Adikavi Nannaya University

Subject: Data Science

Recommended Combination: BSc - Computers-Statistics-Data Science

Eligibility: Intermediate (MPC/MEC) or equivalent course

Semester	Part	Subject	Hrs.	Credits	IA	ES	Total
FIRST YEAR							
SEMESTER I	I	Maths for Data Science	4	3	25	75	100
		Tutorial	2	2	0	50	50
SEMESTER II	II	Introduction to Data Science With R	4	3	25	75	100
		R Lab	2	2	0	50	50
SECOND YEAR							
SEMESTER III	III	Data Mining and Data Analysis	4	3	25	75	100
		Data Mining and Data Analysis using 'R'Lab	2	2	0	50	50
SEMESTER IV	IV	Multivariate Technique for Data Analysis	4	3	25	75	100
		Multivariate Technique for Data Analysis Using 'R' Lab	2	2	0	50	50
THIRD YEAR							
SEMESTER V	V	Big Data Technology	3	3	25	75	100
		Big Data Technology through Hadoop Lab -I	2	2	0	50	50
	VI	Big Data Acquisition	3	3	25	75	100
		Big Data Technology through Hadoop Lab - II	2	2	0	50	50
SEMESTER VI	**VII (A/B)**	**Elective-I**	3	3	25	75	100
		A. Java Programming for Data Analytics					
		B .Python Programming for Data Analytics					
		Lab for Elective –I	2	2	0	50	50
	VIII ClusterA- 1,2,3 Or Cluster B- 1,2,3	**Elective-II(cluster A)**					
		1. Spark Programming +. Spark Programming Lab	3	3	25	75	100
		2.Marketing Analytics + Mongo DB Lab					
		3.Social Network Analytics + Social Network Analytics through 'R' Lab					
		Lab	2	2	20	30	50
		Elective-II(cluster B)					
		1.Data & Information Security + Information Security through Python Lab	3	3	25	75	100
		2. Spark Programming + Spark Programming Lab					
		3.Big Data Security + Mango DB Lab					
		Lab & Project	2	2	20	30	50

Adikavi Nannaya University
Subject: Data Science
Recommended Combination: BSc - Computers-Statistics-Data Science
Eligibility: Intermediate (MPC/MEC) or equivalent course

I YEAR I SEMESTER

MATHS FOR DATA SCIENCE

Objective

The course is a brief overview of the basic tools from Linear Algebra and Multivariable Calculus that will be needed in subsequent course of the program.

Outcome

By completing the course the students will have been reminded of the basic tools of Linear Algebra and Multivariable Calculus needed in subsequent courses in the program notably:

- Fundamental properties of matrices, their norms, and their applications.
- Differentiating/Integrating multiple variable functions and the role of the gradient and the hessian matrix.
- Basic properties of optimization problems involving matrices and functions of multiple variables.

Unit-I

Matrices and Basic Operations, Special structures Matrices and Basic Operations, Interpretation of matrices as linear mappings and some examples.

Square Matrices, Determinants Properties of determinants, singular and non-singular matrices, examples, finding an inverse matrix.

Unit-II

Eigen values and Eigenvectors Characteristic Polynomial, Definition of Left/Right Eigen values and Eigenvectors, Caley – Hamilton theorem, singular value Decomposition, Interpretation of Eigen values/vectors.

Unit-III

Linear Systems Definition, applications, solving linear systems, linear inequalities, linear programming.

Unit-IV

Real-valued functions of two or more variables. Definition, examples, simple demos, applications.

Unit-V

Analysis elements Distance, Limits, Continuity, Differentiability, the gradient and the Gaussian.

Optimization problems Simple examples, motivation, the role of the Hessian maxima and minima and related extreme conditions.

Integration Double integrals, Fubini's theorem, properties, applications.

References

1. Gilbert Strang, *Linear Algebra and its Applications*. Thomson /Brooks Cole (Available in a Greek Translation).
2. Thomas M. Apostol, Calculus, Wiley, 2nd Edition, 1991 ISBN 960-07-0067-2.
3. Michael Spivak. *Calculus,* publish or Perish, 2008, ISBN 978-0914098911.
4. Ross L. Finney, Maurice D.Weir . and Frank R. Giordano. Thomas's Calculus, Pearson 12th Edition 2009.
5. David C. Lay, Linear Algebra and Its Applications, 4th Editoin.
6. Yourself saad, Iterative Methods for spare Linear Systems.

Student Activity:

1. Find the Eigenvectors of A = { 1 1 1 1,2 3 4 5,3 4 5 6}
2. Find orthogonal S =Spam{(1 1 1 1),(1 4 4 0),(-1 4 4 0),(-4 2 2 0)}

Adikavi Nannaya University
Subject: Data Science
Recommended Combination: BSc - Computers-Statistics-Data Science
Eligibility: Intermediate (MPC/MEC) or equivalent course

I YEAR I SEMESTER

MATHS FOR DATA SCIENCE

Tutorial

1. Study various applications of Matrices.
2. Study different polynomial functions and their uses.
3. Take one real world example and apply the Linear System solution.
4. Study some real valued functions and its applications.
5. Study and solve one optimization problem.

Adikavi Nannaya University

Subject: Data Science

Recommended Combination: BSc - Computers-Statistics-Data Science

Eligibility: Intermediate (MPC/MEC) or equivalent course

I YEAR II SEMESTER

INTRODUCTION TO DATA SCIENCE WITH R

Objective

Data Science is a fast-growing interdisciplinary field, focusing on the analysis of data to extract knowledge and insight. This course will introduce students to the collection. Preparation, analysis, modelling and visualization of data, covering both conceptual and practical issues. Examples and case studies from diverse fields will be presented, and hands-on use of statistical and data manipulation software will be included.

