**Assignment 6 - Spark RDDs**

**Submission Deadline: *26th April-2018***

This assignment is a spark RDD tutorial. It will help you to grasp the basic concept of functionalities and operations offered by the APACHE SPARK.

You must use the spark functionality of persisting data in memory and partitioning. Try to write an efficient query.

**Question 1 (35 marks)**

**Consider the format of the input is as follows**

|  |  |  |
| --- | --- | --- |
| RollNo | Course | Grade |
| 17L3021 | DB | A |
| 16K3999 | ITC | B |

**Write spark code for the following:**

1. Input the data as a rdd and filter the data such that year of the roll-number should be in the range of 1995- 2018 that is 95-18.
2. Partition the input data on the base of Campus. (override Spark Partitioner)
3. Persist the partitioned data.
4. For each campus print the total number of students.
5. Provide ordering to compare the two roll-numbers using the following rule
   1. For campus use lexicographic ordering that is F < I < k <L< P
   2. For year follow the rule of year 16<17 and 99 < 01
   3. For last part of roll-number follow int ordering.
6. For each course print the number of failures in each campus.
7. Remove all the duplicate values
8. For each student compute the GPA. Assume only five grades (Grade A GPA=4, Grade B GPA=3, Grade C GPA 2, Grade D GPA 1 and Grade F GPA=0)

**Submission Details:** *You must submit Source code, Jar file and Output file.*

*No form of plagiarism will be tolerated.*