**Big Data Systems:**

|  |  |
| --- | --- |
| **Session** | **Topic / Title** |
| **1** | **Different Types of Data and Storage for Data: Structured Data (Relational Databases), Semi-structured data (Object Stores), and Unstructured Data (File systems) What is Big Data? Characteristics of Big Data. Systems perspective - Processing: In-memory vs. (from) secondary storage vs. (over the) network, Locality of Reference: Principle, examples Impact of Latency: Algorithms and data structures that leverage locality, data organization on disk for better locality** |
| **2** | **Parallel and Distributed Processing: Motivation (Size of data and complexity of processing); Storing data in parallel and distributed systems: Shared Memory vs. Message Passing; Strategies for data access: Partition, Replication, and Messaging, Memory Hierarchy in Distributed Systems: In-node vs. over the network latencies, Locality, Communication Cost. Distributed Systems: Motivation (size, scalability, cost-benefit), Client-Server vs. Peer-to-Peer models, Cluster Computing: Components and Architecture** |
| **3** | **Big Data Analytics: Requirements, constraints, approaches, and technologies, Big Data Systems – Characteristics: Failures; Reliability and Availability; Consistency – Notions of Consistency.** |
| **4** | **CAP Theorem and implications for Big data Analytics, Big Data Lifecycle: Data Acquisition, Data Extraction – Validation and Cleaning, Data Loading, Data Transformation, Data Analysis and Visualization. Case study – Big data application** |
| **5** | **Distributed Computing - Design Strategy: Divide-and-conquer for Parallel / Distributed Systems - Basic scenarios and Implications. Programming Patterns: Data-parallel programs and map as a construct; Tree-parallelism, and reduce as a construct; Map-reduce model: Examples (of map, reduce, map-reduce combinations, and Iterative map-reduce)** |
| **6** | **Hadoop: Introduction, Architecture, and Map-reduce Programming on Hadoop** |
| **7** | **Hadoop: Hadoop Distributed File System (HDFS), Scheduling in Hadoop (using YARN). Example – Hadoop application.** |
| **8** | **Hadoop Ecosystem: Databases and Querying (HBase, Pig, and Hive)** |

**Stream Processing and Analytics:**

|  |  |
| --- | --- |
| **Session** | **Topic / Title** |
| **1** | **Thinking about Data Systems, Reliable, Scalable and Maintainable Data Applications, Properties of Data, Scaling with traditional databases, Big Data Systems, Desired properties of Big Data Systems** |
| **2** | **Data Model for Big Data, Generalized Big Data System Architecture, Real-time systems, Difference between Batch processing and Stream Processing, Difference between real-time and streaming systems** |
| **3** | **Streaming Data Applications, Databases and Streams, Usage patterns of Streaming Data, Sources of Streaming Data, Complex Event Processing Systems** |
| **4** | **Generalized Streaming Data Architecture, Lambda Architecture, Kappa Architecture, Streaming Data System Component, Features of Real-time Architecture, A real-time architecture checklist** |
| **5-6** | **Service Configuration and Coordination Systems, Maintaining the state, Apache ZooKeeper, Kafka Fundamentals Overview, Use-Cases and applications, Architecture, Kafka Topics, Producer and Consumer Using CLI, Programming Kafka, Simple Kafka Producer, Simple Kafka Consumer, Producer, Consumer Configuration, Producer, Consumer Execution, Kafka Consumer Groups** |
| **7-8** | **Streaming Data Processor Concepts, Timing Concepts, Windowing, Joins, Storage for Streaming Data, NoSQL storage Systems, Choosing a Storage technology, Delivery of Streaming Metrics** |

**Natural Language Processing:**

|  |  |
| --- | --- |
| **Session** | **Topic / Title** |
| **1** | **Natural Language Understanding and Generation: The Study of Language, Applications of Natural Language Understanding, Evaluating Language Understanding Systems, The Different Levels of Language Analysis, The Organization of Natural Language Understanding Systems.** |
| **2** | **N-gram Language Modelling: N-Grams, Generalization and Zeros, Smoothing, The Web and Stupid Backoff, Evaluating Language Models, Smoothing, The Web and Stupid Backoff.** |
| **3** | **Neural Network and Neural Language Modelling: Units, The XOR problem, Feed-Forward Neural Networks, Training Neural Nets, Neural Language Models.** |
| **4** | **Vector semantics and Embedding: Lexical semantics, Vector semantics, Word and Vectors, TFIDF, Word2Vec, Skip gram and CBOW, Glove, Visualizing Embedding’s.** |
| **5** | **Part-of-Speech Tagging: (Mostly) English Word Classes, The Penn Treebank Part-of-Speech Tag set, Part-of-Speech Tagging, Markov Chains, The Hidden Markov Model, HMM Part-of-Speech Tagging, Part-of-Speech Tagging for Morphological Rich Languages.** |
| **6** | **Hidden Markov Model Algorithms: Likelihood Computation: The Forward Algorithm, Decoding: The Viterbi Algorithm, HMM Training: The Forward-Backward Algorithm, Maximum Entropy Markov Model, Bidirectionality.** |
| **7** | **Topic modelling: Mathematical foundations for LDA, Multinomial and Dirichlet distributions, Intuition behind LDA, LDA Generative model, Latent Dirichlet Allocation Algorithm and Implementation, Gibbs Sampling.** |

**Deep Learning:**

|  |  |
| --- | --- |
| **Session** | **Topic / Title** |
| **1** | **Fundamentals of Neural Network: Objective of the course, Supervised, unsupervised, semi-supervised and reinforcement learning problems, Why Deep Learning?, Applications of Deep Learning, Biological neuron vs artificial neuron, Connectionism model** |
| **2** | **Fundamentals of Neural Network: Perceptron, Perceptron learning algorithm, Multilayer Perceptron (MLP), MLP on Boolean, reals and continuous values** |
| **3** | **Fundamentals of Neural Network: MLP as classifiers, MLP as Universal approximators, Issue of Depth and Width** |
| **4** | **Deep Feedforward Neural Network: MLP with hidden Layers, Forward Propagation, Backward Propagation, Training a DNN using Gradient Descent algorithm, Computational Graphs** |
| **5** | **Deep Feedforward Neural Network: Activation Functions, Softmax Regression** |
| **6** | **Optimization algorithms for Deep models: Challenges – Saddle points and plateau, Non-convex optimization intuition, Stochastic Gradient Descent (SGD), Minibatch SGD, Overview of Rprop, Quickprop, Momentum, Nastrov’s Accelarated Momentum, Algorithms with Adaptive Learning Rates, Adagrad, RMSprop, ADAM** |
| **7** | **Regularization for Deep models: Model Selection, Underfitting, and Overfitting, L1 and L2 Regularization, Dropout, Challenge - Vanishing and Exploding Gradients, Parameter Initialization, Challenge Covariance Shift, Batch Normalization** |
| **8** | **Convolutional Neural Network: Basics of Computer Vision and Invariance, Convolutions for Images, Learning a Kernel, Padding and stride, Channels, Pooling, Designing a CNN** |