Outcomes

i. Recognize the various discipline that contribute to a successful data science effort.
ii. Understand the processes of data science identifying the problem to be solved, data collection, preparation, modelling, evaluation and visualization.
iii. Be aware of the challenges that arise in data sciences.
iv. Develop an appreciation of the many techniques for data modelling and mining.
v. Be cognizant of ethical issues in many data science tasks.
vi. Be comfortable using commercial and open source tools such as the R language and its associated libraries for data analytics and visualization.

Unit-I

Introduction to the field of data science, different types of data(Data Base data, data Warehouse data, Transaction Data, Stock Exchange Data, Time Series and Bio logical data) ; data collection.

Unit-II

Experimental design; data attributes; data cleaning; data characterization and analysis.

Unit-III

Data modelling and mining techniques; model evaluation; visualization; application of data science introducing to R – R Data structures – Help functions in R

Unit-IV

Vectors-Scalars-Declarations- recycling-Common Vector operations – Using all and any Vectrorized operations-NA and NULL values – Filtering – Vectorized if- then else-Vector Equality – Vector Element names.

Creating matrices –Matrix operations-Applying Functions to Matrix Rows and Columns – Adding and deleting rows and columns.

Unit-V

Vector /Matrix Distinction –Avoiding Dimension Reduction –Higher Dimensional arrays – lists- Creating lists – General list operations – Accessing list components and values – applying functions to lists –recursive lists. Creating Data Frames – Matrix –like operations in frames – Merging Data Frames – Applying function to Data frames.

References

1.Nina Zumel, John Mount, "Practical Data Science with R", Manning Publications, 2014.
2.Jure Leskovec, Anand Rajaraman, Jeffrey D.Ullman, "Mining of Massive Datasets", Cambridge University Press, 2014.
3.Mark Gardener, "Beginning R - The Statistical Programming Language", John Wiley & Sons, Inc., 2012.
4.W. N. Venables, D. M. Smith and the R Core Team, "An Introduction to R", 2013.
5.Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, Abhijit Dasgupta, "Practical Data Science Cookbook", Packt Publishing Ltd., 2014.
6.Nathan Yau, "Visualize This: The FlowingData Guide to Design, Visualization, and Statistics", Wiley, 2011.
7.Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, "Professional Hadoop Solutions", Wiley, ISBN: 9788126551071, 2015.

Student Activity

Databases need to undergo pre-processing to be useful for data mining. Dirty data can cause confusion for the data mining procedure, resulting in unreliable output. Data cleaning includes smoothing noisy data, filling in missing values, identifying and removing outliers, and resolving inconsistencies.

I YEAR II SEMESTER

R LAB

1) Installing R and R studio
2) Basic operations in r
3) Getting data into R, Basic data manipulation
4) Basic plotting
5) Loops and functions

Adikavi Nannaya University
Subject: Data Science
Recommended Combination: BSc - Computers-Statistics-Data Science
Eligibility: Intermediate (MPC/MEC) or equivalent course

II YEAR III SEMESTER

DATA MINING AND DATA ANALYSIS

Objective

- To learn data analysis techniques.
- To understand Data mining techniques and algorithms.
- Comprehend the data mining environments and application.

Outcome

Students who complete this course will be able to

- Compare various conceptions of data mining as evidenced in both research and application.
- Characterize the various kinds of patterns that can be discovered by association rule mining.
- Evaluate mathematical methods underlying the effective application of data mining.

Unit-I

Data mining-KDD versus data mining, Stages of the Data Mining Process-Task primitives., Data Mining Techniques – Data mining knowledge representation.

Unit-II

Data mining query languages- Integration of Data Mining System with a Data Warehouse-Issues, Data pre-processing – Data Cleaning.

Data transformation – Feature selection – Dimensionality reduction – Discretization and generating concept hierarchies – Mining frequent patterns association – correlation.

Unit-III

Classification: Basic Concepts, General Approach to solving a classification problem, Decision Tree Induction: Working of Decision Tree, building a decision tree, methods for expressing an attribute test conditions, measures for selecting the best split, Algorithm for decision tree induction.

Model Over fitting: Due to presence of noise, due to lack of representation samples, evaluating the performance of classifier: holdout method, random sub sampling, cross-validation, bootstrap

Unit-IV

Bayesian Classification – Rule Based Classification – Classification by back propagation – Support Vector Machines –Associative Classification – Lazy Learners – Other Classification Methods-

Unit-V

Clustering techniques – Partitioning methods-k-means-Hierarchical Methods – Distance based agglomerative and divisible clustering – Density – Based Methods – Expectation maximization – Grid Based Methods – Model – Based Clustering – Methods – Constraint – Based Cluster Analysis – Outlier Analysis.

References

1. Adelchi Azzalini, Bruno Scapa, "Data Analysis and Data mining" , 2nd Ediiton, Oxford Univeristy Press Inc., 2012.
2. Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, 2011.
3. Alex Berson and Stephen J. Smith, "Data Warehousing, Data Mining & OLAP", 10th Edition, TataMc Graw Hill Edition , 2007.
4. G.K. Gupta, "Introduction to Data Mining with Case Studies", 1st Edition, Easter Economy Edition, PHI, 2006.

