stage: string (nullable = true)
|-- Stadium: string (nullable = true)
|-- City: string (nullable = true)
|-- Home Team Name: string (nullable = true)
|-- Home Team Goals: string (nullable = true)
|-- Away Team Goals: string (nullable = true)
|-- Away Team Name: string (nullable = true)
|-- Win conditions: string (nullable = true)
|-- Attendance: string (nullable = true)
|-- Half-time Home Goals: string (nullable = true)
|-- Half-time Away Goals: string (nullable = true)
|-- Referee: string (nullable = true)
|-- Assistant 1: string (nullable = true)
|-- Assistant 2: string (nullable = true)
|-- RoundID: string (nullable = true)
|-- MatchID: string (nullable = true)
|-- Home Team Initials: string (nullable = true)
|-- Away Team Initials: string (nullable = true)
+----------+------------+
9: Git
6: TODO
4: Run
Terminal
sbt shell
Build
Event Log
Build completed successfully in 13 s 237 ms (12 minutes ago)
376:1 LF UTF-8 2 spaces
Type here to search
ENG
6:17 PM
5/9/2020

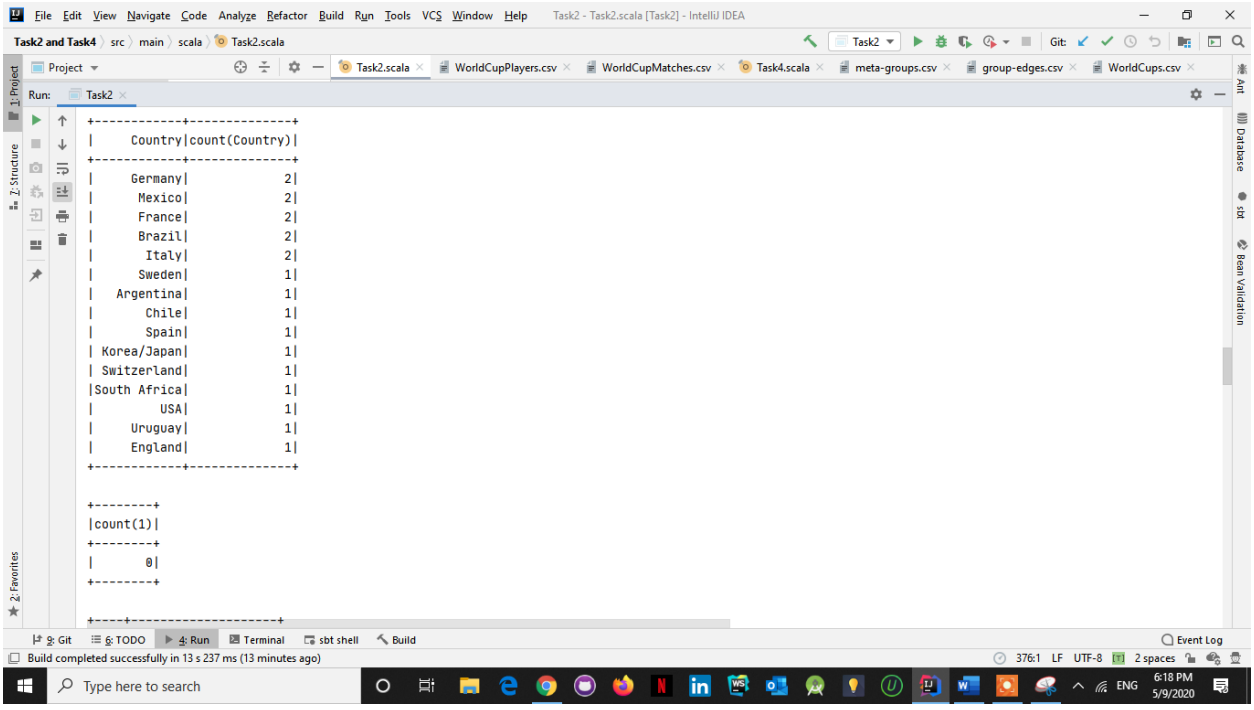


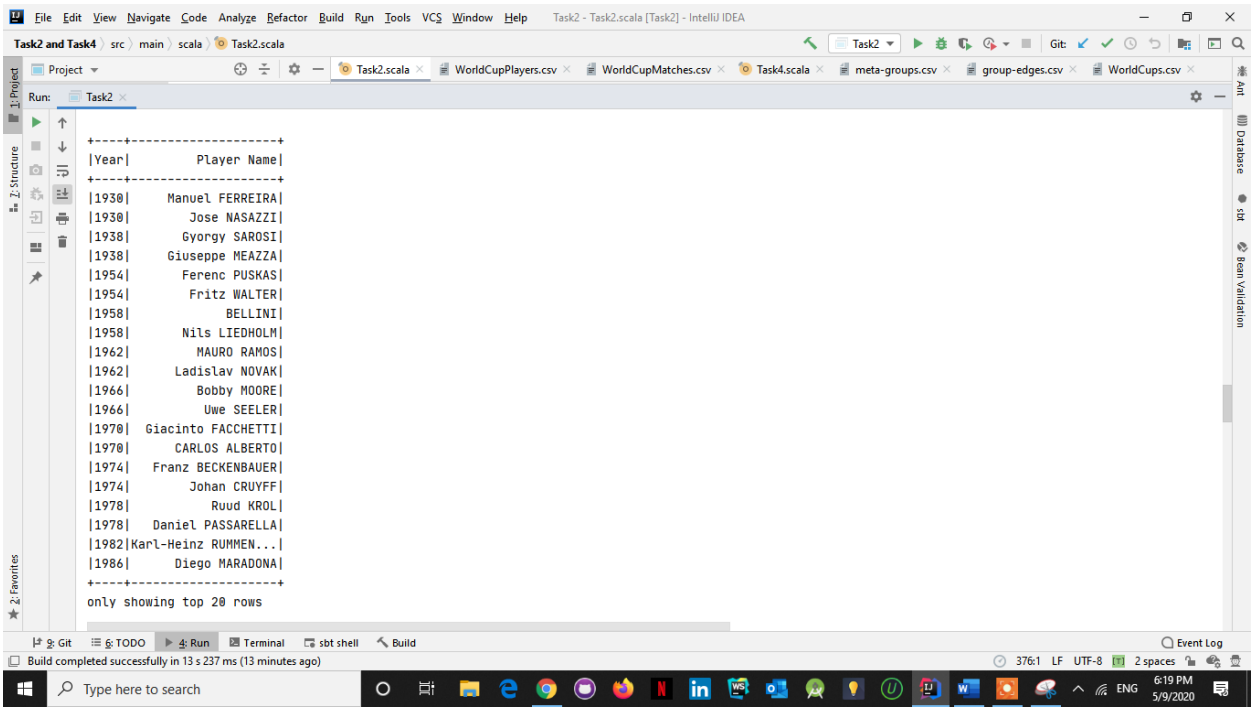
```
+----------+------------+
|    Winner|HighestCount|
+----------+------------+
|    Brazil|           5|
|     Italy|           4|
|Germany FR|           3|
|   Uruguay|           2|
| Argentina|           2|
|   Germany|           1|
|    France|           1|
|     Spain|           1|
|   England|           1|
+----------+------------+

+----+----------+
|Year|Attendance|
+----+----------+
|1962|   893.172|
|1958|   819.810|
|1954|   768.607|
|1930|   590.549|
|1938|   375.700|
|1934|   363.000|
|1994| 3.587.538|
|2014| 3.386.810|
|2006| 3.359.439|
```

