Predictive Analytics using R and Hadoop

Table of Contents:

5. **Introduction to Sqoop** **(Day3)**
    a. Overview
    b. Invoking Sqoop
    c. Configuring Sqoop to import/ export data from HDFS to mysql and vice versa
6. **Lab 4** - Develop a working prototype for analyzing tweets on pre-fixed topics. Define external tables in hive to store raw tweets. Extract user text from raw tweets. Using sqoop, move user text to mysql database, perform sentiment analysis on user text using Naive Bayesian algorithms in Python scripts

7. **Introduction to R and R based Machine Learning** **(Day 4 - 8)**
    a. R Data Types
        a. Installing R, RStudio
        b. R Datatypes
        c. Basic syntax
        d. Variables
        e. Vectors
        f. Matrices
        g. DataFrames
        h. Lists
    b. Data Interfaces
        a. CSV files
        b. Excel files
        c. Text files
        d. Databases
        e. Web links
    c. Charts and graphs
        a. Pie chart
        b. Line charts
        c. Scatter plots
        d. Histograms
    d. Basic statistical concepts
        a. Descriptive statistics (Mean, Median, Mode, Spread measures)
        b. Normal distribution and it's characteristics
        c. Standard scores, confidence levels and intervals
    e. Machine Learning for Predictive Data Analytics
        a. What is predictive Data Analytics
        b. Predictive Data Analytics Project Lifecycle - CRISP-DM
        c. What is machine learning models
    f. Data Exploration / Preparing data for analytics
        a. Identifying data quality problems
        b. Handling data quality issues
        c. Tidy Data
        d. Analysing attributes using tools such as SPLOM (Scatter Plot Matrix)
        e. Dimensionality Reduction concept / method
    g. Supervised Learning Methods (Concepts, creating, evaluating and improvising)
        a. Support Vector Machines
        b. Neural Networks
        c. Decision trees
        d. Naive Bayesian classifiers

    e. K-NN Clustering
    f. Linear Regression

h. Unsupervised Learning Methods (Concepts, creating, evaluating and improvising)
    a. K-Means Clustering
    b. Association rules mining

i. Important concepts of machine learning
    a. Curse of dimensionality
    b. Dimensionality reduction
    c. Linearly separable data, Kernels and kernel tricks
    d. Bias variance & model effectiveness
    e. Model performance measures