PROJECT SOURCE CODE

You should attach your source separately from the main report. Please zip the source code and sample data

Source Code and Sample data files:

Some of the coding requirements are as follows:

- Please include a README file indicating which language and technology you used and how to compile your code.

- You need to use HDFS for at least some part of the project.
  - This could mean that you use HDFS for data extraction, pre-processing, or actual classification or clustering. The key is you have to use HDFS somewhere.

- Your code should be well documented.

- Ideally, you should create a UNIX script such that the entire workflow – data extraction, parsing, pre-processing, analysis, MapReduce, machine learning task – can be run using that script. The script can accept parameters from the command line.

Please attach a sample of your data. This should not be the entire dataset, but just the top few lines, so the TA can run your code. About 1000 lines/records should be fine.