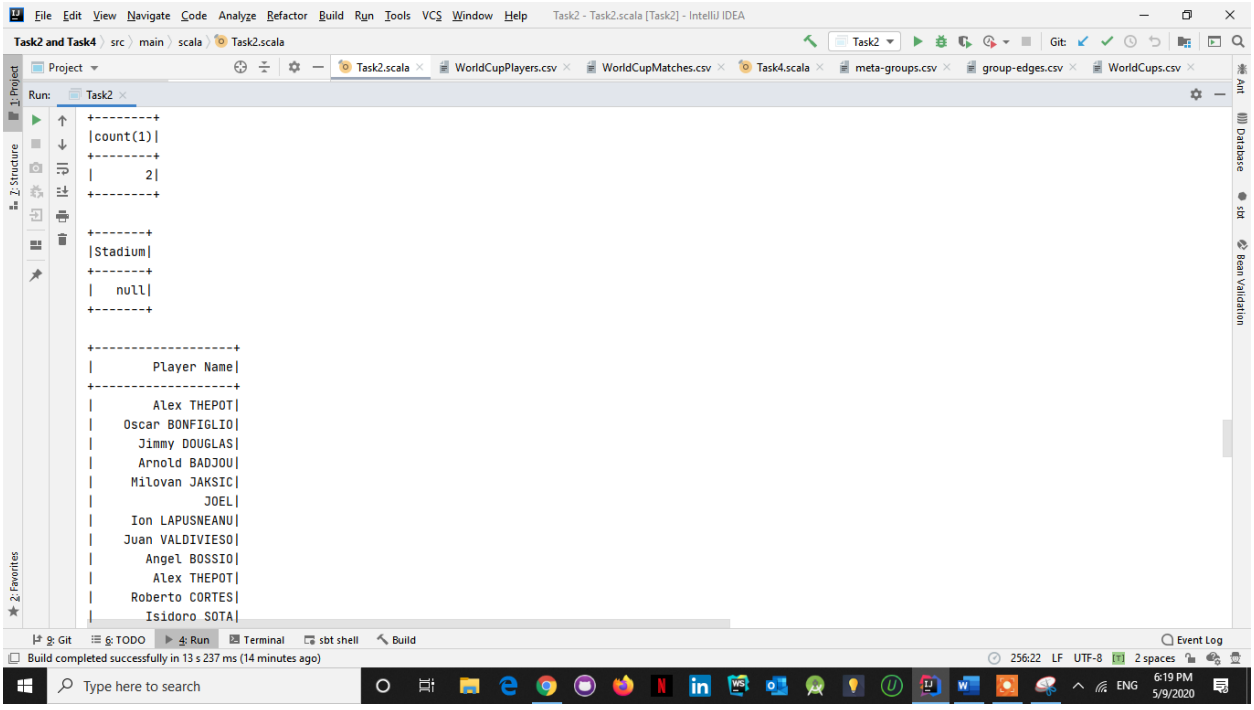


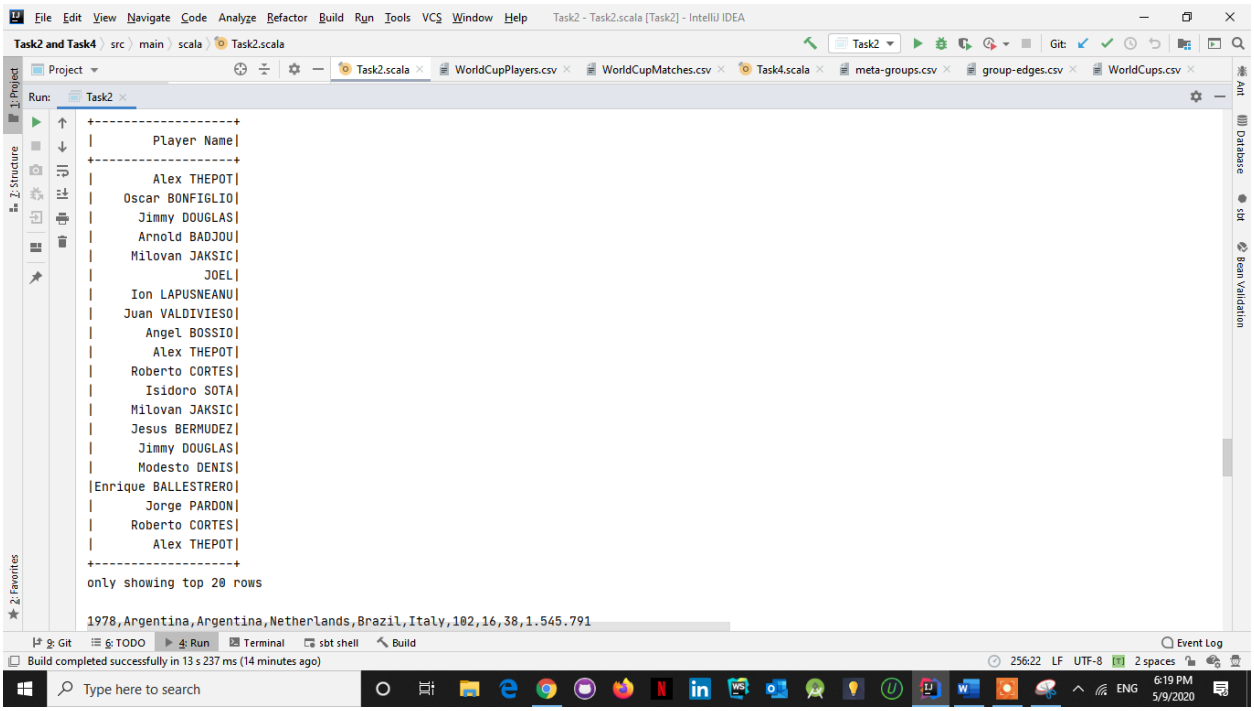
```
+--------------------+-----------------------------------------+-----------------------------------+
|               Stage|sum(CAST(Home Team Goals AS DOUBLE))|sum(CAST(Away Team Goals AS DOUBLE))|
+--------------------+-----------------------------------------+-----------------------------------+
|             Group 1|                                    129.0|                               45.0|
|               Final|                                     43.0|                               26.0|
|             Group H|                                     39.0|                               33.0|
|             Group 5|                                      5.0|                                4.0|
|                null|                                     null|                               null|
|             Group 6|                                     43.0|                               14.0|
|             Group A|                                     77.0|                               83.0|
|         Third place|                                      5.0|                                4.0|
|             Group 3|                                    116.0|                               38.0|
|         First round|                                     30.0|                               14.0|
|             Group E|                                     65.0|                               56.0|
|             Group 4|                                    127.0|                               49.0|
|         Round of 16|                                    114.0|                               68.0|
|   Preliminary round|                                     30.0|                               13.0|
|Play-off for thir...|                                      0.0|                                6.0|
|         Semi-finals|                                     76.0|                               55.0|
|             Group D|                                     73.0|                               50.0|
|Match for third p...|                                     38.0|                               21.0|
|             Group 2|                                    142.0|                               45.0|
|      Quarter-finals|                                    118.0|                               62.0|
+--------------------+-----------------------------------------+-----------------------------------+
only showing top 20 rows
```

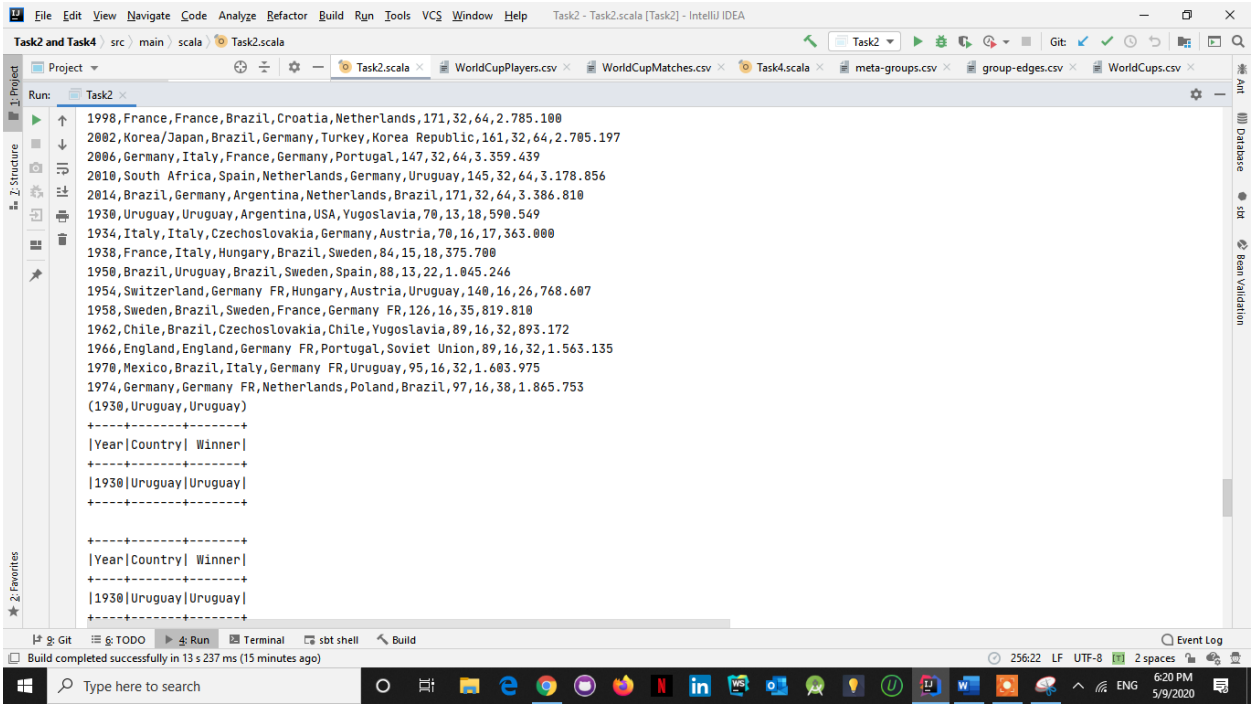

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 › src › main › scala › Task2.scala
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
Run: Task2
+--------------------+--------------+
| Referee|count(Referee)|
+--------------------+--------------+
|Ravshan IRMATOV (...| 10|
|LARRIONDA Jorge (...| 8|
|ARCHUNDIA Benito ...| 8|
| QUINIOU Joel (FRA)| 8|
|RODRIGUEZ Marco (...| 8|
+--------------------+--------------+
+----+----------+------------------------------------+
|Year| Winner|sum(CAST(Home Team Goals AS DOUBLE))|
+----+----------+------------------------------------+
|1930| Uruguay| 62.0|
|1934| Italy| 99.0|
|1938| Italy| 99.0|
|1950| Uruguay| 62.0|
|1954|Germany FR| 99.0|
|1958| Brazil| 180.0|
|1962| Brazil| 180.0|
|1966| England| 54.0|
|1970| Brazil| 180.0|
|1974|Germany FR| 99.0|
|1978| Argentina| 111.0|
|1982| Italy| 99.0|
|1986| Argentina| 111.0|
|1990|Germany FR| 99.0|
9: Git
6: TODO
4: Run
Terminal
sbt shell
Build
Event Log
Build completed successfully in 13 s 237 ms (13 minutes ago)
376:1 LF UTF-8 2 spaces
Type here to search
ENG
6:18 PM
5/9/2020

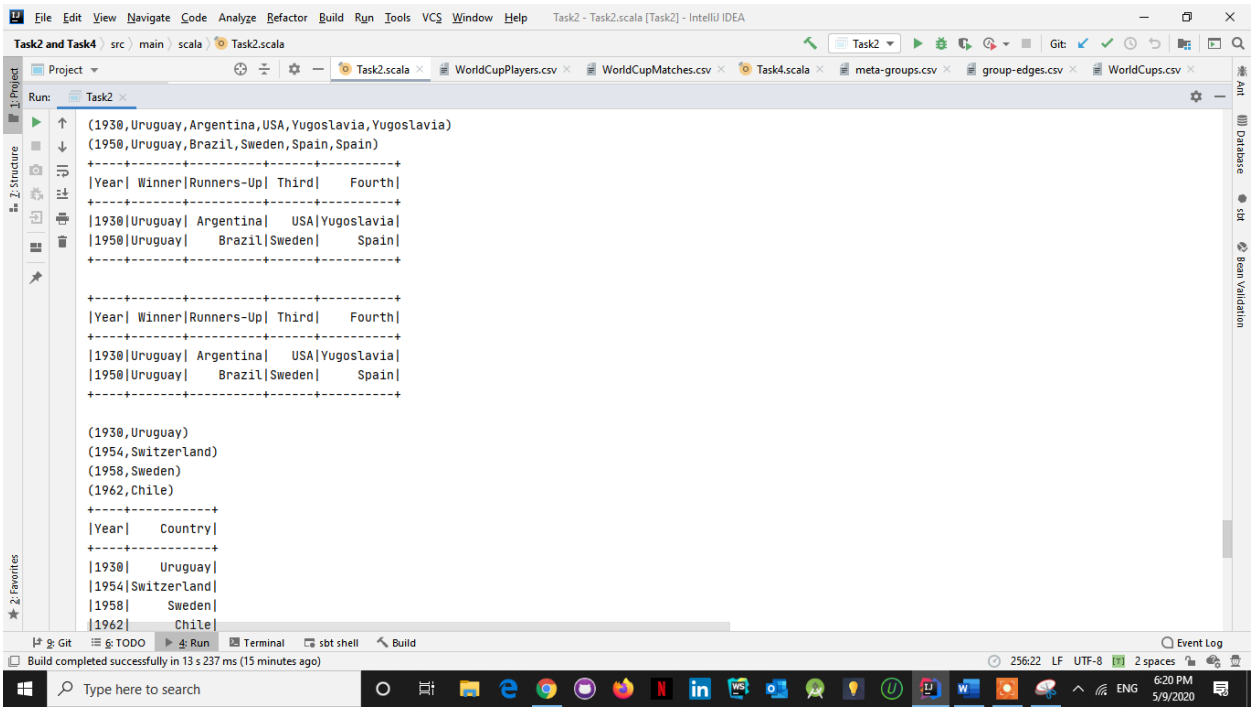

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 › src › main › scala › Task2.scala
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
Run: Task2
+------------+--------------+
| Country|count(Country)|
+------------+--------------+
Germany	2
Mexico	2
France	2
Brazil	2
Italy	2
Sweden	1
Argentina	1
Chile	1
Spain	1
Korea/Japan	1
Switzerland	1
South Africa	1
USA	1
Uruguay	1
England	1
+------------+--------------+
+--------+
|count(1)|
+--------+
| 0|
+--------+
9: Git
6: TODO
4: Run
Terminal
sbt shell
Build
Event Log
Build completed successfully in 13 s 237 ms (13 minutes ago)
376:1 LF UTF-8 2 spaces
Type here to search
ENG
6:18 PM
5/9/2020

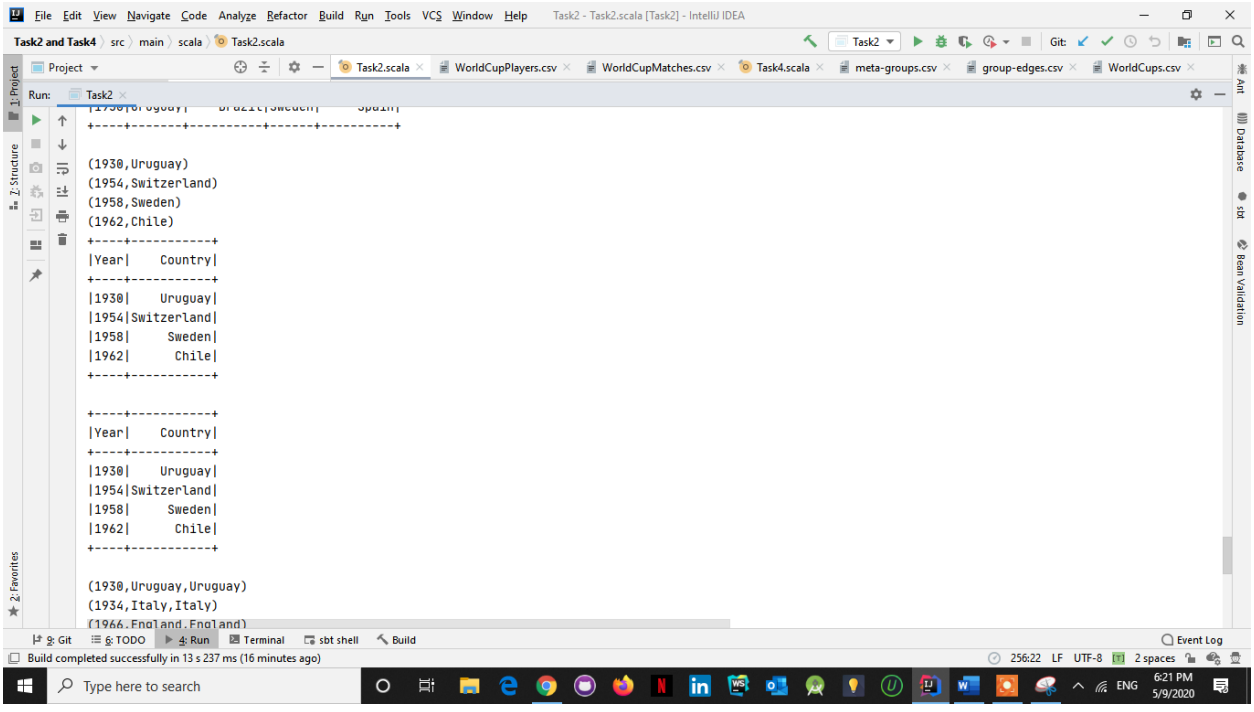

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Task2
Git:
Project
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
Run: Task2
+----+--------------------+
|Year| Player Name|
+----+--------------------+
|1930| Manuel FERREIRA|
|1930| Jose NASAZZI|
|1938| Gyorgy SAROSI|
|1938| Giuseppe MEAZZA|
|1954| Ferenc PUSKAS|
|1954| Fritz WALTER|
|1958| BELLINI|
|1958| Nils LIEDHOLM|
|1962| MAURO RAMOS|
|1962| Ladislav NOVAK|
|1966| Bobby MOORE|
|1966| Uwe SEELER|
|1970| Giacinto FACCHETTI|
|1970| CARLOS ALBERTO|
|1974| Franz BECKENBAUER|
|1974| Johan CRUYFF|
|1978| Ruud KROL|
|1978| Daniel PASSARELLA|
|1982|Karl-Heinz RUMMEN...|
|1986| Diego MARADONA|
+----+--------------------+
only showing top 20 rows
9: Git
6: TODO
4: Run
Terminal
sbt shell
Build
Event Log
Build completed successfully in 13 s 237 ms (13 minutes ago)
376:1 LF UTF-8 2 spaces
Type here to search
ENG
6:19 PM
5/9/2020

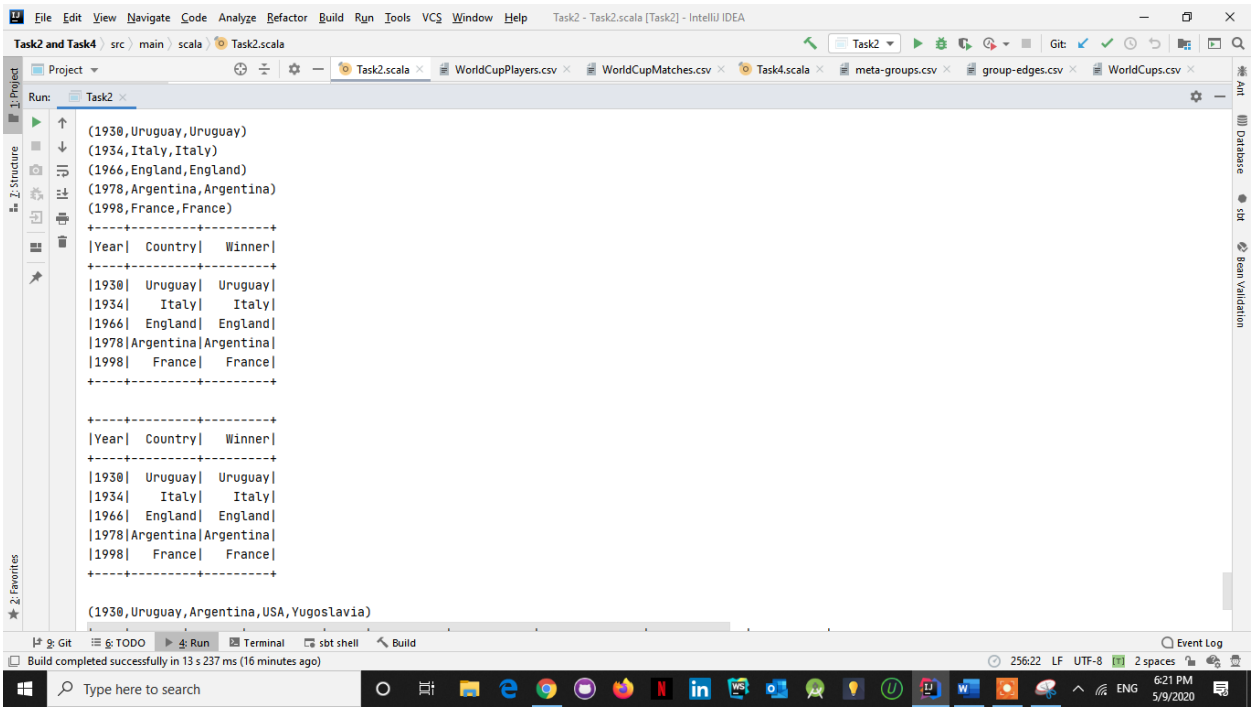

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Task2
Git:
Project
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
Run: Task2
+--------+
|count(1)|
+--------+
| 2|
+--------+
+-------+
|Stadium|
+-------+
| null|
+-------+
+--------------------+
| Player Name|
+--------------------+
| Alex THEPOT|
| Oscar BONFIGLIO|
| Jimmy DOUGLAS|
| Arnold BADJOU|
| Milovan JAKSIC|
| JOEL|
| Ion LAPUSNEANU|
| Juan VALDIVIESO|
| Angel BOSSIO|
| Alex THEPOT|
| Roberto CORTES|
| Isidoro SOTA|
9: Git
6: TODO
4: Run
Terminal
sbt shell
Build
Event Log
Build completed successfully in 13 s 237 ms (14 minutes ago)
256:22 LF UTF-8 2 spaces
Type here to search
ENG
6:19 PM
5/9/2020

