



Министерство науки и высшего образования Российской Федерации
Калужский филиал
федерального государственного бюджетного
образовательного учреждения высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(КФ МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИУК «Информатика и управление»

КАФЕДРА ИУК4 «Программное обеспечение ЭВМ, информационные технологии»

ДОМАШНЯЯ РАБОТА

«Основы Spark. Установка Spark. Основные команды для работы с RDD»

ДИСЦИПЛИНА: «Технологии обработки больших данных»

Выполнил: студент гр. ИУК4-72Б _____ (Карельский М.К.)
(Подпись)

Проверил: _____ (Голубева С.Е.)
(Подпись)

Дата сдачи (защиты):

Результаты сдачи (защиты):

- Балльная оценка:

- Оценка:

Калуга, 2023

Цель: формирование практических навыков работы с платформой Apache Spark для обработки больших данных.

Задачи:

1. Изучить основы Apache Spark.
2. Научиться устанавливать и конфигурировать Spark.
3. Уметь работать с RDD.
4. Получить навыки написания программ для обработки больших данных.

Задание:

Написать скрипт для платформы Apache Spark для решения задачи, указанной в варианте. В качестве входных текстовых файлов можно использовать книги в txt формате из библиотеки Project Gutenberg: <https://www.gutenberg.org>.

Вариант 7

Для двух текстовых файлов подсчитать количество слов, которые встречаются одновременно и в первом, и во втором файле. Результат сохранить в файл в виде пар ключ-значение, где ключ – количество общих слов, значение – само слово.

Листинг:

HW.py

```
from pyspark.sql import SparkSession
import shutil

spark = SparkSession.builder.appName("WordCount").getOrCreate()

text_file_1 = spark.sparkContext.textFile("pg71725.txt")
text_file_2 = spark.sparkContext.textFile("pg71729.txt")

counts_1 = text_file_1.flatMap(lambda line: line.strip().lower().split(" "))\
    .map(lambda word: (word, 1))\
    .reduceByKey(lambda a, b: a + b)
counts_2 = text_file_2.flatMap(lambda line: line.strip().lower().split(" "))\
    .map(lambda word: (word, 1))\
    .reduceByKey(lambda a, b: a + b)

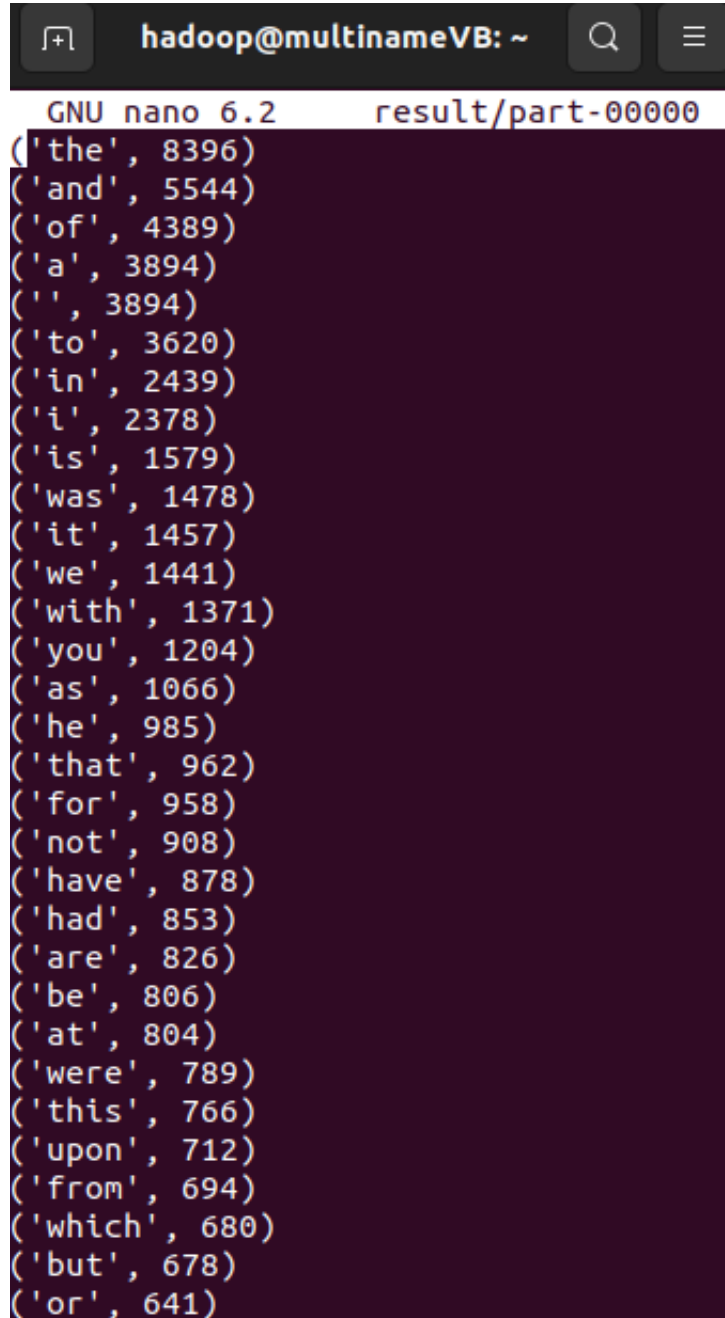
joined = counts_1.join(counts_2)\
    .map(lambda val: (val[0], val[1][0] + val[1][1]))\
    .sortBy(lambda val: val[1], False)

shutil.rmtree("./result")
joined.saveAsTextFile("./result")
spark.stop()
```