Student Activity

Case Study I: Analysis and Forecasting of House Price Indices

Case Study II: Customer Response Prediction and Profit Optimization

Adikavi Nannaya University
Subject: Data Science
Recommended Combination: BSc - Computers-Statistics-Data Science
Eligibility: Intermediate (MPC/MEC) or equivalent course

II YEAR III SEMESTER

DATA MINIG AND DATA ANALYSIS LAB

Objectives

- To Analyze the data using statistical methods
- To understand and demonstrate data mining

List of Experiments

1. Data Analysis – Getting to know the Data (Using ORANGE WEKA)
    - Parametric – Means . T-Test, Correlation
    - Prediction for numerical outcomes – Linear regression
    - Correlation analysis
    - Preparing data for analysis
        - Pre-Processing techniques

2. Data Mining (Using ORANGE WEKA or any source data mining tool)
    - Implement clustering algorithm
    - Implement classification using
        - Decision tree
        - Back Propagation
    - Visualization methods

Outcome

- Use Statistical techniques to carry out the analysis of data
- Gain hands-on skills and experience on data mining tools.

II YEAR IV SEMESTER

MULTIVARIATE TECHNIQUE FOR DATA ANALYSIS

Objective

The Objective of this course is to introduce the students into the field of Multivariate Techniques for analyzing large volumes of data and to take decisions based on inference drawn.

Outcomes

- Data characteristics and form of Distribution of the Data Structures.
- Understanding the usage of multivariate techniques for the problem under the consideration.
- For drawing valid inferences and to plan for future investigation.

Unit-I : Introduction to Multivariate Analysis

Meaning of Multivariate Analysis, Measurements Scales – Metric measurement scales and Non-Metric measurement scales, classification of multivariate techniques (Dependence Techniques and Inter-dependence Techniques), Applications of Multivariate Techniques in different disciplines.

Unit-II : Factor Analysis

Factor Analysis: Meaning, objectives and Assumptions, Designing a factor analysis, Deriving factors and assessing overall factors, Interpreting the factors and validation of factor analysis.

Unit-III: Cluster Analysis

Cluster Analysis: Objectives and Assumptions, Research design in cluster analysis, Deriving clusters and assessing overall fit (Hierarchical Methods, Non Hierarchical Methods and Combinations), Interpretation of clusters and validation of profiling of the clusters.

Unit-IV: Discriminate Analysis

Discriminate Analysis – Concept, objective and applications, Procedure for conducting discriminate analysis, Stepwise discriminate analysis and Mahalanobis procedure. Logit model.

Unit-V: Linear Programming

Linear Programming problem – Formulation, graphical method, simplex method. Integer Programming. Transportation and Assignment problem.

References

1. Joseph F Hair, William C Black etal, "Multivariate Data Analysis", Pearson Education, 7th edition, 2013.
2. T.W Anderson, " An introduction to Multivariate Statistical Analysis, 3rd Edition", Wiley 2003.
3. William r Dillon, John Wiley & Sons, "Multivariate Analysis Methods and Applications", Wiley, 1984.
4. Naresh K Malhotra, Satyabhusan Dash, "Marketing Research Anapplied Orientation", Pearson, 2011.
5. Hamdy A Taha, "Operations Research", Pearson, 2012.
6. S R Yaday, A K Malik, "Operations Research", Oxford, 2014.

II YEAR IV SEMESTER

MULTIVARIATE TECHNIQUE FOR DATA ANALYSIS Using 'R' LAB

1. Navigating the basic operating environment of 'R'
2. Importing network data.
3. Creating and manipulating network objects.
4. Plotting Network Graphs.
5. Network Descriptive Statistics.
6. Hypothesis Testing.

III YEAR V SEMESTER

BIG DATA TECHNOLOGY

Objective

This course provides practical foundation level training that enables immediate and effective participation in big data projects. The course provides grounding in basic and advanced methods to big data technology and tools, including map Reduce and Hadoop and its ecosystem.

Outcome

1. Learn tips and tricks for Big Data use cases and solutions.
2. Learn to build and maintain reliable, scalable, distributed systems with Apache Hadoop.
3. Able to apply Hadoop ecosystem components.

Unit-I

Introduction –distributed file system – Big Data and its importance, Four Vs, Drivers for Big data, Big data analytics, Big data applications.

Unit-II

Big Data – Apache Hadoop & Hadoop Ecosystem – Moving Data in and out of Hadoop – Understanding inputs and outputs of Map reduce- Data Serialization.

Unit-III

Introduction –distributed file system-algorithms using map reduce, Matrix – Vector Multiplication by Map Reduce – Hadoop – Understanding the Map Reduce architecture – Writing Hadoop Map Reduce Programs – Loading data into HDFS – Executing the MAP phase – Shuffling and sorting – Reducing phase execution.

Unit-IV

Hadoop Architecture, Hadoop Storage : HDFS, Common Hadoop Shell Commands, Anatomy of File Write and Read., NameNode, Secondary NameNode, and DataNode, Hadoop Map Reduce paradigm, Map and Reduce tasks, Job, Task trackers –Cluster Setup – SSH & Hadoop Configuration –HDFS Administering – Monitoring & Maintenance.

Unit-V

Hadoop ecosystem components – Schedulers- Fair and Capacity, Hadoop 2.0 New Features – NameNode High Availability, HDFS Federation, MRv2, YARN, Running MRv1 in YARN.