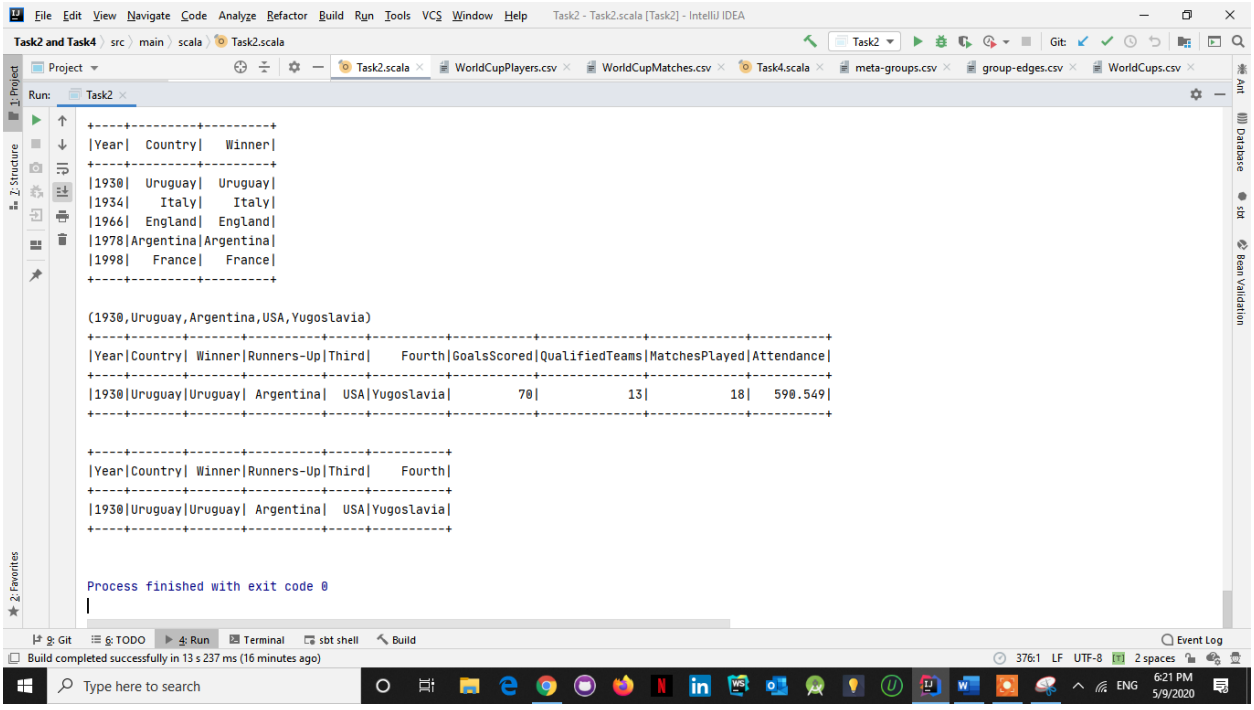

+--------------------+
|Player Name|
+--------------------+
|Alex THEPOT|
|Oscar BONFIGLIO|
|Jimmy DOUGLAS|
|Arnold BADJOU|
|Milovan JAKSIC|
|JOEL|
|Ion LAPUSNEANU|
|Juan VALDIVIESO|
|Angel BOSSIO|
|Alex THEPOT|
|Roberto CORTES|
|Isidoro SOTA|
|Milovan JAKSIC|
|Jesus BERMUDEZ|
|Jimmy DOUGLAS|
|Modesto DENIS|
|Enrique BALLESTRERO|
|Jorge PARDON|
|Roberto CORTES|
|Alex THEPOT|
+--------------------+
only showing top 20 rows
1978,Argentina,Argentina,Netherlands,Brazil,Italy,102,16,38,1.545.791


1998,France,France,Brazil,Croatia,Netherlands,171,32,64,2.785.100
2002,Korea/Japan,Brazil,Germany,Turkey,Korea Republic,161,32,64,2.705.197
2006,Germany,Italy,France,Germany,Portugal,147,32,64,3.359.439
2010,South Africa,Spain,Netherlands,Germany,Uruguay,145,32,64,3.178.856
2014,Brazil,Germany,Argentina,Netherlands,Brazil,171,32,64,3.386.810
1930,Uruguay,Uruguay,Argentina,USA,Yugoslavia,70,13,18,590.549
1934,Italy,Italy,Czechoslovakia,Germany,Austria,70,16,17,363.000
1938,France,Italy,Hungary,Brazil,Sweden,84,15,18,375.700
1950,Brazil,Uruguay,Brazil,Sweden,Spain,88,13,22,1.045.246
1954,Switzerland,Germany FR,Hungary,Austria,Uruguay,140,16,26,768.607
1958,Sweden,Brazil,Sweden,France,Germany FR,126,16,35,819.810
1962,Chile,Brazil,Czechoslovakia,Chile,Yugoslavia,89,16,32,893.172
1966,England,England,Germany FR,Portugal,Soviet Union,89,16,32,1.563.135
1970,Mexico,Brazil,Italy,Germany FR,Uruguay,95,16,32,1.603.975
1974,Germany,Germany FR,Netherlands,Poland,Brazil,97,16,38,1.865.753
(1930,Uruguay,Uruguay)
+----+-------+-------+
|Year|Country| Winner|
+----+-------+-------+
|1930|Uruguay|Uruguay|
+----+-------+-------+
+----+-------+-------+
|Year|Country| Winner|
+----+-------+-------+
|1930|Uruguay|Uruguay|
+----+-------+-------+


File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Task2
Git:
Project
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
Run: Task2
(1930,Uruguay,Argentina,USA,Yugoslavia,Yugoslavia)
(1950,Uruguay,Brazil,Sweden,Spain,Spain)
+----+-------+----------+------+----------+
|Year| Winner|Runners-Up| Third| Fourth|
+----+-------+----------+------+----------+
|1930|Uruguay| Argentina| USA|Yugoslavia|
|1950|Uruguay| Brazil|Sweden| Spain|
+----+-------+----------+------+----------+
+----+-------+----------+------+----------+
|Year| Winner|Runners-Up| Third| Fourth|
+----+-------+----------+------+----------+
|1930|Uruguay| Argentina| USA|Yugoslavia|
|1950|Uruguay| Brazil|Sweden| Spain|
+----+-------+----------+------+----------+
(1930,Uruguay)
(1954,Switzerland)
(1958,Sweden)
(1962,Chile)
+----+-----------+
|Year| Country|
+----+-----------+
|1930| Uruguay|
|1954|Switzerland|
|1958| Sweden|
|1962| Chile|
+----+-----------+
9: Git
6: TODO
4: Run
Terminal
sbt shell
Build
Event Log
Build completed successfully in 13 s 237 ms (15 minutes ago)
256:22 LF UTF-8 2 spaces
Type here to search
ENG
6:20 PM
5/9/2020


File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task2.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task2.scala
Task2
Git:
Project
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
Run: Task2
(1930,Uruguay)
(1954,Switzerland)
(1958,Sweden)
(1962,Chile)
+----+-----------+
|Year| Country|
+----+-----------+
|1930| Uruguay|
|1954|Switzerland|
|1958| Sweden|
|1962| Chile|
+----+-----------+
+----+-----------+
|Year| Country|
+----+-----------+
|1930| Uruguay|
|1954|Switzerland|
|1958| Sweden|
|1962| Chile|
+----+-----------+
(1930,Uruguay,Uruguay)
(1934,Italy,Italy)
9: Git
6: TODO
4: Run
Terminal
sbt shell
Build
Event Log
Build completed successfully in 13 s 237 ms (16 minutes ago)
256:22 LF UTF-8 2 spaces
Type here to search
ENG
6:21 PM
5/9/2020


(1930,Uruguay,Uruguay)
(1934,Italy,Italy)
(1966,England,England)
(1978,Argentina,Argentina)
(1998,France,France)
+----+---------+---------+
|Year| Country| Winner|
+----+---------+---------+
|1930| Uruguay| Uruguay|
|1934| Italy| Italy|
|1966| England| England|
|1978|Argentina|Argentina|
|1998| France| France|
+----+---------+---------+
+----+---------+---------+
|Year| Country| Winner|
+----+---------+---------+
|1930| Uruguay| Uruguay|
|1934| Italy| Italy|
|1966| England| England|
|1978|Argentina|Argentina|
|1998| France| France|
+----+---------+---------+
(1930,Uruguay,Argentina,USA,Yugoslavia)


+----+---------+---------+
|Year| Country| Winner|
+----+---------+---------+
|1930| Uruguay| Uruguay|
|1934| Italy| Italy|
|1966| England| England|
|1978|Argentina|Argentina|
|1998| France| France|
+----+---------+---------+
(1930,Uruguay,Argentina,USA,Yugoslavia)
+----+-------+-------+----------+-----+----------+-----------+---------------+-------------+----------+
|Year|Country| Winner|Runners-Up|Third| Fourth|GoalsScored|QualifiedTeams|MatchesPlayed|Attendance|
+----+-------+-------+----------+-----+----------+-----------+---------------+-------------+----------+
|1930|Uruguay|Uruguay| Argentina| USA|Yugoslavia| 70| 13| 18| 590.549|
+----+-------+-------+----------+-----+----------+-----------+---------------+-------------+----------+
+----+-------+-------+----------+-----+----------+
|Year|Country| Winner|Runners-Up|Third| Fourth|
+----+-------+-------+----------+-----+----------+
|1930|Uruguay|Uruguay| Argentina| USA|Yugoslavia|
+----+-------+-------+----------+-----+----------+
Process finished with exit code 0

Task 3

a. Perform Word-Count on Twitter Streaming Data using Spark.

Code:

Twitter.py | TwitterStreamming.py

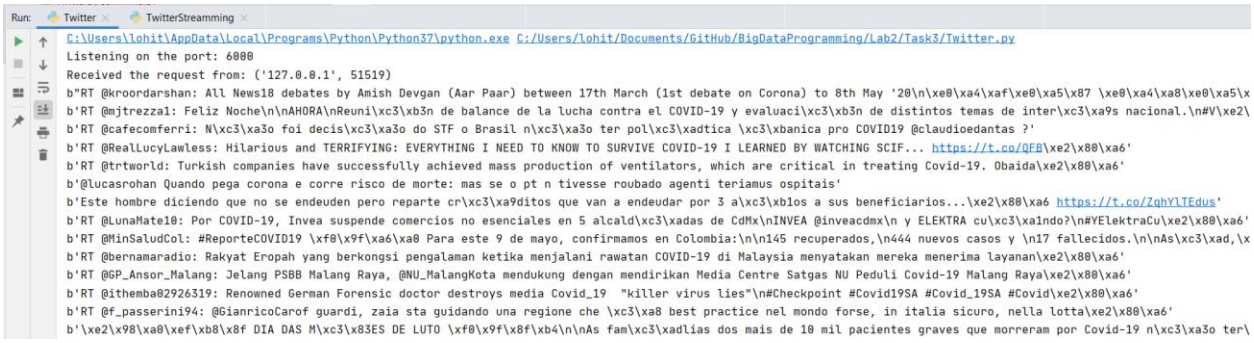
```
from builtins import BaseException, print, str

from tweepy import OAuthHandler
from tweepy import Stream
from tweepy.streaming import StreamListener
import socket
import json
import time

# authorization tokens
access_token = "1045756888510541824-1zbS5Pi4hL6pYUwHuO41pwNcK4CNvm"
access_token_secret = "yer1Vmp2te61Wat52olEpmptjYqLumSZ3ry9XpyKTJ7oH"
consumer_key = "WAlQxKnCMdIFRsx1NyS7bdAEV"
consumer_secret = "cJNYNLYHhb0McLJaQ3sN9uWiAR6DmFDhzmN6CXNUCCIPOezM18"

auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

class StreamListener(StreamListener):

    def __init__(self, socket):
        self.socket = socket

```

```
    def on_data(self, data):
        try:
            msg = json.loads(data)
            print(msg['text'].encode('utf-8'))
            self.socket.send(msg['text'].encode('utf-8'))
            return True
        except BaseException as e:
            print("Encountered error in on_data: %s" % str(e))
        return True

    def on_error(self, status_code):
        print("Encountered streaming error (", status_code, ")")
        return True

def readData(socket):
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, StreamListener(socket))
    tags = ["layoffs","covid","corona","wfh"]
    stream.filter(track=tags)

if __name__ == "__main__":
    s = socket.socket()
    host = "localhost"
    port = 6000
    s.bind((host, port))
    print("Listening on the port: %s" % str(port))
    s.listen(5)
    socket, addr = s.accept()
    print("Received the request from: " + str(addr))
    time.sleep(5)
    readData(socket)
```

