Assignment 2

The overall goal of these assignments is to make you familiar with basics of spark. The first problem will carry 3 marks, while the second problem will carry 5 marks. You need to upload the code in Moodle- two separate files for two problems. The files should be named as follows:
For the first problem: yourRollNo-1.<extension>
For the second problem: youRollNo-2.<extension>

The letters in your roll must be in small. Strictly follow the file naming convention. Otherwise, I may miss your file and you may get zero score (which you certainly do not deserve).

Do not send me your code via email. Upload all through Moodle.

1. You are given a set of feature vectors of real numbers (n-dimensional), one feature vector per line. The first entry of the vector is the class label (1 or 2 or 3 and so on) in which the feature vector belongs. You have to write a spark program that will compute the mean vector for each class.
   (The sample data is given in **mean.txt**).

2. You are given a 2 column data file for two variables X (1st column) and Y (2nd column). The values are real numbers. The numbers are separated by space. Write spark code that will compute the Pearson co-relation coefficient between X and Y.
   (The sample data is given in **cor.txt**).

Submission deadline: Normal: 3rd April. Grace period: 5th April, with 50% deduction.