References

1. Boris lublinsky, Kevin t. Smith Alexey Yakubovich, “Professional Hadoop Solutions”. Wiley, ISBN : 9788126551071, 2015.
2. Chris Eaton, Dirk Deroos et al., “Understanding Big Data”, McGraw Hill , 201.
3. Tom White, “HADOOP” : The definitive Guide”, O Reilly 2012.

Student Activity:

Case Study I: Centers for Medicare & Medicaid Services: The Integrity of Healthcare Data and Secure Payment Processing.

Case Study II: Hadoop and the Data Warehouse: Competitive or Complementary

III YEAR V SEMESTER

BIG DATA TECHNOLOGY Through Hadoop LAB -I

1. Implement the following Data Structures in Java
   a) Linked Lists
   b) Stacks
   c) Queues
   d) Set
   e) Map

2. (i) Perform setting up and Installing Hadoop in its three operating modes:
   Standalone
   Pseudo distributed
   Fully distributed
   (ii) Use web based tools to monitor your Hadoop setup.

3. Implement the following file management tasks in Hadoop:
   Adding files and directories
   Retrieving files
   Deleting files

III YEAR V SEMESTER

BIG DATA ACQUISITION

Objective

- To Understand the complexity and volume of Big Data and their challenges
- To analyse the various methods of data collection.
- To comprehend the necessity for pre-processing Big Data and their issues

Outcome

- Identify the various sources of Big Data
- Design new algorithms for collecting Big Data from various sources
- Design algorithms for pre-processing Big Data other than the traditional approaches
- Design methodologies to extract data from structured and un-structured data for analytics

Unit- I

INTRODUCTION TO BIG DATA ACQUISITION: Big data framework – fundamental concepts of Big Data Management and analytics – Current challenges and trends in Big Data Acquisition.

Unit-II

DATA COLLECTION AND TRANSMISSION: Bid data collection – Strategies – Types of Data Sources – Structured Vs Unstructured data – ELT vs ETL – storage infrastructure requirements – Collection methods – Log files – sensors – Methods for acquiring network data (Libcap-based and zero-copy packet capture technology) – Specialized network monitoring softwares (Wireshark, Smartsniff and Winnetcap) – Mobile equipments – Transmission methods – Issues.

Unit- III

DATA PRE-PROCESSING: Data pre-processing overview-Sampling- Missing Values – Outlier Detection and Treatment – Standardizing Data – Categorization – Weights of Evidence Coding – Variable Selection and Segmentation

Unit-IV

DATA ANALYTICS : Predictive Analytics (Regression, Decision Tree, Neural Networks) – Descriptive Analytics (Association Rules, Sequence Rules), Survival Analysis (Survival Analysis Measurements, Kaplan Meir Analysis, Parametric Survival Analysis) – Social Network Analytics (Social Network Learning – Relational Neighbour Classification).

Unit-V

BIG DATA PRIVACY AND APPLICATIONS: Data Masking – Privately identified Information (PII) – Privacy preservation in Big Data – Popular Big Data Techniques and

tools – Map Reduce paradigm and the Hadoop system – Applications – Social Media Analytics – Recommender Systems – Fraud Detection.

References

1. Bart Baesens, "Analytics in a Big Data World: The Essential Guide to Data Science and its Applications', John Wiley & Sons, 2014.
2. Min Chen. Shiwen Mao, Yin Zhang. Victor CM Leung, Big Data: Related Technologies, Challenges and Future Prospects, Springer, 2014.
3. Michael Minelli, Michele Chambers Ambiga Dhiraj, "Big Data, Big Analytics : Emerging Business Intelligence and Analytic Trends", John Wiley & Sons, 2013.
4. Raj. Pethuru " Handbook of Research on Cloud Infrastructures for Big Data Analytics", IGI Global.

Student Activity:

Case study I: "BankAmeriDeals" provides cash-back offers to credit and debit-card customers based upon analyses of their prior purchases.

Case Study II: ***GOOGLE:*** Working with the U.S. Centers for Disease Control, tracks when users are inputting search terms related to flu topics, to help predict which regions may experience outbreaks.

III YEAR V SEMESTER

BIG DATA TECHNOLOGY Through Hadoop LAB –II

1. Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm

2. Write a Map Reduce program that mines weather data.
   Weather sensors collecting data every hour at many locations across the globe gather a large volume of log data, which is a good candidate for analysis with MapReduce, since it is semi structured and record-oriented.

3. Implement Matrix Multiplication program with Hadoop Map Reduce.

4. Install and Run Pig then write Latin scripts to sort, group, join, project, and filter your data.

5. Install and Run Hive then use Hive to create, alter, and drop databases, tables, views, functions, and indexes

Adikavi Nannaya University
Subject: Data Science
Recommended Combination: BSc - Computers-Statistics-Data Science
Eligibility: Intermediate (MPC/MEC) or equivalent course

III YEAR VI SEMESTER

Elective-I(A)

JAVA PROGRAMMING FOR DATA ANALYTICS

Objective

The rate in which data is exponentially growing has led to the evolvement of many technologies to better utilize this data for timely and accurate decision making. Such data with huge variety. Volume and velocity is coined as big data. The big data platform such as Hadoop is programmed in Java. This course aims at discussing the technical concepts which are the basic building blocks for most of the big data platforms.

Outcome

1. Understanding basic network and distributed programming.
2. Constructing a real world application with data storage and retrieval
3. Leveraging the benefits of reusable components
4. Analyzing basic file modes and operations
5. Applying Map Reduce paradigm to solve problems

Unit-I

NETWORK PROGRAMMING & DISTRIBUTED OBJECTS: Connecting to a Server – Implementing Servers and Clients – Advanced Socket Programming – Intel Address – URL Connections – RMI Programming.

