DA-IICT
IT 508, Winter 2021-2022
Lab Exercise 5
Date: 12/04/2022, Expected by: 18/04/2022

This lab focuses on two things: (1) SPARK SQL, (2) SPARK MLlib. 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. This problem intends to give a basic understanding of the dataframes API in SPARK. You need to try out all the code snippets given in "An End-to-End Example" in chapter 2 from [2]. While preparing the report, in addition to the screen shots, write down about what you understand about the "physical plan" creation in SPARK.

2. This problem intends to give a basic understanding of SPARK SQL by loading and quering tweets. You need to try out the "Basic Query Example", "Sparl SQL UDF" and example 9-40 in chapter 9 from [1]. While preparing the report, in addition to the screen shots, share your ideas about doing sentiment analysis of the tweets in SPARK. Specifically, what will be the basic steps involved in doing so (you need not implement this in SPARK).

3. This problem intends to give a working knowledge of the classification task in SPARK MLlib. You need to try out the "Example: Spam Classification" in chapter 11 from [1]. While preparing the report, in addition to the screen shots, you need to explain "HashingTF", and write a short note on Logistic regression in SPARK MLlib. For reference, you can search for videos on logistic regression on the channel {a}, and also for some use cases for logistic regression on the channel {b}.

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.

References - online for perusal:

{a} *The AI University*, available on: `https://www.youtube.com/c/TheAIUniversity/videos`.
{b} *Data Science for Everyone*, available on: `https://www.youtube.com/c/DataScienceforEveryone/videos`.

Disclaimer: For the video links above, I, or DA-IICT, do not endorse any of the online learning platforms and/or the video creators. Links are provided since the content is freely available and I assume that it might help in learning.

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.