DA-IICT
IT 508, Winter 2021-2022
Lab Exercise 4
Date: 29/03/2022, Expected by: 06/04/2022

This lab focuses on two things: (1) HIVE, (2) SPARK

Lab Problems:

1. The first task is to install Spark and gain basic working knowledge. For the same, you can refer to chapter 2 in [3] and go through sections "Downloading Spark" and "Introduction to Spark's Python and Scala Shells" (you may also refer to the some initial chapters in [4] as well). Subsequently, do the following.
   (a) Create a file "sample.txt" with a few sentences written in English. First, put the file in the local file system and do a word count. Then, with Hadoop installed in pseudo-distributed mode, put the file in HDFS and again do the word count. You can refer to {a} and {b} for a demo on solving this problem.
   (b) Install PySpark and repeat 1-(a) by doing the word count through it.
   (c) Modify your code lines for 1-(a) and 1-(b) such that case-insensitive counting of words is done.
2. In this exercise you need to learn installing HIVE and running some basic queries over a dataset. For the same, refer to chapter 11 (Hive and Hadoop herd) from [1] (you may also refer to the chapter 12 in [2] as well). The task is to re-create the example queries discussed in section 11.1.2. For this, you will be working on the patent dataset used in lab 2 earlier. Note that while preparing the report for this problem, you need to write a short note on "HIVE over Hadoop". Also, for this problem, I am not putting up any reference to some online video, however, you are free to refer to any of your choice.

References - books for perusal:

[1] *Hadoop In Action*, Chuck Lam, Manning Publications Co.
[2] *Hadoop The Definitive Guide*, Tom White, O' Reilly.
[3] *Learning Spark*, H. Karau, A. Konwinski, P. Wendell and M. Zaharia, O'Reilly Media Inc.
[4] *Spark the Definitive Guide*, B. Chambers and M. Zaharia, O'Reilly Media Inc.

References - online for perusal:

{a} *Apache Spark Word Count example - Spark Shell*, available on: `https://www.youtube.com/watch?v=HQTB3hlLD6E`.
{b} *Spark Word Count Example*, available on: `https://www.youtube.com/watch?v=I4B96EngFa8`.

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.