DA-IICT
IT 508, Winter 2021-2022
Lab Exercise 2
Date: 08/02/2022, Expected by: 22/02/2022

This lab focuses on two things: (1) doing changes in the code of the built-in example "WordCount.java" and testing it, (2) downloading a dataset and processing it through the MapReduce framework.

Lab Problems:

1. Do the three modifications, given by (1, (2) and (3) on the lower part of page 17 in [2], to the "WordCount.java". Run the modified code on the same input text file you must have used in lab 1 for WordCount.

2. Refer to chapter 4 in [2], and go through sections 4.1 - 4.3. The task is to download the patent data set from `https://www.nber.org/research/data/us-patents` and use it with listings 4.1 and 4.2. Our aim is to check if we can get the same output as given there.

   You can work in the terminal or Eclipse. If you choose Eclipse and are stuck with running programs using Hadoop libraries in it, you can refer to {e}.

References (books) for perusal:

[1] *The Complete Reference Java*, Herbert Schildt, Tata McGraw-Hill, 7th Edition.
[2] *Hadoop In Action*, Chuck Lam, Manning Publications Co.
[3] *Hadoop The Definitive Guide*, Tom White, O' Reilly.

References (online) for perusal:

{a} *Java Tutorial for Beginners [2020]*, Mosh, available on: `https://www.youtube.com/watch?v=eIrMbAQSU34`
{b} *Java Tutorial for Beginners in Hindi*, Great Learning, available on:`https://www.youtube.com/watch?v=eKRM-053ei4`
{c} *Compiling Hadoop Code with Eclipse*, available on: `https://coursys.sfu.ca/2021fa-cmpt-732-g1/pages/EclipseHadoop`.
{d} *How to Configure the Eclipse with Apache Hadoop?*, available on: `https://www.geeksforgeeks.org/how-to-configure-the-eclipse-with-apache-hadoop/`.
{e} *Running MapReduce WordCount Using Eclipse*, available on: `https://www.youtube.com/watch?v=g7C_iEeMkrM`.

Disclaimer: For the video links above, I, or DA-IICT, do not endorse any of the online learning platforms and 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 for the final stage you 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.