Twitter.py | TwitterStreamming.py

```python
import os

from pyspark import SparkContext
from pyspark.streaming import StreamingContext

from collections import namedtuple

os.environ["SPARK_HOME"] = "C:\Spark"

def main():
    sc = SparkContext(appName="WordCount")
    ssc = StreamingContext(sc, 5)
    lines = ssc.socketTextStream("localhost", 6000)
    fields = ("word", "count")
    Tweet = namedtuple('Text', fields)

    counts = lines.flatMap(lambda text: text.split(" "))\
        .map(lambda word: (word, 1))\
        .reduceByKey(lambda x, y: x + y).map(lambda rec: Tweet(rec[0], rec[1]))
    counts.pprint()
    ssc.start()
    ssc.awaitTermination()

if __name__ == "__main__":
    main()
```

Output:



```
Run:  Twitter  TwitterStreamming
05/09 17:55:37 WARN BlockManager: Block input-0-1589064937200 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
e: 2020-05-09 17:55:10
-------------------------------------------
t(word='March', count=1)
t(word='न्यूज़', count=1)
t(word='अ…RT', count=1)
t(word='', count=15)
t(word='el', count=4)
t(word='y', count=5)
t(word='Não', count=1)
t(word='foi', count=1)
t(word='Brasil', count=1)
t(word='ter', count=1)
```

Task 4

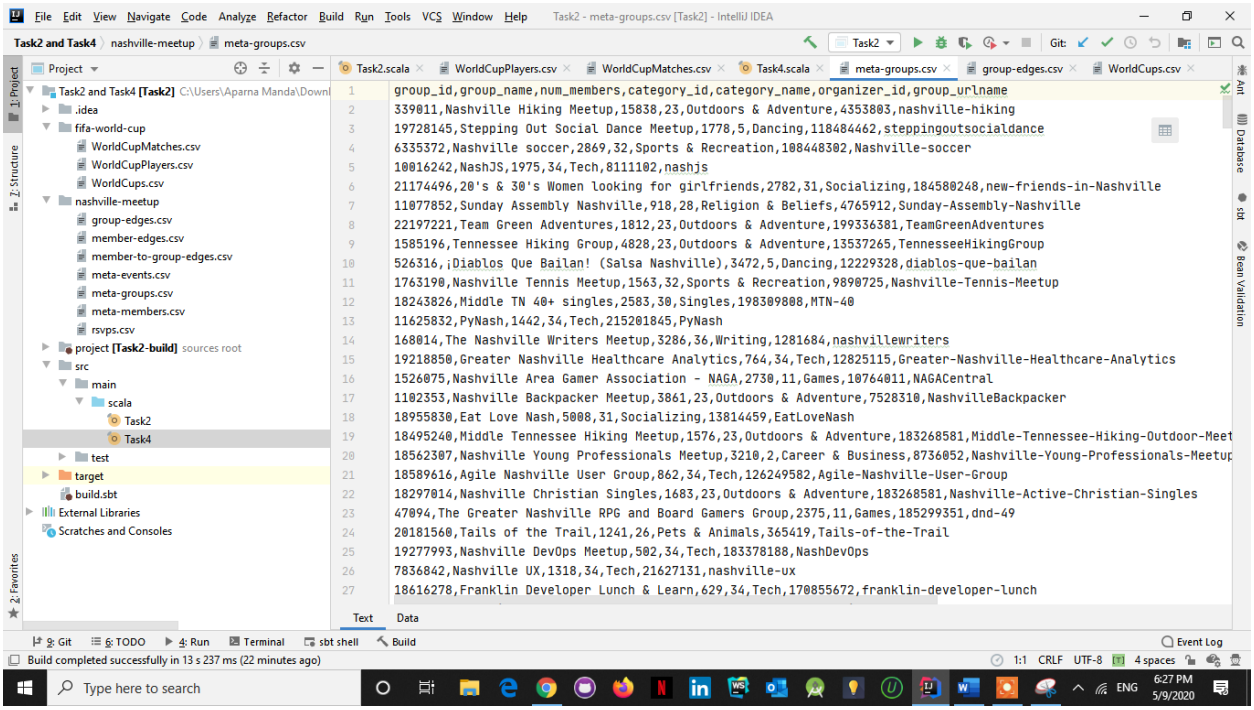
a. Perform Page Rank

b. State importance of using graphx on the chosen dataset.

