CS 6350

Big Data Management & Analytics

Toxic Comment Classification

Names of students in your group:

Ch Muhammad Talal Muneer
cxm180004

Chirag Shahi
cxs180005

Number of free late days used: 2

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

README FILE

Link to the PySpark notebook on AWS: https://bdprojtoxic.s3-us-west-1.amazonaws.com/e-DSA0Y016H9GEF2SO7QSJ525OC/Untitled2.ipynb

Link to the S3 bucket to access data: https://bdprojtoxic.s3-us-west-1.amazonaws.com/train.csv

- Go to AWS EMR and create a cluster with Spark added using advanced options
- Navigate to notebooks in EMR and click on create a new notebook and open a new PySpark notebook in Jupyter lab
- Using the link given, either manually enter the code in the notebook or import into your new notebook
- Now change the variable *'input_bucket'* with the link provided above of the S3 bucket and change the variable *'input_path'* to ''
- Execute the rest of the program as it is