Task 1

1.Write a program to read a text file and print the number of rows of data in the document.

2. Write a program to read a text file and print the number of words in the document.

3. 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.

```
package spark.basic.cl18
import org.apache.spark.sql.SparkSession


 object task1 extends App {

   val sparkSession = SparkSession.builder.master("local")
     .appName("spark").getOrCreate()

   val sparkcontetxt = sparkSession.sparkContext

   val fc = sparkcontetxt.textFile("/home/acadgild/Downloads/19_Dataset.txt")

   val rowcount = fc.count()

   val words = fc.flatMap(x => x.split(","))

   val wc = fc.flatMap(x => x.split("-"))
   println("Row Count = " + rowcount)
   println("Word Count = " + words.count())
   println("WC with sep - = " + wc.count())

 }
```

Output:

```
18/07/29 22:08:40 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0)
18/07/29 22:08:40 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks ha
18/07/29 22:08:40 INFO DAGScheduler: ResultStage 0 (count at task1.scala:14)
18/07/29 22:08:40 INFO DAGScheduler: Job 0 finished: count at task1.scala:14,
Row Count = 22
18/07/29 22:08:40 INFO SparkContext: Starting job: count at task1.scala:20
18/07/29 22:08:40 INFO DAGScheduler: Got job 1 (count at task1.scala:20) with
18/07/29 22:08:40 INFO DAGScheduler: Final stage: ResultStage 1 (count at tas
18/07/29 22:08:40 INFO DAGScheduler: Parents of final stage: List()
18/07/29 22:08:40 INFO DAGScheduler: Missing parents: List()
```

```
18/07/29 22:08:40 INFO TaskSetManager: Finished task 0.0
18/07/29 22:08:40 INFO DAGScheduler: ResultStage 1 (cour
18/07/29 22:08:40 INFO DAGScheduler: Job 1 finished: cou
Word Count = 110
18/07/29 22:08:40 INFO TaskSchedulerImpl: Removed TaskSe
18/07/29 22:08:40 INFO SparkContext: Starting job: count
18/07/29 22:08:40 INFO DAGScheduler: Got job 2 (count at
18/07/29 22:08:40 INFO DAGScheduler: Final stage: Result
18/07/29 22:08:40 INFO DAGScheduler: Parents of final st
```

```
18/07/29 22:08:40 INFO LineRecordReader: Found UTF-8 BOM and skipped
18/07/29 22:08:40 INFO Executor: Finished task 0.0 in stage 2.0 (TID
18/07/29 22:08:40 INFO TaskSetManager: Finished task 0.0 in stage 2.0
18/07/29 22:08:40 INFO DAGScheduler: ResultStage 2 (count at task1.sc
18/07/29 22:08:40 INFO DAGScheduler: Job 2 finished: count at task1.s
WC with sep - = 44
18/07/29 22:08:40 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose
18/07/29 22:08:40 INFO SparkContext: Invoking stop() from shutdown ho
18/07/29 22:08:40 INFO BlockManagerInfo: Removed broadcast_3_piece0 o
18/07/29 22:08:40 INFO SparkUI: Stopped Spark web UI at http://192.16
18/07/29 22:08:40 INFO MapOutputTrackerMasterEndpoint: MapOutputTrack
18/07/29 22:08:40 INFO MemoryStore: MemoryStore cleared
```

Task 2

Problem Statement 1:

1. Read the text file, and create a tupled rdd.

2. Find the count of total number of rows present.

3. What is the distinct number of subjects present in the entire school

4. What is the count of the number of students in the school, whose name is Mathew and marks is 55

```
package spark.basic.cl18

import org.apache.spark.sql.SparkSession

object task2 extends App {

  val sparkSession = SparkSession.builder.master("local")
    .appName("spark").getOrCreate()

  val sparkcontetxt = sparkSession.sparkContext

  val fc = sparkcontetxt.textFile("/home/acadgild/Downloads/19_Dataset.txt")

  //task2_1
  val tup = fc.map(x =>
  {
    val row = x.split(",").toList
    (row.apply(0),row.apply(1),row.apply(2),row.apply(3).toInt,row.apply(4).toInt )

  })

  tup.foreach(println)

  println("row count = " + tup.count())

  val sub = tup.map(_._2)

  println("Distinct Subjects = "+ sub.distinct().count())

  val fil = tup.filter(_._1 == "Mathew").filter(_._4 == 55).count()

  println("Count of Mathew and 55 = "+ fil)


}
```