Input

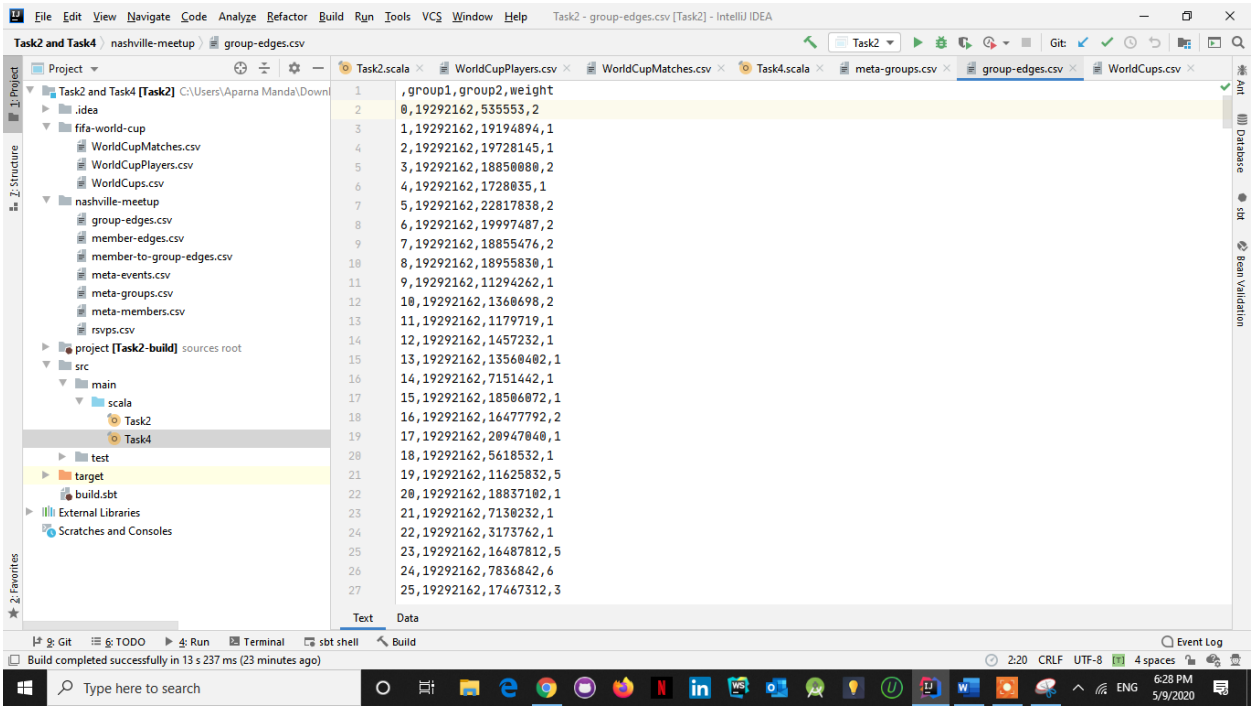
Meta-groups.csv as vertices View

Vertices of graph are 100 records of Meta-groups.csv

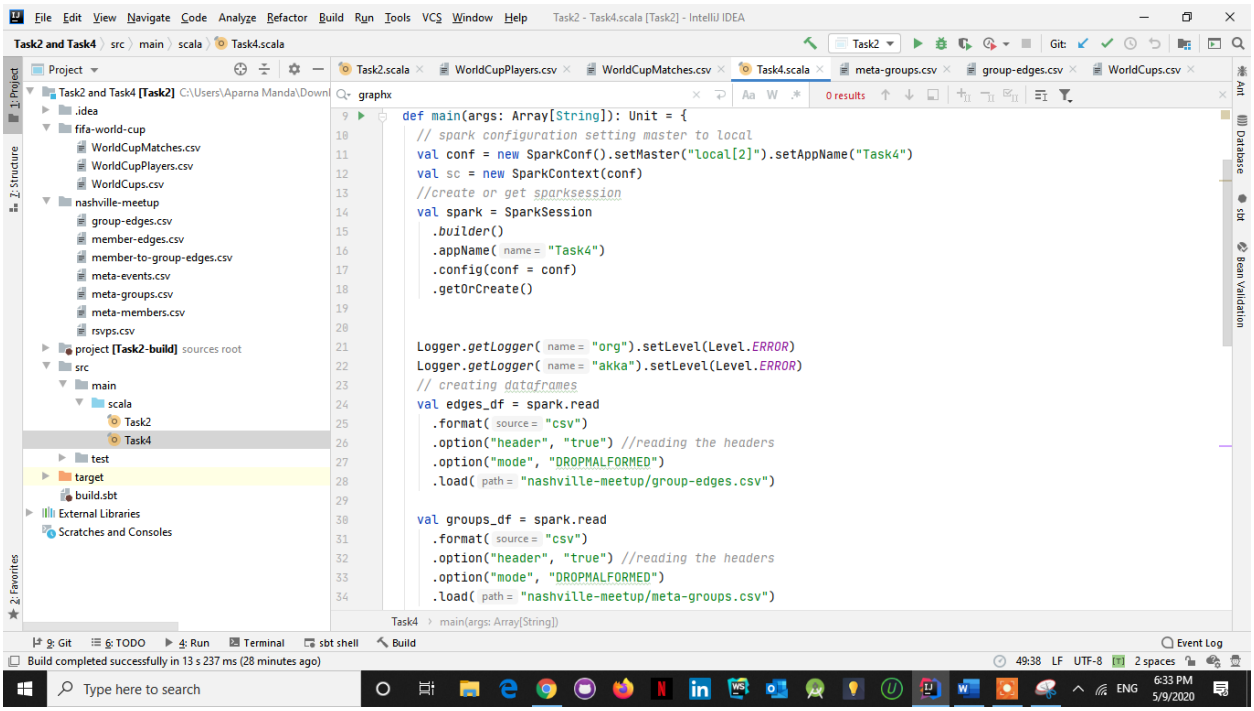


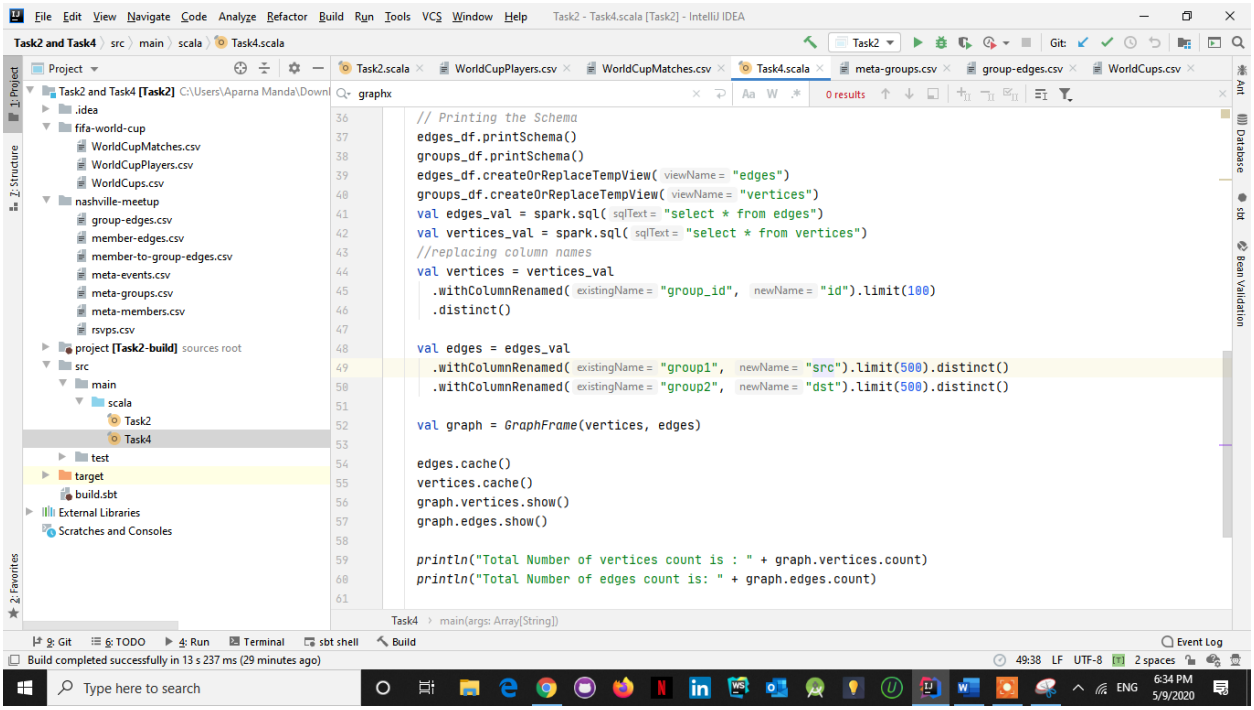
Group-edges.csv as edges View

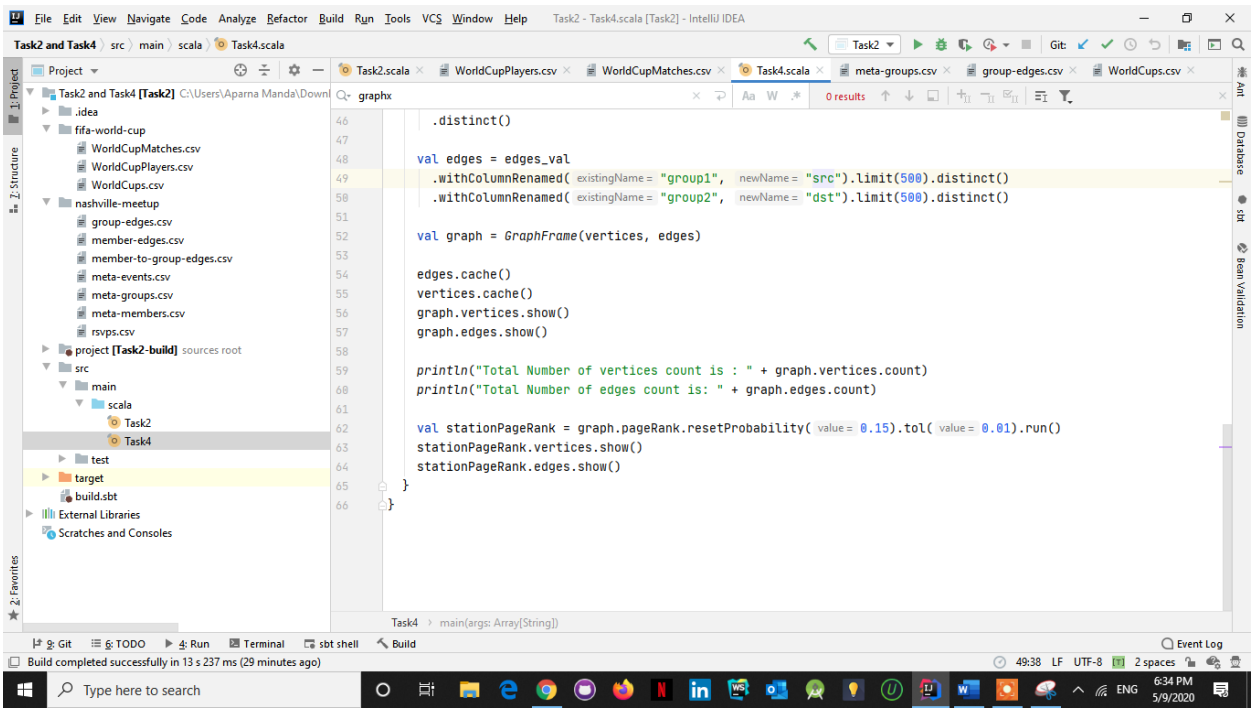
Edges of graph are 500 records of Group-edges.csv