Unit-II

CONNECTING TO DATABASE: The design of JDBC – Basic Concepts – Executing Queries – Prepared Statements – Result Sets – Metadata – Transactions.

Unit-III

JAVABEANS: The bean – Writing Process – Using Beans to Build Application – Bean Property Types- Property Editors – Customizers.

Unit-IV

STREAMS AND FILES: Streams – Text Input and Output – Reading and Writing Binary Data – Zip Archives – Object Streams and Serialization – Memory Mapped Files.

Unit-V

PROGRAMMING MAP REDUCE: Map Reduce program in Java – Map Reduce API – Programming Examples – Combiner Functions- Distributed Map Reduce Job.

References

1. White. "Hadoop: The Definitive Guide". Third Edition – 2012 – O'Reilly – ISBN: 9789350237564.
2. Cay S.Horstmann. Gary Cornell. "Core Java™ 2: Volume II-Advanced Features". Prentice Hall. 9th edition. ISBN: 978-0137081608.
3. Jean Dollimore. Tim Kindberg. George Coulouris. "Distributed Systems Concepts and Design". 4th Edition. Jun 2005. Hardback. 944 pages. ISBN: 9780321263544.
4. Y. Daniel Liang. Introduction to Java Programming. Tenth Edition. Pearson, 2015.

Student Activity:

Case Study I: Create a school Data Base.
Case Study II: How to install a JAVABEANS
Case Study III: Analyze life-threatening risks

II YEAR VI SEMESTER

Elective-I(A) Lab

JAVA PROGRAMMING FOR DATA ANALYTICS Lab

1. Write a Java Mapper program for Word Count.
2. Write a Java reducer program for Word Count.
3. Implement Java program for

   Word Count: given a collection of text documents, find the number of occurrences of each word in the collection.
4. Wirte a Java Mapper program for

   Max Temp: given a file containing temperature measurements, find the maximum temperature recording per year.
5. Implement Java program reducer for

   Max Temp: given a file containing temperature measurements, find the maximum temperature recording per year.
6. Write a Java program for

   Max Temp: given a file containing temperature measurements, find the maximum temperature recording per year.

III YEAR VI SEMESTER

Elective-I(B)
PYTHON PROGRAMMING FOR DATA ANALYTICS

Objective

Data. Which is available in abundance and in accessible forms, if analysed in an efficient manner unfolds many patterns and promising solutions. Data has to be pre-processed, converted to required format and fed to appropriately chosen algorithm to yield better results. This course aims at applying such techniques to raw data. Using Python, to arrive at meaningful result.

Outcome

1. Understanding the basic concepts of Python
2. Preparing and pre-processing data
3. Understanding the data aggregation and grouping concepts
4. Leveraging web scraping
5. Visualizing the results of analytics effectively

Unit-I

PYTHON CONCEPTS, DATA STRUCTURES CLASSES: Interpreter – Program Execution – Statements- Expressions – Flow Controls – Functions – Numeric Types – Sequences – Strings, Tuples, Lists and – Class Definition – Constructors – Inheritance – Overloading – Text & Binary Files – Reading and Writing.

Unit-II

DATA WRANGLING: Combining and Merging Data Sets – Reshaping and Pivoting – Data Transformation – String Manipulation, Regular Expressions.

Unit-III

DATA AGGREGATION, GROUP OPERATIONS , TIMESERIES : GroupBy Mechanics – Data Aggregation – GroupWise Operations and Transformations – Pivot Tables and Cross Tabulations - Date and Time Date Type tools – Time Series Basics – Data Ranges , Frequencies and Shifting.

Unit-IV

WEB SCRAPING: Data Acquisition by Scraping Web applications – Submitting a form – Fetching Web pages – Downloading Web pages through form submissions – CSS Selectors.

Unit-V

VISUALIZAITON IN PYTHON: Matplotlib package – Plotting Graphs – Controlling Graph – Adding Text – More Graph Types – Getting and Setting values – Patches.

References

1. Mark Lutz. "Programming Python". O'Reilly Media, 4th edition, 2010.
2. Mark Lutz. "Learning Python". O'Reilly Media, 5th edition, 2013
3. Tim Hall and J-P Stacey. "Python 3 for Absolute Beginners". Apress. 1st edition, 2009
4. Magnus Lie Hetland. "Beginning Python: From Novice to Progessional". Aprèss. Second Edition, 2005.
5. Shai Vaingast. "Beginning Python Visualizing Crafting Visual Transformation Scripts". Après. 2nd edition. 2014.
6. Wes Mc Kinney, "Python for Data Analysis". O'Reilly Media, 2012.
7. White. "Hadoop: The Definitive Guide". Third Edition – O'Reilly, 2012.
8. Brandon Rhodes and John Goerzen. "Foundations of Python Network Programming: The Comprehensive Guide to Building Network Application with Python". Apress, Second Edition, 2010

III YEAR VI SEMESTER

Elective-I(B)

PYTHON PROGRAMMING FOR DATA ANALYTICS Lab

1. Write a Python Mapper program for Word Count.
2. Write a Python reducer program for Word Count.
3. Implement Python program for

   Word Count: given a collection of text documents, find the number of occurrences of each word in the collection.
4. Write a Python Mapper program for

   Max Temp: given a file containing temperature measurements, find the maximum temperature recording per year.
5. Implement Python program reducer for

   Max Temp: given a file containing temperature measurements, find the maximum temperature recording per year.
6. Write a Java program for

   Max Temp: given a file containing temperature measurements, find the maximum temperature recording per year.

