Qinyun Lin – SEC01 (NUID 001582464)

Big Data System Engineering with Scala
Spring 2022
Assignment No. 1

NORTHEASTERN UNIVERSITY

18 98

LVX
VERITAS
VIRTVS

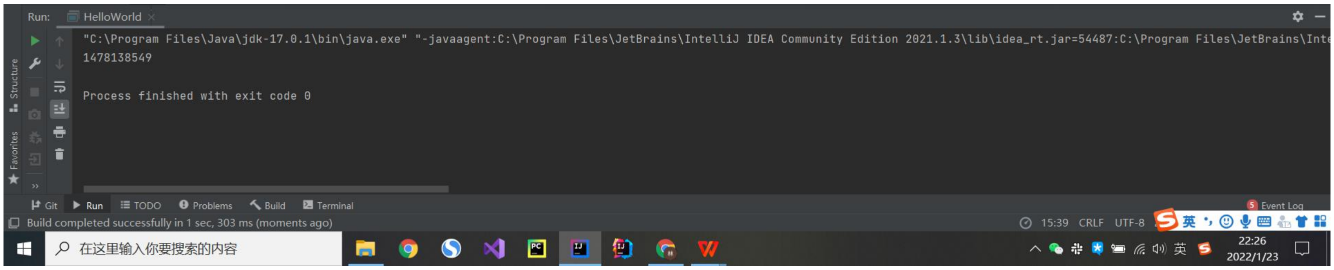
Task

Ensure a development environment can compile, test and run Scala programs

1. Check the result of "scala.util.Random.nextInt"
2. Run "HelloWord"
3. Run "Ingest.scala" and view the list of movies
4. Modify "Ingest.scala" and find movies of "New Zealand"
5. Get "Spark ready" and familiar with interacting with spark program

Solution

1. Check the result of "scala.util.Random.nextInt"

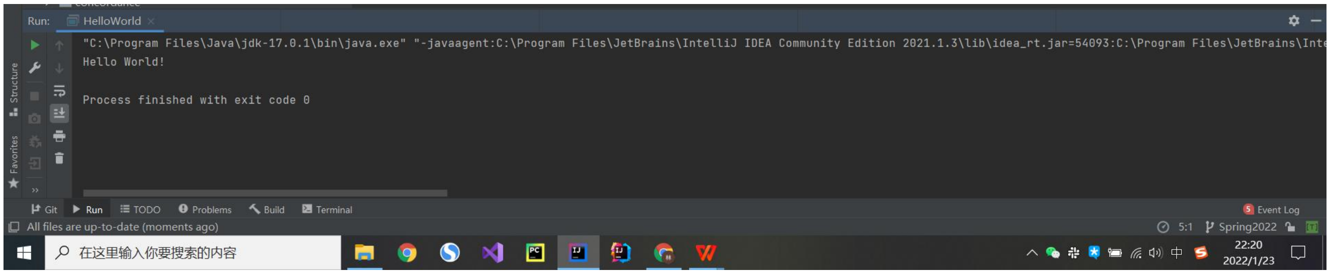


```
"C:\Program Files\Java\jdk-17.0.1\bin\java.exe" "-javaagent:C:\Program Files\JetBrains\IntelliJ IDEA Community Edition 2021.1.3\lib\idea_rt.jar=54487:C:\Program Files\JetBrains\Inte
1478138549

Process finished with exit code 0
```

Build completed successfully in 1 sec, 303 ms (moments ago)