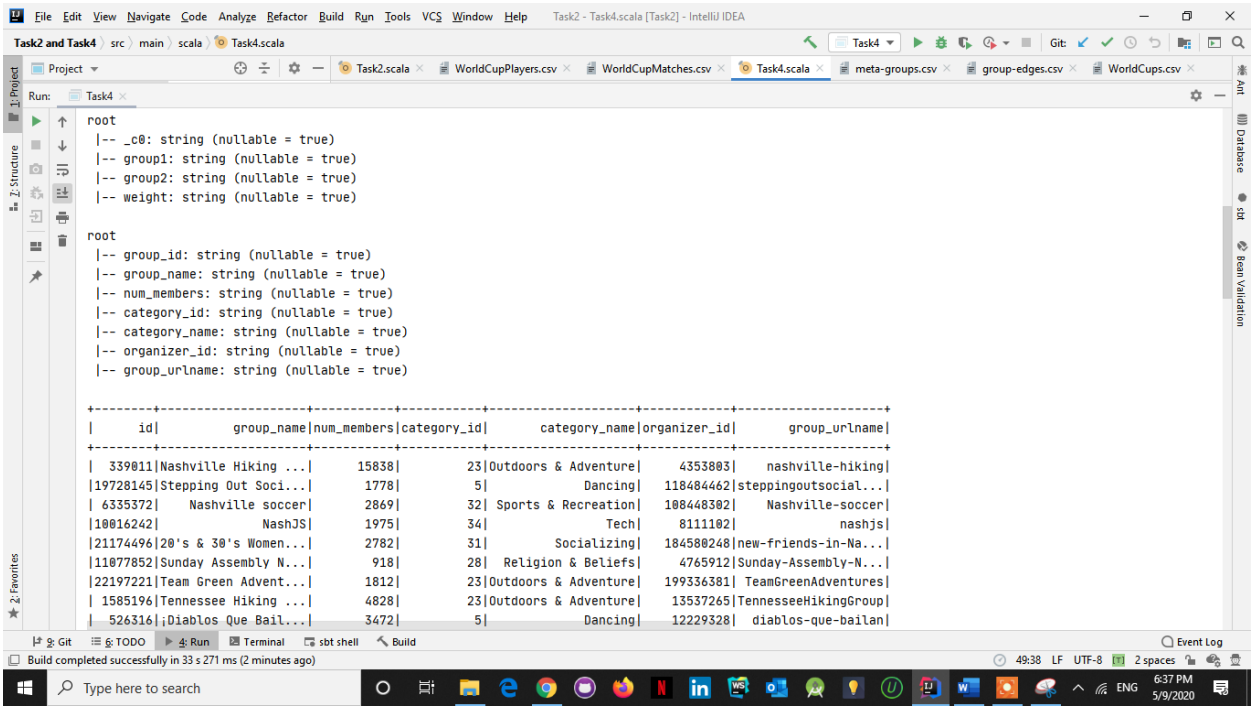
Code



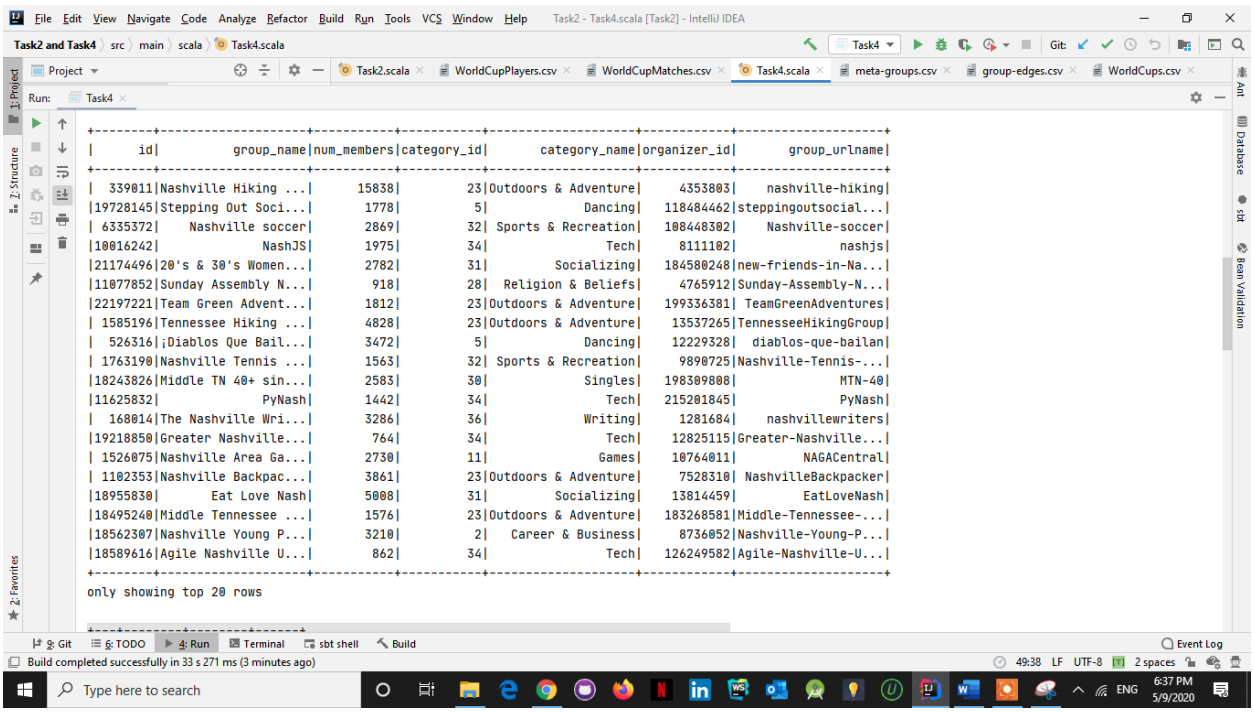

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task4.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task4.scala
Task2
Git:
Project
Task2 and Task4 [Task2] C:\Users\Aparna Manda\Down
.idea
fifa-world-cup
WorldCupMatches.csv
WorldCupPlayers.csv
WorldCups.csv
nashville-meetup
group-edges.csv
member-edges.csv
member-to-group-edges.csv
meta-events.csv
meta-groups.csv
meta-members.csv
rsvps.csv
project [Task2-build] sources root
src
main
scala
Task2
Task4
test
target
build.sbt
External Libraries
Scratches and Consoles
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
graphx
0 results
// Printing the Schema
edges_df.printSchema()
groups_df.printSchema()
edges_df.createOrReplaceTempView(viewName = "edges")
groups_df.createOrReplaceTempView(viewName = "vertices")
val edges_val = spark.sql(sqlText = "select * from edges")
val vertices_val = spark.sql(sqlText = "select * from vertices")
//replacing column names
val vertices = vertices_val
.withColumnRenamed(existingName = "group_id", newName = "id").limit(100)
.distinct()
val edges = edges_val
.withColumnRenamed(existingName = "group1", newName = "src").limit(500).distinct()
.withColumnRenamed(existingName = "group2", newName = "dst").limit(500).distinct()
val graph = GraphFrame(vertices, edges)
edges.cache()
vertices.cache()
graph.vertices.show()
graph.edges.show()
println("Total Number of vertices count is : " + graph.vertices.count)
println("Total Number of edges count is: " + graph.edges.count)
Task4 main(args: Array[String])
9: Git
6: TODO
4: Run
Terminal
sbt shell
Build
Event Log
Build completed successfully in 13 s 237 ms (29 minutes ago)
49:38 LF UTF-8 2 spaces
Type here to search
ENG
6:34 PM
5/9/2020


File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task4.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task4.scala
Task2
Git:
Project
Task2 and Task4 [Task2] C:\Users\Aparna Manda\Down
.idea
fifa-world-cup
WorldCupMatches.csv
WorldCupPlayers.csv
WorldCups.csv
nashville-meetup
group-edges.csv
member-edges.csv
member-to-group-edges.csv
meta-events.csv
meta-groups.csv
meta-members.csv
rsvps.csv
project [Task2-build] sources root
src
main
scala
Task2
Task4
test
target
build.sbt
External Libraries
Scratches and Consoles
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
graphx
0 results
.distinct()
val edges = edges_val
.withColumnRenamed(existingName = "group1", newName = "src").limit(500).distinct()
.withColumnRenamed(existingName = "group2", newName = "dst").limit(500).distinct()
val graph = GraphFrame(vertices, edges)
edges.cache()
vertices.cache()
graph.vertices.show()
graph.edges.show()
println("Total Number of vertices count is : " + graph.vertices.count)
println("Total Number of edges count is: " + graph.edges.count)
val stationPageRank = graph.pageRank.resetProbability(value = 0.15).tol(value = 0.01).run()
stationPageRank.vertices.show()
stationPageRank.edges.show()
}
}
Task4 main(args: Array[String])
9: Git
6: TODO
4: Run
Terminal
sbt shell
Build
Event Log
Build completed successfully in 13 s 237 ms (29 minutes ago)
49:38 LF UTF-8 2 spaces
Type here to search
ENG
6:34 PM
5/9/2020