III YEAR VI SEMESTER

Elective –II (Cluster A)

1.SPARK PROGRAMMING

Objective

1. Student can understand the depth of fundamental concepts, design principles, and system architectures of Apache Spark.
2. Students can understand and learn processing and analysing big data sets.

Outcome

- Learn to use the Apache Spark framework for the purpose of Such Big Data Management and analysis..
- Focus on Fundamental Concepts.
- Focus on Architecture and, interfaces and various components.

Unit-I

Introduction to big data, properties of data different processing frame works. Introduction to Hadoop and Spark.

Unit-II

Programming with Scale, data types, conditional and control statements, functional & non functional programming.

Unit-III

Introduction to Spark, RPD supporting operators. Architecture of Spark, Working with data sets.

Unit-IV

Spark Libraries, Creating machine learning and predictive models using MLib.

Unit-V

Processing Streaming data and graph structured data using spark streaming and Graph.

REFERENCES

1. **Learning Spark: Lightning-Fast Big Data Analysis by** Holden Karau, Andy Konwinski, Patrick Wendell, O'Reilly Publishers
2. **Advanced Analytics with Spark: Patterns for Learning from Data at Scale** By Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills

3. **Spark in Action by** Petar Zecevic, Marko Bonaci Manning Publications Company, 2016

III YEAR VI SEMESTER

Elective –II (Cluster A)

1.SPARK PROGRAMMING LAB
(Using either Python, Scala (or) Java)

1.
   (a) Scala Installation in Windows platform
   (b) Write a program to implement Arithematic operators.
   (c) Write a profram to find biggest of two numbers

2. Write a program to implement
   Word Count: given a collection of text documents, find the number of occurrences of each word in the collection.
3. Write a program to implement

   Max Temp: given a file containing temperature measurements, find the maximum temperature recording per year

4. Write a program to implement Pie estimation

III YEAR VI SEMESTER

Elective –II (Cluster A)

2.MARKETING ANALYTICS

Objective

The objective of this course is to provide through knowledge required to address fundamental marketing decision problems. It will also train to view marketing process and relationships systematically and analytically. The techniques discussed in this course are useful in market segmentation, targeting, and mapping market structure and product design.

Outcome

1. Learn how to tap a simple and cost-effective tool. Microsoft Excel, to solve specific business problems using powerful analytic techniques.
2. Helps to forecast sales and improve response rates for marketing campaigns.
3. Explore how to optimize price points for produces and service, optimize store layouts, and improve online advertising.

Unit-I

MARKETING DATA SUMMARIZATION : Slicing and Dicing Marketing Data with Pivot Tables – Using Excel Charts to Summarize Marketing Data- Using Excel Functions to Summarize Marketing Data.

Unit-II

FORECASTING TECHNIQUES: Simple Linear Regression and Correlation – Using Multiple Regression to Forecast Sales- Forecasting in the Presence of special Events- Modelling Trend and Seasonality- Ratio to Moving Average Forecasting Method – Winter's Method – Using Neural Networks to Forecast Sales.

Unit-III

CUSTOMER NEEDS : Conjoint Analysis - Logistic Regression – Discrete Choice Analysis – Customer Value- Introduction to Customer Value , Benefits.

Unit-IV

MARKET SEGEMENTATION : Cluster Analysis – User Based Collaborative Filtering – Collaborative Filtering – Using Classification Trees for Segmentation.

Unit-V

RETAILING AND MARKET RESERCH TOOLS: Retailing – Introduction to retailing, Market Basket Analysis and Lift – Marketing Research Tools – Principal Components Analysis.

References

1. Wayne.L.Winston, " Marketing Analysis: Data driven Techniques with MS-Excel", Wiley, 1st ed. 2014.
2. Stephan Sorger. "Marketing Analytics: Strategic models and Metrics ", Create Space Independent Publishing Platform 1st ed., 2013.

Student Activity:

Case Study I: How one company's thought leader ship content in driving new business exposure?

Case Study II: How IBM offset the impact of a down economy on event attendance.

III YEAR VI SEMESTER

Elective –II (Cluster A)

2.Mongo DB Lab

1. Learn the basics of Mongo DB.
2. Installation steps for Mongo DB.
3. Use the following commands
   (a) DATABASE_NAME.
   (b) Drop Database( )
   (c) create Collection
   (d) insert( )
   (e) drop( )
   (f) find( )
4. Differentiate between SQL and Mongo DB.
5. Write a program to update a collection in Mongo DB
6. Write a program to remove specific document from Mongo DB.
7. Write a program to implement aggregate function in Mongo DB
8. Apply the Map reduce operation to find total salary of each department assuming employee collection is already exists.

III YEAR VI SEMESTER

Elective –II (Cluster A)

3. SOCIAL NETWORK ANALYTICS

Objective

This course will be used for social network analysts, both its theory and computational tools, to make sense of the social and information networks that have been fuelled and rendered accessible by the internet.

Outcome

1. Analyze the structure and evolution of networks.
2. Able to knowledge from disciplines as diverse as sociology, mathematics, computer science.
3. Understand the Online interactive demonstrations and hands-on analysis of real-world data sets.

