BIG DATA ANALYTICS

ASSIGNMENT 1

SPRING 2022

Due Date: 29th March 2022 (Submit Code file online on google classroom)

Instructions:

- *The name of the file should be your rollnumber-Question number*
- *Do not copy the work of your peers. If cheating is detected, your case will be referred to DC.*

Question 1: (10 marks) We have a huge data file of FAST students. The file consists of the roll number of the FAST students, followed by the subject code and the student grade in that subject.

Consider an input file in the following format

```
L22-2100 DB D
K21-1601 SE F
I21-1601 OS F
K21-1702 DS B
L21-1705 OS A
L22-2101 DB D
K21-1601 OS F
L21-1601 SE F
L21-1702 SE B
L21-1705 DB A
```

Write spark code for the following. Consider each part as separate. Input the data from the text file in an RDD

i. Select the records of students from the Lahore campus. Display a few records and print the count of the students from Lahore.
ii. Filter the records of the students from the year in the range of 1995- 2018.
iii. Display the count of students on each Campus.
iv. Partition the input data on the base of Campus. (override Spark Partitioner).
v. For each course, print the number of failures on each Campus.
vi. Remove all the duplicate rows from input data.
vii. Find the minimum and maximum grades in each subject. The ordering of grades is as follows A >B >C >D >F
viii. We wish to sort the file based on the roll number (hint work with sortByKey). The two roll-numbers are compared using the following rule
    a. For Campus use lexicographic ordering that is F < I < k <L< P
    b. For year follow the rule of year 16<17 and 99 < 01
    c. For the last part of roll-number, follow int ordering.

ix. 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)
x. Convert grades to GPA as mentioned in part viii and find the average GPA of each Subject