2. Run "HelloWord"

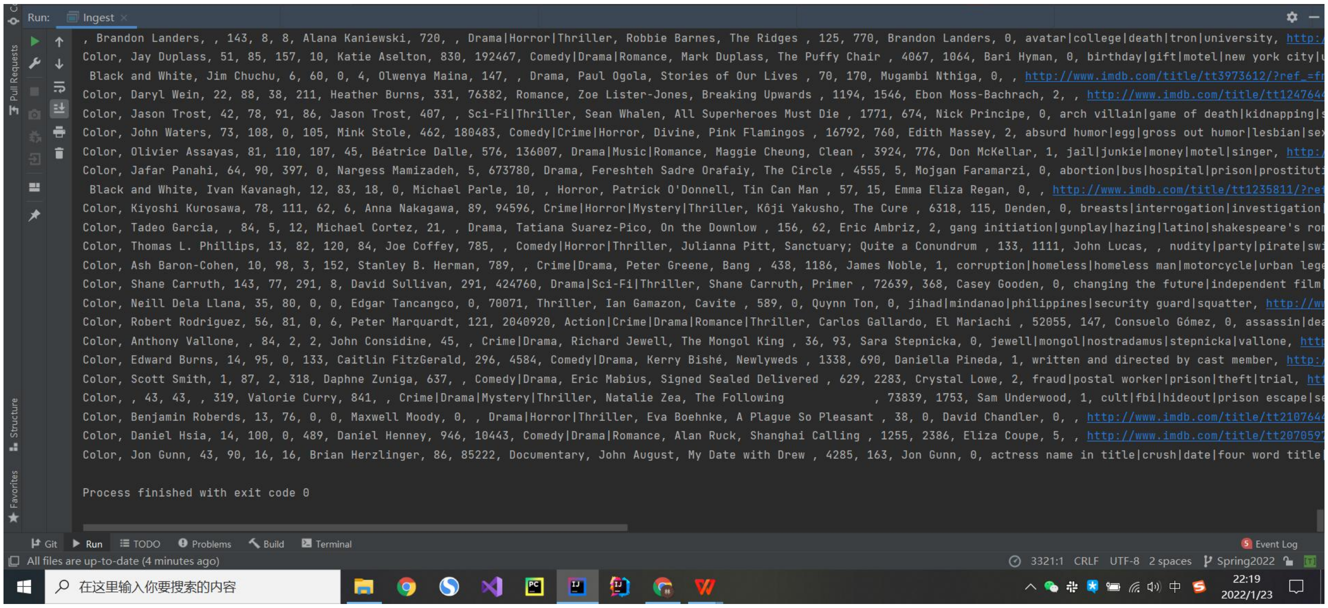


```
"C:\Program Files\Java\jdk-17.0.1\bin\java.exe" "-javaagent:C:\Program Files\JetBrains\IntelliJ IDEA Community Edition 2021.1.3\lib\idea_rt.jar=54093:C:\Program Files\JetBrains\Inte
Hello World!

Process finished with exit code 0
```

All files are up-to-date (moments ago)