Unit-I

INTRODUCTION : Overview; Social network data-Formal methods – Paths and Connectivity – Graphs to represent social relations – Working with network data – Network Datasets – Strong and Weak ties – Closure, Structural Holes, and Social Capital.

Unit-II

SOCIAL INFLUENCE: Homophile; Mechanisms Underlying Homophile, Selection and Social Influence. Affiliation, Tracking Link Formation in Online Data, Spatial Model of Segregation – Positive and Negative Relationship – Structural Balance – Applications of Structural Balance. Weaker Form of Structural Balance.

Unit-III

INFORMATION NETWORKS AND THE WORLD WIDE WEB : The structure of the Web- World Wide Web – Information Networks, Hypertext and Associative Memory – Web as a Directed Graph. Bow-Tie Structure of the Web-Link Analysis and Web Search – Searching the web; Ranking Link Analysis using Hubs and Authorities – Page Link Analysis in Modern Web Search, Applications, Special Analysis Random Walks and Web Search.

Unit-IV

SOCIAL NETWORK MINING: Clustering of Social Network graphs; Betweens, Girvan Newman algorithm – Discovery of Communities – Cliques and Bipartite graph-Graph partitioning methods – Matrices-Eigen values – simrank..

Unit-V

NETWORK DYNAMICS: Cascading Behaviour in Networks: Diffusion in Networks, Modelling Diffusion- Cascades and Cluster, Thresholds, Extensions of the Basic Cascade Model – Six Degrees of Separation – Structure and Randomness, Decentralized Search – Empirical Analysis and Generalized Models – Analysis of Decentralized Search.

References

1. Easley and Kleinberg, "Networks, Crowds, and Markets: Reasoning about a highly connected World", Cambridge Univ. Press. 2010..
2. Robert A. Hamneman and Mark Riddle. " Introduction to social network methods", University of California, 2005.
3. Jure Leskovec. Stanford Univ. Anand Rajaraman, Milliway Labs, Jeffery D. Ullman "Mining of massive Datasets", Cambridge University Press. 2 edition , 2014.
4. Wasseman, S. & Faust, K. "Social Network Analysis:, Methods and Applications", Cambridge University Press., 1st edition, 1994.
5. Borgatti, S.P. Everest, M.G. & Johnson, J.C., " Analyzing Social networks", SAGE Publications Ltd., 1 edition, 2013.
6. John Scott, , Social Network Analysis: A Handbook " – SAGE Publications Ltd., 2nd edition, 2000.

Student Activity:

Case Study I: How Twitter Will Work.
Case Study II: Write a report on some of social networks in our day to day life

III YEAR VI SEMESTER

Elective –II (Cluster 1)

3.SOCIAL NETWORK ANALYTICS through 'R' Lab

1. Navigating the basic operating environment of 'R'
2. Importing network data.
3. Creating and manipulating network objects.
4. Plotting Network Graphs.
5. Network Descriptive Statistics.
6. Hypothesis Testing.

III YEAR VI SEMESTER

Elective –II (Cluster B)

1.DATA & INFORMATION SECURITY

Unit -I

Overview of Security: Protection versus security; aspects of security – data integrity, data availability, privacy; security problems, user authentication, Orange Book.

Unit -II

Security Threats: Program threats, worms, viruses, Trojan horse, trap door, stack and buffer overflow; system threats- intrudes; communication threats- tapping and piracy.

Unit -III

Cryptography: Substitution, transposition ciphers, symmetric-key algorithms – Data Encryption Standard, advanced encryption standards, public key encryption – RSA; Diffie-Hellman key exchange, ECC cryptography, Message Authentication – MAC, has functions.

Unit -IV

Digital Signatures: Symmetric key signatures, public key signatures, message digests, public key infrastructures.

Unit -V

Security Mechanism: Intrusion detection, auditing and logging, tripwire, system –call monitoring.

References

1. W. Stallings, Cryptography and Network Security Principles and Practices (4th ed.,), Prentice – Hall of India, 2006.
2. C. Pfleeger and SL Pfleeger, Security in Computing (3rd ed.,), Prentice- Hall of India, 2007.
3. D. Gollamann, Computer Security, John Wiley and Sons, Ny, 2002.
4. J. Piwprzyk, T. Hardjono and J. Seberry, Fundamentals of Computer Security, Springer- Verlag Berling, 2003.
5. J.M. Kizza, Computer Netwrok Security, Springer, 2007
6. M. Merkow and J. Breithaupt, Information Security: Principles and Practices, Pearson Education, 2006.

Student Activity

Case Study I: Transform Data from one format to another format using Cryptography.
Case Study II: How mails are hacked.

Adikavi Nannaya University
Subject: Data Science
Recommended Combination: BSc - Computers-Statistics-Data Science
Eligibility: Intermediate (MPC/MEC) or equivalent course

III YEAR VI SEMESTER

Elective –II (Cluster B)

1.DATA & INFORMATION SECURITY Through Python Lab

1. Implement Ceiser Cipher encryption in Python.
2. Implement Ceiser Cipher decryption in Python.
3. Implement Transposition technique encryption in Python.
4. Implement Substitution cipher encryption in Python.
5. Implement Substitution cipher decryption in Python.
6. Implement One time Pad cipher in Python.
7. Implement DES encryption in Python.
8. Implement RSA Public Key encryption in Python.

III YEAR VI SEMESTER

Elective –II (Cluster B)

2.SPARK PROGRAMMING

Objective

1. Student can understand the depth of fundamental concepts, design principles, and system architectures of Apache Spark.
2. Students can understand and learn processing and analysing big data sets.