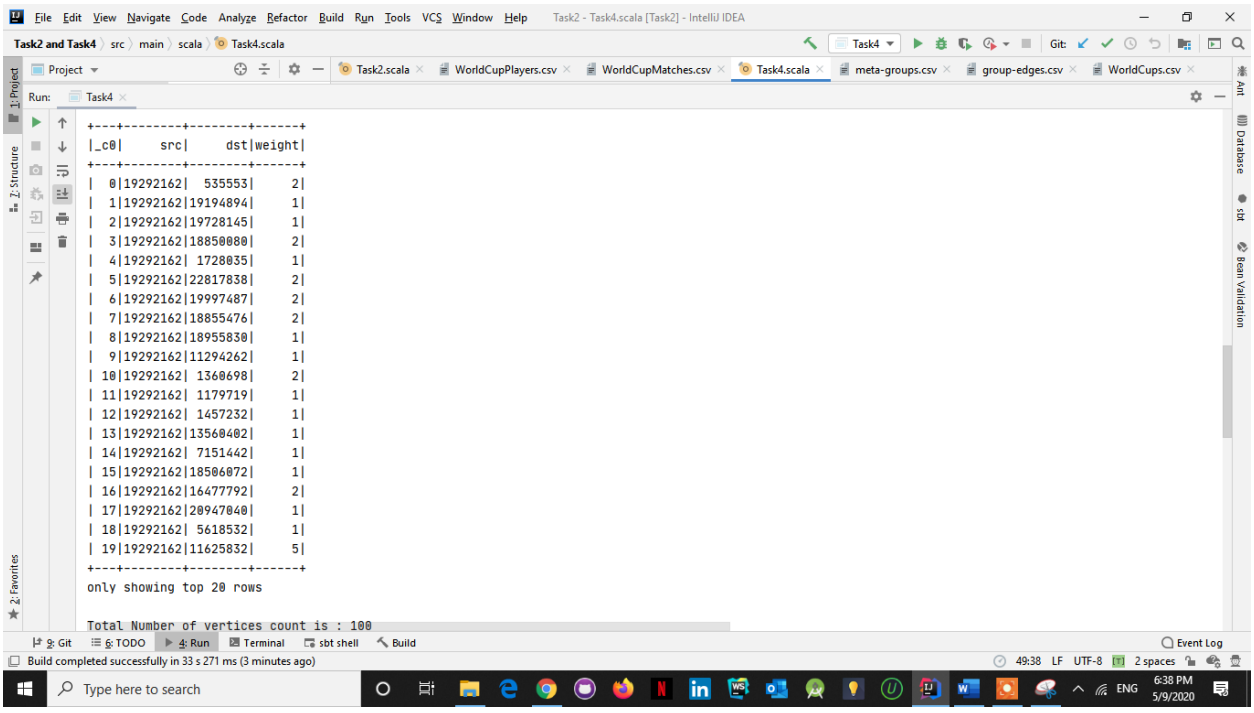
Output

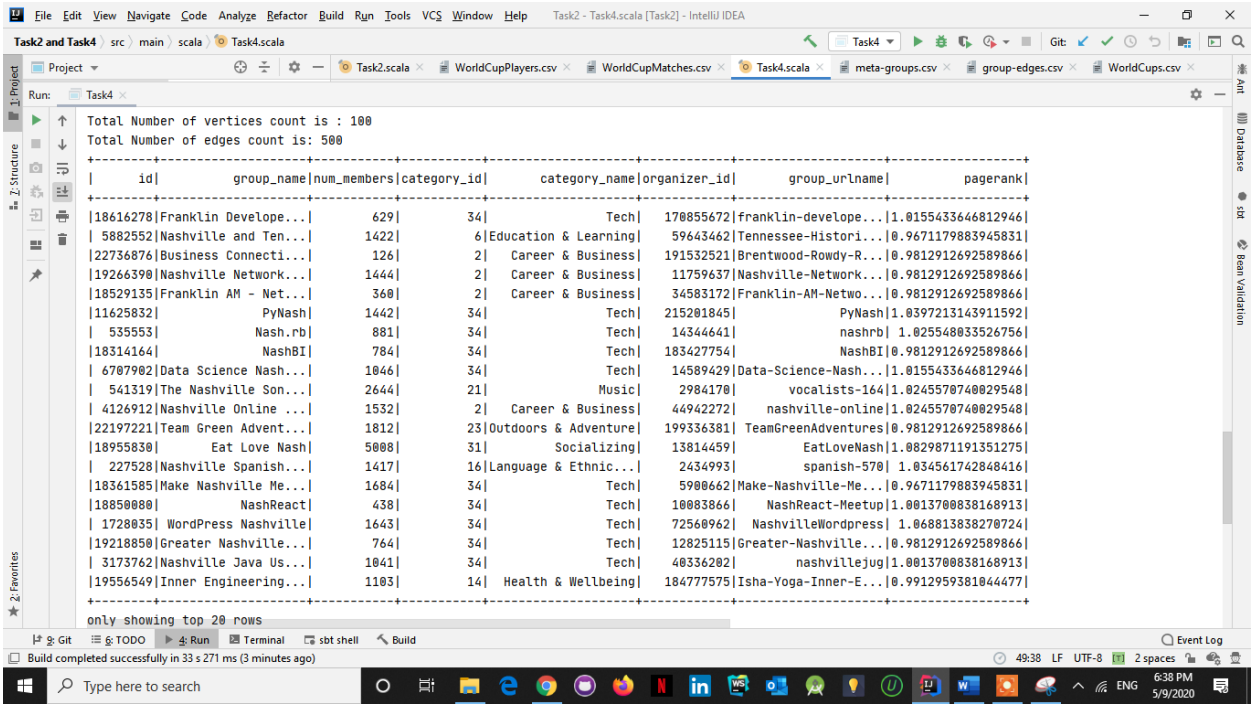


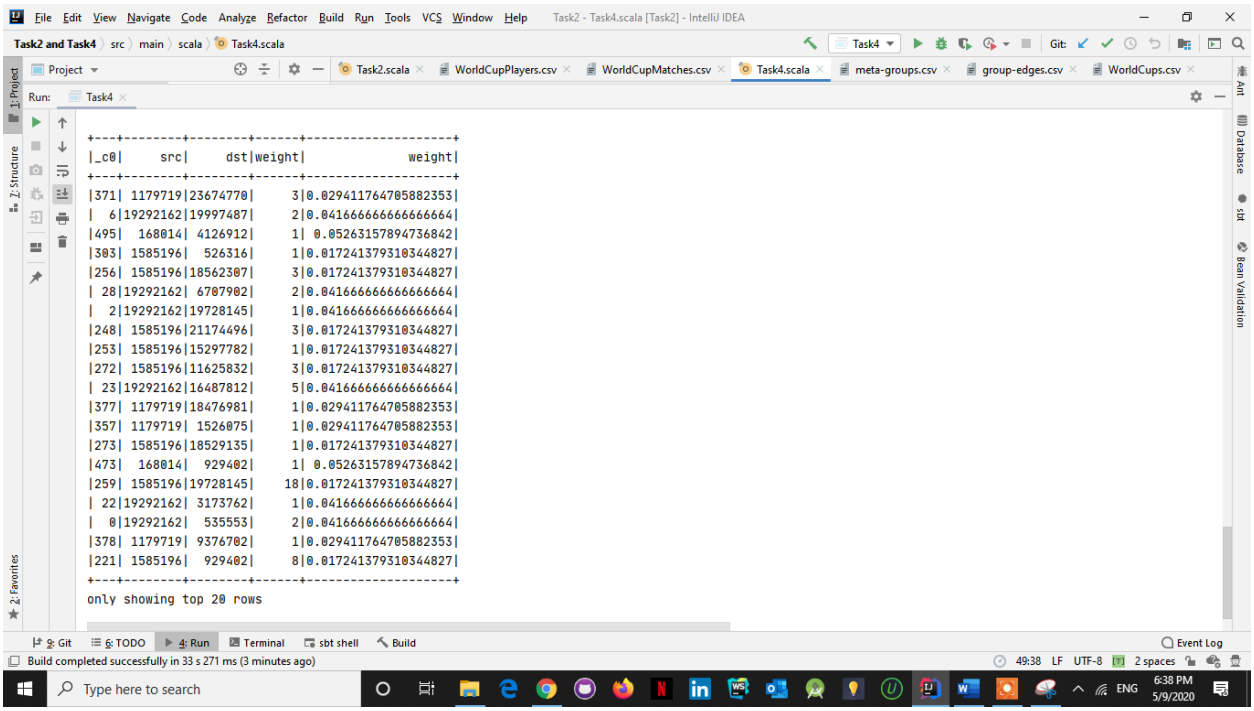
```
root
 |-- _c0: string (nullable = true)
 |-- group1: string (nullable = true)
 |-- group2: string (nullable = true)
 |-- weight: string (nullable = true)

root
 |-- group_id: string (nullable = true)
 |-- group_name: string (nullable = true)
 |-- num_members: string (nullable = true)
 |-- category_id: string (nullable = true)
 |-- category_name: string (nullable = true)
 |-- organizer_id: string (nullable = true)
 |-- group_urlname: string (nullable = true)

+--------+--------------------+-----------+-----------+--------------------+------------+--------------------+
|      id|          group_name|num_members|category_id|       category_name|organizer_id|       group_urlname|
+--------+--------------------+-----------+-----------+--------------------+------------+--------------------+
|  339011|Nashville Hiking ...|      15838|         23|Outdoors & Adventure|     4353803|    nashville-hiking|
|19728145|Stepping Out Soci...|       1778|          5|             Dancing|   118484462|steppingoutsocial...|
| 6335372|    Nashville soccer|       2869|         32| Sports & Recreation|   108448302|    Nashville-soccer|
|10016242|              NashJS|       1975|         34|                Tech|     8111102|              nashjs|
|21174496|20's & 30's Women...|       2782|         31|         Socializing|   184580248|new-friends-in-Na...|
|11077852|Sunday Assembly N...|        918|         28|  Religion & Beliefs|     4765912|Sunday-Assembly-N...|
|22197221|Team Green Advent...|       1812|         23|Outdoors & Adventure|   199336381| TeamGreenAdventures|
| 1585196|Tennessee Hiking ...|       4828|         23|Outdoors & Adventure|    13537265|TennesseeHikingGroup|
|  526316|¡Diablos Que Bail...|       3472|          5|             Dancing|    12229328|  diablos-que-bailan|
```



```
+--------+--------------------+-----------+-----------+--------------------+------------+--------------------+
|      id|          group_name|num_members|category_id|       category_name|organizer_id|       group_urlname|
+--------+--------------------+-----------+-----------+--------------------+------------+--------------------+
|  339011|Nashville Hiking ...|      15838|         23|Outdoors & Adventure|     4353803|    nashville-hiking|
|19728145|Stepping Out Soci...|       1778|          5|             Dancing|   118484462|steppingoutsocial...|
| 6335372|    Nashville soccer|       2869|         32| Sports & Recreation|   108448302|    Nashville-soccer|
|10016242|              NashJS|       1975|         34|                Tech|     8111102|              nashjs|
|21174496|20's & 30's Women...|       2782|         31|         Socializing|   184580248|new-friends-in-Na...|
|11077852|Sunday Assembly N...|        918|         28|  Religion & Beliefs|     4765912|Sunday-Assembly-N...|
|22197221|Team Green Advent...|       1812|         23|Outdoors & Adventure|   199336381| TeamGreenAdventures|
| 1585196|Tennessee Hiking ...|       4828|         23|Outdoors & Adventure|    13537265|TennesseeHikingGroup|
|  526316|¡Diablos Que Bail...|       3472|          5|             Dancing|    12229328|  diablos-que-bailan|
| 1763190|Nashville Tennis ...|       1563|         32| Sports & Recreation|     9890725|Nashville-Tennis-...|
|18243826|Middle TN 40+ sin...|       2583|         30|             Singles|   198309888|              MTN-40|
|11625832|              PyNash|       1442|         34|                Tech|   215201845|              PyNash|
|  168014|The Nashville Wri...|       3286|         36|             Writing|     1281684|    nashvillewriters|
|19218850|Greater Nashville...|        764|         34|                Tech|    12825115|Greater-Nashville...|
| 1526075|Nashville Area Ga...|       2730|         11|               Games|    10764011|         NAGACentral|
| 1102353|Nashville Backpac...|       3861|         23|Outdoors & Adventure|     7528310| NashvilleBackpacker|
|18955830|       Eat Love Nash|       5008|         31|         Socializing|    13814459|         EatLoveNash|
|18495240|Middle Tennessee ...|       1576|         23|Outdoors & Adventure|   183268581|Middle-Tennessee-...|
|18562307|Nashville Young P...|       3210|          2|   Career & Business|     8736052|Nashville-Young-P...|
|18589616|Agile Nashville U...|        862|         34|                Tech|   126249582|Agile-Nashville-U...|
+--------+--------------------+-----------+-----------+--------------------+------------+--------------------+
only showing top 20 rows
```


File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task4.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task4.scala
Task4
Git:
Project
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
Run: Task4
+---+--------+--------+------+
|_c0| src| dst|weight|
+---+--------+--------+------+
0	19292162	535553	2
1	19292162	19194894	1
2	19292162	19728145	1
3	19292162	18850080	2
4	19292162	1728035	1
5	19292162	22817838	2
6	19292162	19997487	2
7	19292162	18855476	2
8	19292162	18955830	1
9	19292162	11294262	1
10	19292162	1360698	2
11	19292162	1179719	1
12	19292162	1457232	1
13	19292162	13560402	1
14	19292162	7151442	1
15	19292162	18506072	1
16	19292162	16477792	2
17	19292162	20947840	1
18	19292162	5618532	1
19	19292162	11625832	5
+---+--------+--------+------+
only showing top 20 rows
Total Number of vertices count is : 100
9: Git 6: TODO 4: Run Terminal sbt shell Build
Event Log
Build completed successfully in 33 s 271 ms (3 minutes ago)
49:38 LF UTF-8 2 spaces
Type here to search
ENG
6:38 PM
5/9/2020


File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Task2 - Task4.scala [Task2] - IntelliJ IDEA
Task2 and Task4 src main scala Task4.scala
Task4
Git:
Project
Task2.scala
WorldCupPlayers.csv
WorldCupMatches.csv
Task4.scala
meta-groups.csv
group-edges.csv
WorldCups.csv
Run: Task4
Total Number of vertices count is : 100
Total Number of edges count is: 500
+--------+--------------------+----------+-----------+--------------------+-----------+--------------------+--------------------+
| id| group_name|num_members|category_id| category_name|organizer_id| group_urlname| pagerank|
+--------+--------------------+----------+-----------+--------------------+-----------+--------------------+--------------------+
18616278	Franklin Develope...	629	34	Tech	170855672	franklin-develope...	1.0155433646812946
5882552	Nashville and Ten...	1422	6	Education & Learning	59643462	Tennessee-Histori...	0.9671179883945831
22736876	Business Connecti...	126	2	Career & Business	191532521	Brentwood-Rowdy-R...	0.9812912692589866
19266390	Nashville Network...	1444	2	Career & Business	11759637	Nashville-Network...	0.9812912692589866
18529135	Franklin AM - Net...	360	2	Career & Business	34583172	Franklin-AM-Netwo...	0.9812912692589866
11625832	PyNash	1442	34	Tech	215201845	PyNash	1.0397213143911592
535553	Nash.rb	881	34	Tech	14344641	nashrb	1.025548033526756
18314164	NashBI	784	34	Tech	183427754	NashBI	0.9812912692589866
6707902	Data Science Nash...	1046	34	Tech	14589429	Data-Science-Nash...	1.0155433646812946
541319	The Nashville Son...	2644	21	Music	2984170	vocalists-164	1.0245570740029548
4126912	Nashville Online ...	1532	2	Career & Business	44942272	nashville-online	1.0245570740029548
22197221	Team Green Advent...	1812	23	Outdoors & Adventure	199336381	TeamGreenAdventures	0.9812912692589866
18955830	Eat Love Nash	5008	31	Socializing	13814459	EatLoveNash	1.0829871191351275
227528	Nashville Spanish...	1417	16	Language & Ethnic...	2434993	spanish-570	1.034561742848416
18361585	Make Nashville Me...	1684	34	Tech	5900662	Make-Nashville-Me...	0.9671179883945831
18850080	NashReact	438	34	Tech	10083866	NashReact-Meetup	1.0013700838168913
1728035	WordPress Nashville	1643	34	Tech	72560962	NashvilleWordpress	1.068813838270724
19218850	Greater Nashville...	764	34	Tech	12825115	Greater-Nashville...	0.9812912692589866
3173762	Nashville Java Us...	1041	34	Tech	40336202	nashvillejug	1.0013700838168913
19556549	Inner Engineering...	1103	14	Health & Wellbeing	184777575	Isha-Yoga-Inner-E...	0.9912959381044477
+--------+--------------------+----------+-----------+--------------------+-----------+--------------------+--------------------+
only showing top 20 rows
9: Git 6: TODO 4: Run Terminal sbt shell Build
Event Log
Build completed successfully in 33 s 271 ms (3 minutes ago)
49:38 LF UTF-8 2 spaces
Type here to search
ENG
6:38 PM
5/9/2020



b. Importance of GraphX

The importance of a group can be measured the weights of the groups it is connected to and thus GraphX would be the best-fit for the dataset and Page Rank algorithm would be the perfect algorithm to determine the importance of the groups.