Session 17

Assignment 1

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Session 17 – Assignment 1 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Session 17 – Assignment 1** |
|  |  |
| **Start Date:** | 16/10/2017 |
|  |  |
|  |  |



# 

# Contents

[Contents 2](#_Toc495936927)

[Change History 3](#_Toc495936928)

[1. Problem Statement 4](#_Toc495936929)

[2. Sample document : 4](#_Toc495936930)

[3. Solution 5](#_Toc495936931)

[4. Result 6](#_Toc495936932)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 16/10/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Problem Statement

* Write a program to read a text file and print the number of rows of data in the document.
* Write a program to read a text file and print the number of words in the document.
* We have a document where the word separator is -, instead of space. Write a spark code, to obtain the count of the total number of words present in the document.

# Sample document :

This-is-my-first-assignment.

It-will-count-the-number-of-lines-in-this-document.

The-total-number-of-lines-is-3

# Solution

To enhance my learning I have written a solution that answers all questions in one using the suggested dataset (**Mad or what**).

**Code created**

*package com.duncb.spark*

*import org.apache.spark.\_*

*import org.apache.spark.SparkContext.\_*

*import org.apache.log4j.\_*

*/\*\* Count up how many of each word appears in a book as simply as possible. \*/*

*object WCount {*

*/\*\* Our main function where the action happens \*/*

*def main(args: Array[String]) {*

*// Set the log level to only print errors*

*Logger.getLogger("org").setLevel(Level.ERROR)*

*// Create a SparkContext using every core of the local machine*

*val sc = new SparkContext("local[\*]", "WordCount")*

*// Read each line of my file into an RDD*

*val input = sc.textFile("file:///N:/Datasets/spark1.txt")*

*// Split into words separated by an \_ character*

*val words = input.flatMap(x => x.split("-"))*

*words.foreach(println)*

*val numberOfLInes= input.count*

*val numberofWords = words.count*

*println("Number of lines = " + numberOfLInes)*

*println ("Number of Words = " + numberofWords)*

*// Count up the occurrences of each word*

*val wordCounts = words.countByValue()*

*// Print the results.*

*wordCounts.foreach(println)*

*}*

*}*

# Result

*Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties*

*The*

*total*

*number*

*of*

*lines*

*is*

*3*

*This*

*is*

*my*

*first*

*assignment*

*It*

*will*

*count*

*the*

*number*

*of*

*lines*

*in*

*this*

*document*

*Number of lines = 3*

*Number of Words = 22*

*(number,2)*

*(this,1)*

*(count,1)*

*(in,1)*

*(is,2)*

*(This,1)*

*(document,1)*

*(total,1)*

*(The,1)*

*(my,1)*

*(will,1)*

*(first,1)*

*(It,1)*

*(assignment,1)*

*(lines,2)*

*(3,1)*

*(of,2)*

*(the,1)*