Outcome

- Learn to use the Apache Spark framework for the purpose of Such Big Data Management and analysis..
- Focus on Fundamental Concepts.
- Focus on Architecture and, interfaces and various components.

Unit-I

Introduction to big data, properties of data different processing frame works. Introduction to Hadoop and Spark.

Unit-II

Programming with Scale, data types, conditional and control statements, functional & non functional programming.

Unit-III

Introduction to Spark, RPD supporting operators. Architecture of Spark, Working with data sets.

Unit-IV

Spark Libraries, Creating machine learning and predictive models using MLib.

Unit-V

Processing Streaming data and graph structured data using spark streaming and Graph.

REFERENCES

1. **Learning Spark: Lightning-Fast Big Data Analysis by** Holden Karau, Andy Konwinski, Patrick Wendell, O'Reilly Publishers
2. **Advanced Analytics with Spark: Patterns for Learning from Data at Scale** By Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills
3. **Spark in Action by** Petar Zecevic, Marko Bonaci Manning Publications Company, 2016

III YEAR VI SEMESTER

Elective –II (Cluster B)

2.SPARK PROGRAMMING LAB

(Using either Python, Scala (or) Java)

1.
   (a) Scala Installation in Windows platform
   (b) Write a program to implement Arithmetic operators.
   (c) Write a program to find biggest of two numbers

2. Write a program to implement
   Word Count: given a collection of text documents, find the number of occurrences of each word in the collection.
3. Write a program to implement
   Max Temp: given a file containing temperature measurements, find the maximum temperature recording per year
4. Write a program to implement Pie estimation

Adikavi Nannaya University
Subject: Data Science
Recommended Combination: BSc - Computers-Statistics-Data Science
Eligibility: Intermediate (MPC/MEC) or equivalent course

III YEAR VI SEMESTER

Elective –II (Cluster B)

3.BIG DATA SECURITY

Objective

With the data generated from electronic devices growing exponentially, the need to analysed data on a large scale is important. Such data are of many types like financial, personal etc. Big data environment also created significant security challenges. When trying to make quick decisions. Data breach poses many complications. This course aims at introducing concepts related to big data security.

Outcome

1. Understanding significance of privacy, ethics in big data environment.
2. Analyzing the steps to secure big data.
3. Analyzing data security and event logging.

Unit-I

BIG DATA PRIVACY, ETHICS AND SECURITY: Privacy- Re identification of Anonymous people – Why Big Data Privacy is self regulating? – Ethics – Ownership – Ethical Guidelines - Big Data Security – Organizational Security.

Unit-II

SECUTIY, COMPLIANCE, AUDITING, AND PROTECTION: Steps to secure big data – Classifying Data – Protecting – Big Data Compliance – Intellectual Configuration..

Unit-III

HADOOP SECURITY DESIGN: Kerberos – Default Hadoop Model Without security-Hadoop Kerberos Security Implementation & Configuration.

Unit-IV

HADOOP ECOSYSTEM SECURITY: Configuring Kerberos for Hadoop ecosystem components – Pig. Hive. Oozie, Flume, HBase, Scoop.

Unit-V

HADOOP ECOSYSTEM SECURITY: Integrating Hadoop with Enterprise Security Systems- Securing Sensitive Data in Hadoop – SIEM System – Setting up audit logging in hadoop cluster.

References

1. Mark Van Rijmenam, "Think Bigger: Developing a successful Big Data Strategy for your Business", Amazon, 1 edition , 2014.
2. Frank Ohihorst John Wiley & Sons, " Big Data Analytics: Turning Big Data into Big Money", John Wiley & Sons 2013.
3. Sherif Sakr, " Large Scale and Big Data: Processing and Management", CRC Press. 2014.
4. Sudeesh Narayanan, "Securing Hadoop", Pacjt Publishing – 2013.
5. Ben Spivey, Joe Echeverria. "Hadoop Security Protecting Your Big Data Problem", O'Reilly Media , 2015.
6. Top Tips for Securing Big Data Environments : e-book

III YEAR VI SEMESTER

Elective –II (Cluster B)

3.Mongo DB Lab

1. Learn the basics of Mongo DB.
2. Installation steps for Mongo DB.
3. Use the following commands
   (g) DATABASE_NAME.
   (h) Drop Database( )
   (i) create Collection
   (j) insert( )
   (k) drop( )
   (l) find( )
4. Differentiate between SQL and Mongo DB.
5. Write a program to update a collection in Mongo DB
6. Write a program to remove specific document from Mongo DB.
7. Write a program to implement aggregate function in Mongo DB
8. Apply the Map reduce operation to find total salary of each department assuming employee collection is already exists.

PROJECT & VIVA-VOCE

The objective of the project is to motivate them to work in emerging/latest technologies, help the students to develop ability, to apply theoretical and practical tools/techniques to solve real life problems related to industry, academic institutions and research laboratories.

The project is of 2 hours/week for one (semester VI) semester duration and a student is expected to do planning, analyzing, designing, coding, and implementing the project. The initiation of project should be with the project proposal. The synopsis approval will be given by the project guides.

The project proposal should include the following:

- Title
- Objectives
- Input and output
- Details of modules and process logic
- Limitations of the project
- Tools/platforms, Languages to be used
- Scope of future application

The Project work should be either an individual one or a group of not more than three members and submit a project report at the end of the semester. The students shall defend their dissertation in front of experts during viva-voce examinations.