Task 1

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

code-

```
import org.apache.spark.SparkContext
import org.apache.spark.SparkConf

object Acadgild_RDD {

  def main(args:Array[String]){
    println("Hello")
    System.setProperty("hadoop.home.dir", "C://winutils")
    //System.setProperty("spark.sql.warehouse.dir", "file:///C:/spark-warehouse");

    val conf=new SparkConf()
    conf.setAppName("Word Count")
    conf.setMaster("local")
    val sc=new SparkContext(conf)
    //val textfile=sc.textFile("file:///C:/Users/Udayan/Desktop/textfile.txt")
    val textfile=sc.textFile(args(0))
    textfile.foreach(println)
```

Output-

```
Console
<terminated> Acadgild_RDD$ [Scala Application] C:\Program Files\Ja
19/02/15 21:09:29 INFO HadoopRDD: Input split: fi
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
```

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

Input file-

```
textfile - Notepad
File  Edit  Format  View  Help
hadoop is fast
hive is sql on hdfs
spark is superfast
spark is awesome
```

Code-

```
import org.apache.spark.SparkConf

object Word_Count {

   def main(args:Array[String]){
    println("Hello")
    System.setProperty("hadoop.home.dir", "C://winutils")
    //System.setProperty("spark.sql.warehouse.dir", "file:///C:/spark-warehouse");

    val conf=new SparkConf()
    conf.setAppName("Word Count")
    conf.setMaster("local")
    val sc=new SparkContext(conf)
    //val textfile=sc.textFile("file:///C:/Users/Udayan/Desktop/textfile.txt")
    val textfile=sc.textFile(args(0))
    val words=textfile.flatMap(f=>f.split(" "))
    //print number of words in document
    words.foreach(println);
```

Output-

```
Console
<terminated> Word_Count$ [Scala Application] C:
19/02/15 21:17:00 INFO TaskSetManage
19/02/15 21:17:00 INFO Executor: Run
19/02/15 21:17:00 INFO HadoopRDD: In
hadoop
is
fast
hive
is
sql
on
hdfs
spark
is
superfast
spark
is
awesome
19/02/15 21:17:00 INFO Executor: Fin
```

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.

Input-

```
hadoop-is-fast
hive-is-sql-on-hdfs
spark-is-superfast
spark-is-awesome
```

Code-

```
    val conf=new SparkConf()
    conf.setAppName("Word Count")
    conf.setMaster("local")
    val sc=new SparkContext(conf)
    //val textfile=sc.textFile("file:///C:/Users/Udayan/Desktop/textfile.txt")
    val textfile=sc.textFile(args(0))
    val words=textfile.flatMap(f=>f.split("-"))
    //print number of words in document
    words.foreach(println);
    //count of total words
    println(words.count())
```

Output-

```
19/02/15 21:29:54 INFO DAGS
19/02/15 21:29:54 INFO DAGS
14
19/02/15 21:29:54 INFO Spar
19/02/15 21:29:54 INFO DAGS
```

Task 2

Problem Statement 1:

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

Code-

```
object Acadgild_RDD {

  def main(args:Array[String]){
    println("Hello")
    System.setProperty("hadoop.home.dir", "C://winutils")
    //System.setProperty("spark.sql.warehouse.dir", "file:///C:/spark-warehouse");

    val conf=new SparkConf()
    conf.setAppName("Word Count")
    conf.setMaster("local")
    val sc=new SparkContext(conf)
    //val textfile=sc.textFile("file:///C:/Users/Udayan/Desktop/textfile.txt")
    val textfile=sc.textFile(args(0))
    val tuple_data=textfile.map(f=>f.split(",")).map(x=>(x(0),x(1),x(2),x(3),x(4)))
    tuple_data.foreach(println)
```

Output-

```
<terminated> Acadgild_RDD$ [Scala Applicatio
19/02/16 09:04:46 INFO Executor: F
19/02/16 09:04:46 INFO HadoopRDD:
(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)
```

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

Code-

```
val textfile=sc.textFile(args(0))
val tuple_data=textfile.map(f=>f.split(",")).map(x=>(x(0),x(1),x(2),x(3),x(4)))
tuple_data.foreach(println)
//count of number of rows
println(tuple_data.count())
```

Output-

```
19/02/16 09:11:29
19/02/16 09:11:29
22
19/02/16 09:11:29
19/02/16 09:11:29
```

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

Code-

```
//distinct subject in school
tuple_data.map(f=>f._2).distinct().foreach(println)
```

Output-

```
19/02/16 09:16::
maths
history
science
19/02/16 09:16::
```

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

code-

```
//rcords whose name is matews and marks=55
println(tuple_data.filter(f=>f._1=="Mathew" && f._4.toDouble==55).count())
```

Output-

```
19/02/16 09:
19/02/16 09:
2
19/02/16 09:
```

Problem Statement 2:

1. What is the count of students per grade in the school?

2. Find the average of each student (Note - Mathew is grade-1, is different from Mathew in some other grade!)

3. What is the average score of students in each subject across all grades?

4. What is the average score of students in each subject per grade?

5. For all students in grade-2, how many have average score greater than 50?

Problem Statement 3:

Are there any students in the college that satisfy the below criteria:

1. Average score per student_name across all grades is same as average score per