3. Run "Ingest.scala" and view the list of movies

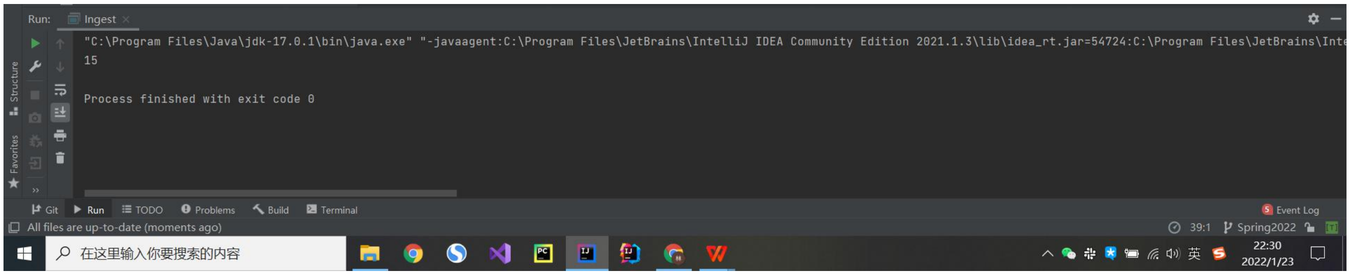


```
, Brandon Landers, , 143, 8, 8, Alana Kaniewski, 720, , Drama|Horror|Thriller, Robbie Barnes, The Ridges , 125, 770, Brandon Landers, 0, avatar|college|death|tron|university, http:/
Color, Jay Duplass, 51, 85, 157, 10, Katie Aselton, 830, 192467, Comedy|Drama|Romance, Mark Duplass, The Puffy Chair , 4067, 1064, Bari Hyman, 0, birthday|gift|motel|new york city|
 Black and White, Jim Chuchu, 6, 60, 0, 4, Olwenya Maina, 147, , Drama, Paul Ogola, Stories of Our Lives , 70, 170, Mugambi Nthiga, 0, , http://www.imdb.com/title/tt3973612/?ref_=f
Color, Daryl Wein, 22, 88, 38, 211, Heather Burns, 331, 76382, Romance, Zoe Lister-Jones, Breaking Upwards , 1194, 1546, Ebon Moss-Bachrach, 2, , http://www.imdb.com/title/tt124764
Color, Jason Trost, 42, 78, 91, 86, Jason Trost, 407, , Sci-Fi|Thriller, Sean Whalen, All Superheroes Must Die , 1771, 674, Nick Principe, 0, arch villain|game of death|kidnapping|
Color, John Waters, 73, 108, 0, 105, Mink Stole, 462, 180483, Comedy|Crime|Horror, Divine, Pink Flamingos , 16792, 760, Edith Massey, 2, absurd humor|egg|gross out humor|lesbian|sex
Color, Olivier Assayas, 81, 110, 107, 45, Béatrice Dalle, 576, 136007, Drama|Music|Romance, Maggie Cheung, Clean , 3924, 776, Don McKellar, 1, jail|junkie|money|motel|singer, http:/
Color, Jafar Panahi, 64, 90, 397, 0, Nargess Mamizadeh, 5, 673780, Drama, Fereshteh Sadre Orafaiy, The Circle , 4555, 5, Mojgan Faramarzi, 0, abortion|bus|hospital|prison|prostitut
 Black and White, Ivan Kavanagh, 12, 83, 18, 0, Michael Parle, 10, , Horror, Patrick O'Donnell, Tin Can Man , 57, 15, Emma Eliza Regan, 0, , http://www.imdb.com/title/tt1235811/?ref
Color, Kiyoshi Kurosawa, 78, 111, 62, 6, Anna Nakagawa, 89, 94596, Crime|Horror|Mystery|Thriller, Kôji Yakusho, The Cure , 6318, 115, Denden, 0, breasts|interrogation|investigation|
Color, Tadeo Garcia, , 84, 5, 12, Michael Cortez, 21, , Drama, Tatiana Suarez-Pico, On the Downlow , 156, 62, Eric Ambriz, 2, gang initiation|gunplay|hazing|latino|shakespeare's ro
Color, Thomas L. Phillips, 13, 82, 120, 84, Joe Coffey, 785, , Comedy|Horror|Thriller, Julianna Pitt, Sanctuary; Quite a Conundrum , 133, 1111, John Lucas, , nudity|party|pirate|sw
Color, Ash Baron-Cohen, 10, 98, 3, 152, Stanley B. Herman, 789, , Crime|Drama, Peter Greene, Bang , 438, 1186, James Noble, 1, corruption|homeless|homeless man|motorcycle|urban leg
Color, Shane Carruth, 143, 77, 291, 8, David Sullivan, 291, 424760, Drama|Sci-Fi|Thriller, Shane Carruth, Primer , 72639, 368, Casey Gooden, 0, changing the future|independent film|
Color, Neill Dela Llana, 35, 80, 0, 0, Edgar Tancangco, 0, 70071, Thriller, Ian Gamazon, Cavite , 589, 0, Quynn Ton, 0, jihad|mindanao|philippines|security guard|squatter, http://ww
Color, Robert Rodriguez, 56, 81, 0, 6, Peter Marquardt, 121, 2040920, Action|Crime|Drama|Romance|Thriller, Carlos Gallardo, El Mariachi , 52055, 147, Consuelo Gómez, 0, assassin|dea
Color, Anthony Vallone, , 84, 2, 2, John Considine, 45, , Crime|Drama, Richard Jewell, The Mongol King , 36, 93, Sara Stepnicka, 0, jewell|mongol|nostradamus|stepnicka|vallone, http
Color, Edward Burns, 14, 95, 0, 133, Caitlin FitzGerald, 296, 4584, Comedy|Drama, Kerry Bishé, Newlyweds , 1338, 690, Daniella Pineda, 1, written and directed by cast member, http:/
Color, Scott Smith, 1, 87, 2, 318, Daphne Zuniga, 637, , Comedy|Drama, Eric Mabius, Signed Sealed Delivered , 629, 2283, Crystal Lowe, 2, fraud|postal worker|prison|theft|trial, htt
Color, , 43, 43, , 319, Valorie Curry, 841, , Crime|Drama|Mystery|Thriller, Natalie Zea, The Following             , 73839, 1753, Sam Underwood, 1, cult|fbi|hideout|prison escape|se
Color, Benjamin Roberds, 13, 76, 0, 0, Maxwell Moody, 0, , Drama|Horror|Thriller, Eva Boehnke, A Plague So Pleasant , 38, 0, David Chandler, 0, , http://www.imdb.com/title/tt210764
Color, Daniel Hsia, 14, 100, 0, 489, Daniel Henney, 946, 10443, Comedy|Drama|Romance, Alan Ruck, Shanghai Calling , 1255, 2386, Eliza Coupe, 5, , http://www.imdb.com/title/tt207059
Color, Jon Gunn, 43, 90, 16, 16, Brian Herzlinger, 86, 85222, Documentary, John August, My Date with Drew , 4285, 163, Jon Gunn, 0, actress name in title|crush|date|four word title|

Process finished with exit code 0
```

