DA-IICT
IT 508, Winter 2021-2022
Lab Exercise 6
Date: 19/04/2022, Expected by: 03/05/2022

This lab focuses on two things: (1) Basic RDD operations, (2) Data Preprocessing. Note that some of the problems below will require you to write explanation about the code lines/concepts used in the solution.

Lab Problems:

1. (a) By taking an RDD of your choice, try out all the functions in Tables 3-2, 3-3 and 3-4 in chapter 3, [1]. If you cannot choose an RDD, you may use the example RDDs given in the chapter.
   (b) Try persisting an RDD with a replication factor of 2 on a Hadoop cluster. You may do it on Scala or Python. For reference, you may see example 3-40 in chapter 3, [1].

2. Study initial details of the "PageRank" algorithm, and implement it over Spark. You should see the example implementation in example 4-25 in chapter 4, [1]. While preparing the report, in addition to the screen shots, you need to write a short note on the "PageRank" algorithm

3. (a) This exercise focuses on data preprocessing and feature engineering. Refer to chapter 25 in [2]. Study (i) "StandardScaler", (ii) "MinMaxScaler", (iii) "MaxAbsScaler", (iv) "ElementwiseProduct", and (v) "Normalizer", and try out each of the operation on data set of your choice.
   (b) From the section "Text Data Transformers", try out "Removing Common Words" and "Creating Word Combinations".

   While solving the above two parts, you need to understand the logic and implementation of each of the operations.

References - books for perusal:

[1] *Learning Spark*, H. Karau, A. Konwinski, P. Wendell and M. Zaharia, O'Reilly Media Inc.
[2] *Spark the Definitive Guide*, B. Chambers and M. Zaharia, O'Reilly Media Inc.

General Instructions:

- There is a lot of help available online. You should definitely search for your queries online to get an early and a better resolution.
- Your lab report must contain a list of steps you took to run the programs for the two problems above and the output. For putting the output, use the screen shot. Although it is desired that you solve the problems completely, but if this does not happen, you can give the output up to the stage you could reach while solving the problems.
- The lab is intentionally made from the text books and refers to a lot of online content, so that you have ample resources to refer to and learn.