Результат:

```
hadoop@multinameVB: $ spark-submit HW.py
23/10/14 14:16:57 WARN Utils: Your hostname, multinameVB resolves to a loopback address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
23/10/14 14:16:57 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
23/10/14 14:17:00 INFO SparkContext: Running Spark version 3.5.0
23/10/14 14:17:00 INFO SparkContext: OS info Linux, 6.2.0-32-generic, amd64
23/10/14 14:17:00 INFO SparkContext: Java version 1.8.0_382
23/10/14 14:17:00 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/10/14 14:17:00 INFO ResourceUtils: =====
23/10/14 14:17:00 INFO ResourceUtils: No custom resources configured for spark.driver.
23/10/14 14:17:00 INFO ResourceUtils: =====
23/10/14 14:17:00 INFO SparkContext: Submitted application: WordCount
```

Рис. 1. Запуск программы



```
hadoop@multinameVB: ~
GNU nano 6.2 result/part-00000
('the', 8396)
('and', 5544)
('of', 4389)
('a', 3894)
('', 3894)
('to', 3620)
('in', 2439)
('i', 2378)
('is', 1579)
('was', 1478)
('it', 1457)
('we', 1441)
('with', 1371)
('you', 1204)
('as', 1066)
('he', 985)
('that', 962)
('for', 958)
('not', 908)
('have', 878)
('had', 853)
('are', 826)
('be', 806)
('at', 804)
('were', 789)
('this', 766)
('upon', 712)
('from', 694)
('which', 680)
('but', 678)
('or', 641)
```

Рис. 2. Результат

Вывод: в ходе выполнения домашней работы были получены практические навыки работы с платформой Apache Spark для обработки больших данных.

ОСНОВНАЯ ЛИТЕРАТУРА

1. Федин Ф.О. Анализ данных. Часть 1. Подготовка данных к анализу [Электронный ресурс] : учебное пособие / Ф.О. Федин, Ф.Ф. Федин. — Электрон. текстовые данные. — М. : Московский городской педагогический университет, 2012. — 204 с. — 2227-8397. — Режим доступа: <http://www.iprbookshop.ru/26444.html>
2. Федин Ф.О. Анализ данных. Часть 2. Инструменты Data Mining [Электронный ресурс] : учебное пособие / Ф.О. Федин, Ф.Ф. Федин. — Электрон. текстовые данные. — М. : Московский городской педагогический университет, 2012. — 308 с. — 2227-8397. — Режим доступа: <http://www.iprbookshop.ru/26445.html>
3. Чубукова, И.А. Data Mining [Электронный ресурс] : учеб. Пособие — Электрон. дан. — Москва : , 2016. — 470 с. — Режим доступа: <https://e.lanbook.com/book/100582>. — Загл. с экрана.
4. Воронова Л.И. Big Data. Методы и средства анализа [Электронный ресурс] : учебное пособие / Л.И. Воронова, В.И. Воронов. — Электрон. текстовые данные. — М. : Московский технический университет связи и информатики, 2016. — 33 с. — 2227-8397. — Режим доступа: <http://www.iprbookshop.ru/61463.html>

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА

5. Волкова Т.В. Разработка систем распределенной обработки данных [Электронный ресурс] : учебно-методическое пособие / Т.В. Волкова, Л.Ф. Насейкина. — Электрон. текстовые данные. — Оренбург: Оренбургский государственный университет, ЭБС АСВ, 2012. — 330 с. — 2227-8397. — Режим доступа: <http://www.iprbookshop.ru/30127.html>
6. Кухаренко Б.Г. Интеллектуальные системы и технологии [Электронный ресурс] : учебное пособие / Б.Г. Кухаренко. — Электрон. текстовые данные. — М. : Московская государственная академия водного транспорта, 2015. — 116 с. — 2227-8397. — Режим доступа: <http://www.iprbookshop.ru/47933.html>
7. Воронова Л.И. Интеллектуальные базы данных [Электронный ресурс] : учебное пособие / Л.И. Воронова. — Электрон. Текстовые данные. — М. : Московский технический университет связи и информатики, 2013. — 35 с. — 2227-8397. — Режим доступа: <http://www.iprbookshop.ru/63324.html>
8. Николаев Е.И. Базы данных в высокопроизводительных информационных системах [Электронный ресурс] : учебное пособие / Е.И. Николаев. — Электрон. текстовые данные. — Ставрополь: Северо-Кавказский федеральный университет, 2016. — 163 с. — 2227-8397. — Режим доступа: <http://www.iprbookshop.ru/69375.html>

Электронные ресурсы:

9. <http://hadoop.apache.org/> (англ.)
10. <https://spark.apache.org/> (англ.)