PROCESSING OF BIG DATA

SPARK

SESSION-3


WOW
creativity
work ?
SELF-LEARNING
help
curiosity
time
communication
go!
music
inspiration

Theme

Agenda

- Automate spark jobs
- Generate surrogate key
- Prototype an end-to-end mini project
- Broadcast variable and accumulators
- Configuration parameters
- Machine learning project end-to-end flow
- Q&A

ML project

- Train model with train.csv dataset
- Save trained model on disk
- Exchange your saved model with your peer
- You will test your peer's model with test.csv and measure the accuracy
- Accuracy should be above 80%
- Instructor has the rights to make final judgment ;-)

Reference

- https://spark.apache.org/docs/latest/ml-guide.html
- http://spark.apache.org/docs/2.4.0/api/python/pyspark.ml.html
- https://stattrek.com/chi-square-test/independence.aspx