Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Інститут прикладного системного аналізу

Кафедра системного проектування

Оброблення надвеликих масивів даних

Лабораторна робота №2

Виконала:

студентка групи ДА-11мп

Молчанова В.С.

Київ – 2021 р.

## Завдання:

Pеалізація та розуміння алгоритму “inverted index”

## Хід роботи:

### Текст програми:

Файл BigData.scala

import org.apache.spark.{SparkConf, SparkContext}  
  
object BigData {  
 def main(args: Array[String]) = {  
 val (lab, output) = *parseParameters*(args)  
 val conf = new SparkConf().setAppName("appName").setMaster("local")  
 val sc = new SparkContext(conf)  
 val result =  
 if (lab == "2" ) *getInvertedIndices*(sc, "D:/Programming/spark-3.2.0-bin-hadoop3.2/input/lab2/")  
 else *getPageRank*(sc)  
 sc.stop()  
 *writeToFile*(result, output)  
 }  
  
 def parseParameters(args: Array[String]): (String, String) = {  
 val lab = if (args.length > 0) args(0) else "2"  
 val output = if (args.length > 1) args(1) else "output.txt"  
 (lab, output)  
 }  
  
 def writeToFile(result: String, filePath: String): Unit ={  
 val pw = new PrintWriter(new File(filePath ))  
 pw.write(result)  
 pw.close()  
 }  
}

Файл InvertedIndex.scala

import org.apache.spark.SparkContext  
object InvertedIndex {  
 def getInvertedIndices(sc: SparkContext, inputFolder: String): String = {  
 val files = sc.wholeTextFiles(inputFolder+"\*")  
 val wordsInFiles = files.map(p => (p.\_1, p.\_2.split("\\W+"))).flatMap(*getWordsIndices*)  
 .reduceByKey((a, b) => s"**$**a, **$**b")  
 val res = wordsInFiles.collect()  
 res.map(p => "%1$-10s \t%2$s".format(p.\_1, *formatFileName*(p.\_2, inputFolder)))  
 .mkString("Result:\n", "\n", "")  
 }  
  
 private def getWordsIndices(pair: (String, Array[String])): Array[(String, String)] = {  
 val (file, words) = pair  
 words.zipWithIndex.map(p => (p.\_1, s"**$**file@**$**{p.\_2}"))  
 }  
  
 private def formatFileName(file: String, inputFolder: String): String = {  
 file.replace(inputFolder, "").replace("file:/", "")  
 }  
}

Файл build\_and\_run.bat

@ECHO OFF  
CALL D:  
CALL cd D:\Programming\Scala\SparkBigData  
CALL sbt package  
set outputPath=.\output.txt  
CALL D:\Programming\spark-3.2.0-bin-hadoop3.2\bin\spark-submit --class "BigData" --master local[4] D:\Programming\Scala\SparkBigData\target\scala-2.12\spark-big-data\_2.12-0.1.jar %1 %outputPath%  
CALL %outputPath%  
PAUSE

### Вхідні дані

Файл test1.txt

hello my

dear friend

this file has some words in it

and some of this words are in another file too

Файл test2.txt

and hello again

here are some words that are in other file

### Результат виконання

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated