e-ITEC course on Big Data Technologies and Machine Learning

27 Sep 2021 – 11 Nov 2021 (Monday – Friday) – 5 hrs per day

Tentative Day wise Schedule - 14:00 HRS - 20:00 HRS(IST)

1. Big Data Technologies

1. Introduction to Linux (6h)

The Evolution of Linux operating system, The Architecture and Structure of Linux, Installation, Introduction to Linux File system, File processing commands, Text Processing Commands, Backup and recovery commands, Network commands, Basic of I/O commands, Inter Process communication, Introduction to Users and Groups, Essentials of Effective User, Group, and Password Management, understanding permissions, Access control list and chmod command, chown and chgrp commands.

2. Introduction to Big data

a. Introduction to Big data (3h)

Introduction to big data platform, Big data challenges, Big Data Applications, Types of Big Data Technologies, Limitations and Solution of Big data Architecture, Introduction to different Big data Architectures

3. Hadoop Environment

a. Introduction to Hadoop and Hadoop Architecture (3h)

What is Hadoop, Brief History and Evolution of Hadoop, Hadoop Distributions and Vendors. Hadoop Architecture, Core components of Hadoop

b. Hadoop Distributed File System (3h)

What is HDFS, Core components of HDFS, Hadoop Server Roles: Name Node, Secondary Name Node, and Data Node

HDFS Architecture overview, The HDFS command line and web interfaces, Analyzing the Data with Hadoop, Scaling Out, high availability and Name Node federation, HDFS – Monitoring & Maintenance.

c. Hadoop Environment (6h)

Demonstration to cloudera quickstart virtual machine

How to set up Hadoop cluster and Install on Virtual Machine, Hadoop Configuration, Security in Hadoop, Administering Hadoop, common hadoop shell commands, Security in a cloudera cluster (HDFS, Hive)

d. Big data analytics with Map Reduce Framework (6h)

Hadoop Map Reduce paradigm, Map Reduce Execution Framework, Anatomy of a Map Reduce Job, Partitioners and Combiners, Input Formats (Input Splits and Records, Text Input, Binary Input, Multiple Inputs) Output Formats (Text Output, Binary Output, Multiple Outputs)

e. Big data analytics with PIG (12h)

Introduction to PIG, Pig Execution Modes, Basics of PIG Latin Programming Conventions, Data Types, Arithmetic and Relational Operators, UDF Statements, PIG Latin Scripting, PIG Built-In Functions ,Eval Functions, Load/Store Functions, Math Functions, String Functions, Date Time Functions, Writing a PIG UDF, Piggy Bank, PIG Macros, Real-Time Data Analytics using PIG

f. Big data analytics with Hive (12h)

The Hive Data-ware House, Basics of Hive Query Language, Working with Hive QL, Operators and Functions, Importing Data, Querying Data & Managing Outputs, Hive Tables (Managed Tables and Extended Tables), Partitions and Buckets, Aggregating, Joins Views, Data manipulation with Hive, User Defined Functions, Writing HQL scripts.

g. Big data analytics with Spark (12h)

Initializing Spark, Spark Components and Architecture, Resilient Distributed Datasets (RDDs), RDD Operations, Passing Functions to Spark, Working with Key-Value Pairs, Shuffle operations, RDD Persistence, Shared Variables, Working with Spark with Hadoop, Spark SQL, Dataframes and Datasets, Spark Streaming

h. Big data analytics with MongoDB (12h)

Overview of SQL (DDL, DML, TCL), Introduction to NoSQL, Difference between SQL and NoSQL, working with MongoDB (Installation, CRUD operations, Aggregation pipeline, Indexing, Data Modeling)

2. Machine learning with python programming

1. Python Programming

a. Introduction to Python Programming (12h)

Installing Python, Introduction to Python Basic Syntax, Data Types, Variables, Operators, Input/output, Python data structure, Introduction to Strings, Lists, Tuples, Dictionaries, Sets. Flow of Control (Modules, Branching) If, If- else, Nested if-else Looping, For, While, Nested loops Control Structure, Uses of Break & Continue ,Functions and methods and Exception Handling, OOPs Concepts, Python classes and objects, Introduction and Installation of Machine learning packages like PANDAS, NUMPY, SKLearn, Matplotlib, Seaborn. Mathematical Computing with Python (NumPy), Data Manipulation with Pandas, Machine Learning with Scikit–Learn.

b. Data Visualization in Python (3h)

Introduction to Data Visualization in Python (i.e. matplotlib, Seaborn)

2. Machine learning

a. Introduction to Machine Learning (6h)

What is machine learning? Types of learning, Applications of Machine learning, Evaluating ML techniques.

b. Data Preprocessing techniques (3h)

Data cleaning, scaling of continuous features, encoding of categorical features, train and test split

c. Supervised Algorithms (6h)

Linear Regression, Decision Trees, Decision Trees case study, Naive bayes classifier, assigning probabilities and calculating results, Naïve Bayes case study, K-Nearest Neighbors Algorithm and case study. Ensemble Learning: Concept of model ensembling, Random forest, Gradient boosting Machines, Model Stacking, Support Vector Machines, Neural Network and its applications, Single layer neural Network, Constructing Neural Networks model, Overview of Feed Forward Neural Network, Back propagation, Activation Functions: Sigmoid, Hyperbolic Tangent

d. Unsupervised Algorithms (6h)

Different type of Unsupervised Machine Learning Algorithms, clustering, K-mean, agglomerative clustering, Association rule mining, apriori algorithm

e. Introduction to Deep Learning (9h)

Introduction to deep Learning, Why Deep Learning is taking off? Deep Learning Architecture (Hyperparameter tuning, Attention mechanism, transfer learning, GAN), Introduction to Tensorflow, Introduction to Keras, Building blocks of deep neural networks, Activation Functions, why non-linear activation functions? Computer Vision: Introduction to Convolutional Neural Network. Sequence Modeling: Recurrent Neural Network. Real world case studies for convolutional neural networks and recurrent neural network model.

f. Introduction to NLP (6h)

Overview of NLP, Shallow Parsing, Deep Parsing, NLP with Machine Learning and Deep Learning, Pre-processing, Need of Pre-processing Data, Introduction to NLTK, Using Python Scripts, Word2Vec models, Building NLP Application.

3. Implementation of Machine learning , Deep learning and NLP with Python Programming

a. Supervised Algorithms with Python (6h) – Hands-on
b. Unsupervised Algorithms with Python (6h) – Hands-on
c. Deep Learning with Python (6h) – Hands-on
d. NLP with Python (6h) – Hands-on