Introduction to Workshop

Miha Torkar

Jozef Stefan Institute

17th October 2019

miha.torkar@ijs.si

Part 1: Spark and Big Data

- Spark in general

- RDD

- DataFrames

- Spark SQL

Part 2: Spark and Machine Learning
Applications to finance

- Mlib in Spark
  - Online resources

- Feature Extraction

- Spark for time series
  - Presentation of various methods
  - Predicting the stock prices

- Spark for classification
  - Presentation of methods
  - Credit default prediction

Part 3: Deep learning

- Theory of Neural networks

- Keras

- Model tuning

- More complex layers (LSTMs)

- Tensorflow cost learning example

Installation check

- Apache Spark installed and working
  - Personal installation
  - Inside Virtual Box

- Jupyter notebook integration

Additional Resources

- Book: Bill Chambers, Matei_Zaharia: Spark The Definite Guide

- API Docs: https://spark.apache.org/docs/latest/api/python/index.html

- Guides: https://spark.apache.org/docs/latest/index.html