Instructions

Extract BigData551-finalsubmission-source.zip

It contains 4 subfolders namely datapreprocessing, R, python and Hadoop containing source codes for each of these implementations.

Configurations

Configure eclipse to be able to run hadoop code.

This step is simple and only consists of including the hadoop related jars to the build path

[This](http://blog.cloudera.com/blog/2009/04/configuring-eclipse-for-hadoop-development-a-screencast/) tutorial outlines all necessary steps

Load the project in eclipse

Running the project

Click the run button to be able to run code.

Project Flow

1. The input to the project is a file called steponeinput.txt. This step multiplies the matrix with it’s transpose and generates an output called steponeoutput.txt
2. This file is taken as input, and the inverse is found and written to a file called steptwooutput.txt
3. The output from the previous step along with the transpose of the original matrix is multiplied and written to stepthreeoutput.txt
4. Finally, this is multiplied with the response vector Y and written to the file called stepfouroutput.txt which contains the final co-efficients, the output of Linear Regression.
5. In order to compute MAE and RMSE values, then execute the R script, compute.R