All files are up-to-date (4 minutes ago)

4. Modify “Ingest.scala” and find number of movies of “New Zealand”



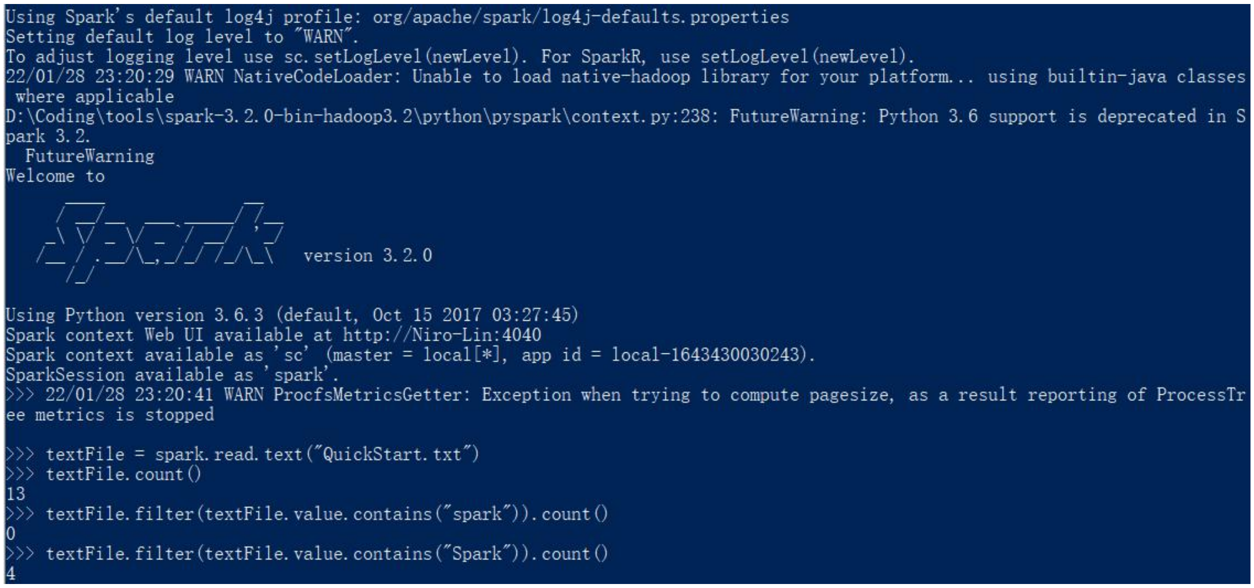
```
"C:\Program Files\Java\jdk-17.0.1\bin\java.exe" "-javaagent:C:\Program Files\JetBrains\IntelliJ IDEA Community Edition 2021.1.3\lib\idea_rt.jar=54724:C:\Program Files\JetBrains\Inte
15

Process finished with exit code 0
```

All files are up-to-date (moments ago)

There are 15 movies are from “New Zealand” in the list.