Output:

```
18/07/30 01:01:48 INFO HadoopRDD: Input split: file:/home/aca
18/07/30 01:01:48 INFO LineRecordReader: Found UTF-8 BOM and s
(Mathew,science,grade-3,45,12)
(Mathew,history,grade-2,55,13)
(Mark,maths,grade-2,23,13)
(Mark,science,grade-1,76,13)
(John,history,grade-1,14,12)
(John,maths,grade-2,74,13)
(Lisa,science,grade-1,24,12)
(Lisa,history,grade-3,86,13)
(Andrew,maths,grade-1,34,13)
(Andrew,science,grade-3,26,14)
(Andrew,history,grade-1,74,12)
(Mathew,science,grade-2,55,12)
(Mathew,history,grade-2,87,12)
(Mark,maths,grade-1,92,13)
(Mark,science,grade-2,12,12)
(John,history,grade-1,67,13)
(John,maths,grade-1,35,11)
(Lisa,science,grade-2,24,13)
(Lisa,history,grade-2,98,15)
(Andrew,maths,grade-1,23,16)
(Andrew,science,grade-3,44,14)
(Andrew,history,grade-2,77,11)
18/07/30 01:01:48 INFO Executor: Finished task 0.0 in stage 0.
```

```
18/07/30 01:01:48 INFO Executor: Running task 0.0 in stage
18/07/30 01:01:48 INFO HadoopRDD: Input split: file:/home/a
18/07/30 01:01:48 INFO LineRecordReader: Found UTF-8 BOM ar
18/07/30 01:01:48 INFO Executor: Finished task 0.0 in stage
row count = 22
18/07/30 01:01:48 INFO TaskSetManager: Finished task 0.0 in
18/07/30 01:01:48 INFO DAGScheduler: ResultStage 1 (count a
18/07/30 01:01:48 INFO DAGScheduler: Job 1 finished: count
18/07/30 01:01:48 INFO TaskSchedulerImpl: Removed TaskSet 1
```

```
18/07/30 01:01:50 INFO TaskSetManager: Starting task 0.0 i
18/07/30 01:01:50 INFO Executor: Running task 0.0 in stage
18/07/30 01:01:50 INFO ShuffleBlockFetcherIterator: Gettin
18/07/30 01:01:50 INFO ShuffleBlockFetcherIterator: Starte
18/07/30 01:01:50 INFO Executor: Finished task 0.0 in stag
Distinct Subjects = 3
18/07/30 01:01:50 INFO TaskSetManager: Finished task 0.0 i
18/07/30 01:01:50 INFO DAGScheduler: ResultStage 3 (count
18/07/30 01:01:50 INFO DAGScheduler: Job 2 finished: coun
18/07/30 01:01:50 INFO TaskSchedulerImpl: Removed TaskSet
```

```
18/07/30 01:01:50 INFO TaskSchedulerImpl: Removed TaskSet 4.0, whose tasks
18/07/30 01:01:50 INFO DAGScheduler: ResultStage 4 (count at task2_1.scala
18/07/30 01:01:50 INFO DAGScheduler: Job 3 finished: count at task2_1.sca
Count of Mathew and 55 = 2
18/07/30 01:01:50 INFO SparkContext: Invoking stop() from shutdown hook
18/07/30 01:01:50 INFO SparkUI: Stopped Spark web UI at http://192.168.0.1
18/07/30 01:01:50 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMas
18/07/30 01:01:50 INFO MemoryStore: MemoryStore cleared
18/07/30 01:01:50 INFO BlockManager: BlockManager stopped
```