Paper – VII (A): Big Data

[4 HPW :: 4 Credits :: 100 Marks (External:80, Internal:20)]

UNIT – I

Getting an overview of Big Data: Introduction to Big Data, Structuring Big Data, Types of Data, Elements of Big Data, Big Data Analytics, and Advantages of Big Data Analytics.

Introducing Technologies for Handling Big Data: Distributed and Parallel Computing for Big Data, Cloud Computing and Big Data, Features of Cloud Computing, Cloud Deployment Models, Cloud Services for Big Data, Cloud Providers in Big Data Market.

UNIT – II

Understanding Hadoop Ecosystem: Introducing Hadoop, HDFS and MapReduce, Hadoop functions, Hadoop Ecosystem. **Hadoop Distributed File System**- HDFS Architecture, Concept of Blocks in HDFS Architecture, Namenodes and Datanodes, Features of HDFS. MapReduce.

Introducing HBase- HBase Architecture, Regions, Storing Big Data with HBase, Combining HBase and HDFS, Features of HBase, Hive, Pig and Pig Latin, Sqoop, ZooKeeper, Flume, Oozie.

UNIT- III

Understanding MapReduce Fundamentals and HBase: The MapReduceFramework ,Exploring the features of MapReduce, Working of MapReduce, Techniques to optimize MapReduce Jobs, Hardware/Network Topology, Synchronization, File system, Uses of MapReduce, Role of HBase in Big Data Processing- Characteristics of HBase.

Understanding Big Data Technology Foundations: Exploring the Big Data Stack, Data Sources Layer, Ingestion Layer, Storage Layer, Physical Infrastructure Layer, Platform Management Layer, Security Layer, Monitoring Layer, Visualization Layer.

UNIT – IV

Storing Data in Databases and Data Warehouses: RDBMS and Big Data, Issues with Relational Model, Non – Relational Database, Issues with Non Relational Database, Polyglot Persistence, Integrating Big Data with Traditional Data Warehouse, Big Data Analysis and Data Warehouse.

NoSQL Data Management: Introduction to NoSQL, Characteristics of NoSQL, History of NoSQL, Types of NoSQL Data Models- Key Value Data Model, Column Oriented Data Model, Document Data Model, Graph Databases, Schema-Less Databases, Materialized Views, CAP Theorem.

Reference

1. BIG DATA, Black Book TM, DreamTech Press, 2016 Edition.

Suggested Reading:

2. Seema Acharya, SubhasniChellappan , "BIG DATA and ANALYTICS", Wiley publications, 2016
3. Nathan Marz and James Warren, "BIG DATA- Principles and Best Practices of Scalable Real-Time Systems", 2010

Practical – 7(A): Big Data (Lab)

[3 HPW:: 1 Credit :: 25 Marks]

Objectives:

- Installation and understanding of working of HADOOP
- Understanding of MapReduce program paradigm.
- Writing programs in Python using MapReduce
- Understanding working of Pig, Hive
- Understanding of working of Apache Spark Cluster

1. Setting up and Installing Hadoop in its two operating modes:
   - Pseudo distributed,
   - Fully distributed.
2. Implementation of the following file management tasks in Hadoop:
   - Adding files and directories
   - Retrieving files
   - Deleting files
3. Implementation of Word Count Map Reduce program
   - Find the number of occurrence of each word appearing in the input file(s)
   - Performing a MapReduce Job for word search count (look for specific keywords in a file)
4. Map Reduce Program for Stop word elimination:
   - Map Reduce program to eliminate stop words from a large text file.
5. Map Reduce program that mines weather data. Weather sensors collecting data every hour at many locations across the globe gather large volume of log data, which is a good candidate for analysis with MapReduce, since it is semi structured and record-oriented. Data available at: https://github.com/tomwhite/hadoop-book/tree/master/input/ncdc/all.
   - Find average, max and min temperature for each year in NCDC data set?
   - Filter the readings of a set based on value of the measurement, Output the line of input files associated with a temperature value greater than 30.0 and store it in a separate file.
6. Install and Run Pig then write Pig Latin scripts to sort, group, join, project, and filter your data.
7. Write a Pig Latin script for finding TF-IDF value for book dataset (A corpus of eBooks available at: Project Gutenberg)
8. Install and Run Hive then use Hive to create, alter, and drop databases, tables, views, functions, and indexes.
9. Install, Deploy & configure Apache Spark Cluster. Run apache spark applications using Scala.
10. Perform Data analytics using Apache Spark on Amazon food dataset, find all the pairs of items frequently reviewed together.