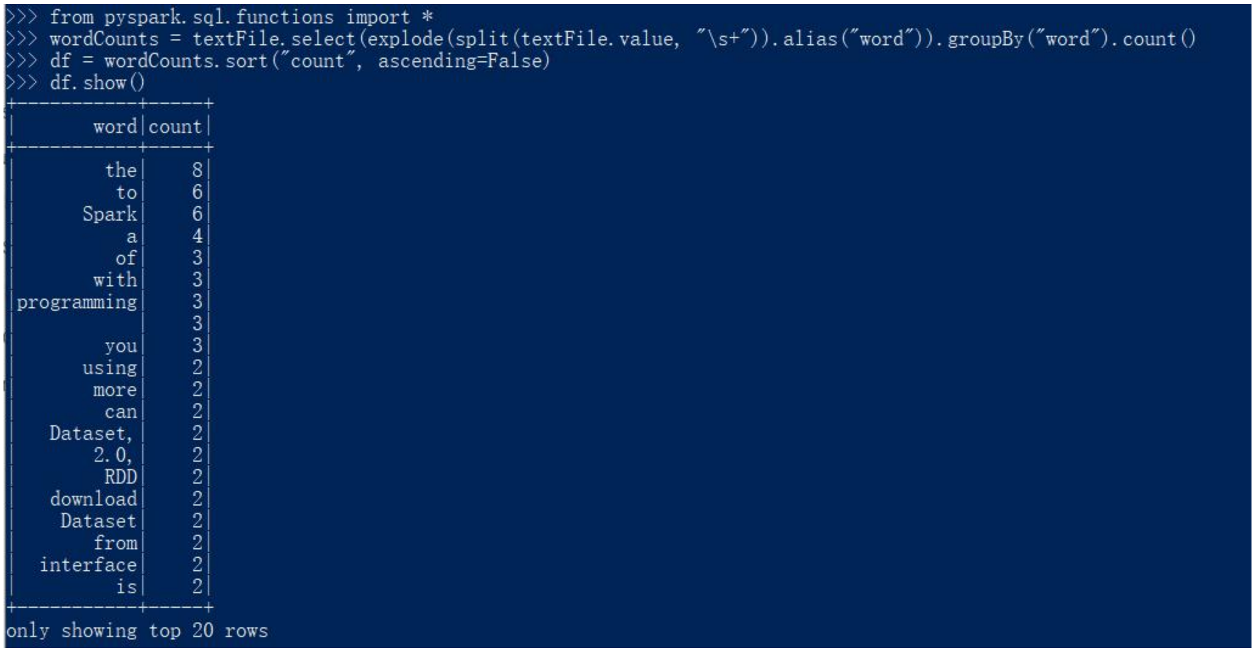
5. Run simple spark program(word count) in Shell



```
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/01/28 23:20:29 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
 where applicable
D:\Coding\tools\spark-3.2.0-bin-hadoop3.2\python\pyspark\context.py:238: FutureWarning: Python 3.6 support is deprecated in S
park 3.2.
  FutureWarning
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.2.0
      /_/

Using Python version 3.6.3 (default, Oct 15 2017 03:27:45)
Spark context Web UI available at http://Niro-Lin:4040
Spark context available as 'sc' (master = local[*], app id = local-1643430030243).
SparkSession available as 'spark'.
>>> 22/01/28 23:20:41 WARN ProcfsMetricsGetter: Exception when trying to compute pagesize, as a result reporting of ProcessTr
ee metrics is stopped

>>> textFile = spark.read.text("QuickStart.txt")
>>> textFile.count()
13
>>> textFile.filter(textFile.value.contains("spark")).count()
0
>>> textFile.filter(textFile.value.contains("Spark")).count()
4
```



```
>>> from pyspark.sql.functions import *
>>> wordCounts = textFile.select(explode(split(textFile.value, "\s+")).alias("word")).groupBy("word").count()
>>> df = wordCounts.sort("count", ascending=False)
>>> df.show()
+-----------+-----+
|       word|count|
+-----------+-----+
|        the|    8|
|         to|    6|
|      Spark|    6|
|          a|    4|
|         of|    3|
|       with|    3|
|programming|    3|
|           |    3|
|        you|    3|
|      using|    2|
|       more|    2|
|        can|    2|
|   Dataset,|    2|
|       2.0,|    2|
|        RDD|    2|
|   download|    2|
|    Dataset|    2|
|       from|    2|
|  interface|    2|
|         is|    2|
+-----------+-----+
only showing top 20 rows
```

Project Source

https://github.com/MrNiro/CSYE7200/tree/Spring2022/assignment-helloworld/src/main/scala/edu/neu/coe/csye7200/assthw