Paper – VII (B) :Deep Learning

[4 HPW:: 4 Credits :: 100 Marks (External:80, Internal:20)]

Objective: The main objective of this course is to give a practical introduction to Deep Learning using Keras. It covers the concepts of deep learning and their implementation.

Outcomes:

At the end of the course the student will be able to

1. Understand the basics of deep learning
2. Understand the usage of tensors in deep learning
3. Use Python deep-learning framework Keras, with Tensor-Flow as a backend engine.

Unit-I

Introduction: History, Hardware, Data, Algorithms

Neural Networks, Data representations for neural networks, Scalars (0D tensors), Vectors (1D tensors), Matrices (2D tensors), 3D tensors and higher-dimensional tensors, Key attributes,. Manipulating tensors in Numpy, The notion of data batches, Real-world examples of data tensors, Vector data, Time series data or sequence data, Image data, Video data

Unit-II

Tensor operations: Element-wise operations, Broadcasting, Tensor dot, Tensor reshaping, Geometric interpretation of tensor operations, a geometric interpretation of deep learning,

Unit-III

Gradient-based optimization, Derivative of a tensor operation, Stochastic gradient descent,. Chaining derivatives: the Backpropagation algorithm

Neural networks: Anatomy, Layers, Models, Loss functions and optimizers

Unit-IV

Introduction to Keras, Keras, TensorFlow, Theano, and CNTK

Recurrent neural networks: A recurrent layer in Keras, Understanding the LSTM and GRU layers

Reference:

1. FrançoisChollet. Deep Learning with Python. Manning Publications, 2018

Suggested Reading:

2. AurélienGéron. Hands on Machine Learning with SciKit-Learn, Keras and Tensor Flow. O'Reily, 2019
3. Andrew W. Trask. Grokking Deep Learning.Manning Publications, 2019

Practical – 7(B): Deep Learning (Lab)

[3 HPW :: 1 Credit :: 25 Marks]

Objectives: The main objective of this lab is to develop deep learning models using Keras

Deep Learning Tools
Students are expected to learn Keras deep-learning framework (https://keras.io), which is open source and free to download. They should have access to a UNIX machine; though it's possible to use Windows, too. It is also recommended that they work on a recent NVIDIA GPU

Note: The exercises should following **Keras workflow** consisting of four steps
1. Define your training data: input tensors and target tensors
2. Define a network of layers (or *model*) that maps your inputs to your targets
3. Configure the learning process by choosing a loss function, an optimizer, and some metrics to monitor
4. Iterate on your training data by calling the fit() method of your model

Exercise 1:
Dataset:
IMDB dataset, a set of 50,000 highly polarized reviews from the Internet Movie Database. They're split into 25,000 reviews for training and 25,000 reviews for testing, each set consisting of 50% negative and 50% positive reviews. the IMDB dataset comes packaged with Keras
Binary Classification Task:
Build a network to classify movie reviews as positive or negative, based on the text content of the reviews.

Exercise 2:
Dataset:
Reuters dataset, a set of short newswires and their topics, published by Reuters in 1986. It's a simple, widely used toy dataset for text classification. There are 46 different topics; some topics are more represented than others, but each topic has at least 10 examples in the training set. Reuters dataset comes packaged as part of Keras.
Single-label Multi class Classification Task:
Build a network to classify Reuters newswires into 46 mutually exclusive topics. Each data point should be classified into only one category (in this case, topic),. The problem is more specifically an instance of *single-label, multiclass classification*.

Exercise 3:
Dataset:
The Boston Housing Price dataset has an interesting difference from the two previous examples. It has relatively few data points: only 506, split between 404 training samples and 102 test samples. And each *feature* in the input data (for example, the crime rate) has a

different scale. For instance, some values are proportions, which take values between 0 and 1; others take values between 1 and 12, others between 0 and 100, and so on.

Regression Task:

The two previous examples were classification problems, where the goal was to predict a single discrete label of an input data point. Another common type of machine-learning problem is *regression*, which consists of predicting a continuous value instead of a discrete label. You'll attempt to predict the median price of homes in a given Boston suburb in the mid-1970s, given data points about the suburb at the time, such as the crime rate, the local property tax rate, and so on.

4. More exercises can be defined on similar lines.