**University of Central Punjab**

**Faculty of Information Technology**



**Programming for Big Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Course Code** | CSDS4423 | | | |
| **Course Title** | Programming for Big Data | | | |
| **Credit Hours** | 3 | | | |
| **Assessment Instruments with Weights** (quizzes, midterms, final, assignments, presentation, lab work, etc.) | Assignments (13%)  Quizzes (12%)  Class Participation (5%)  Mid Term (30%)  Final Term (40%) | | | |
| **Course Instructors** | Mr. Saeed Iqbal, Mr. Saad Azhar Saeed | | | |
| **Course Coordinator** | Mr. Saeed Iqbal | | | |
| **Office Hours** | To be announced later. | | | |
| **Plagiarism Policy** | **All the parties involved in first cheating case in Midterm or Final term will be awarded ‘F’ for the course.** | | | |
| **Textbook** | * Mining of Massive Datasets * Data Analysis with open source tools * Learning Apache Spark with Python | | | |
| **Reference Books** | * Python for Data Analysis * Python for Programmers * Big Data, Mining, and Analytics | | | |
| **Course Goals** | This course is for students who have some programming and database experience. The objective of this course is to give students some experience in data analysis and developing applications that utilize the vast amount of data that is available to general public to create programs that provides information used in improving the standard of application performance. Discovering how the efficiency of applications can be improved by understanding the data. | | | |
| **Learning Outcomes** | 1. Understanding of Python 2. Data manipulation using python 3. Hands on practice with python libraries (Numpy, Pandas and Matplot). 4. Data Visualization 5. Introduction Big Data 6. Introduction to Spark 7. Introduction to RDDs 8. Introduction to DataFrames 9. Advanced Spark Topics 10. Introduction to Spark MLlib | | | |
| **Topics Covered in the Course, with Number of Lectures on Each Topic** | Attached | | | |
| **Lab Projects** | Yes | | | |
| **Theory & Lab Exam** | Mid Term: 1.5 Hours  Final Term: 3 Hours | | | |
| **Class Time Spent on** (in credit hours) | **Theory** | **Problem Analysis** | **Solution Design** | **Labs** |
| 1 | 1 | 1 | 1 |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| **Week** | **Topics Covered** | **Instruments** |
| 1 | * Discussion on Python and its market position. * Motivation regarding learning aspects of this course * Setting up environment for Python. * Installation of Anaconda * What is Big Data? * Characteristics of Big Data * What are the V’s of Big Data? * The Impact of Big Data * Big Data - Beyond the Hype, Big Data Examples, Sources of Big Data * Big Data Adoption, The Big Data and Data Science * The Big Data Platform, Big Data and Data Science. Skills for Data Scientists |  |
| 2 | * Types of IDE(s) and WIDE that will be used in the duration of this course. e.g. Spyder, Jupyter etc * Hello World Program “Print Command” * Keyword Types * Expressions and Variables * Input Method * Conditions and Branching * Loops |  |
| 3 | * String Operations * Lists and Tuples * Sets * Dictionaries |  |
| 4 | * Reading and Writing files * Functions * Objects and Classes |  |
| 5 | * Introduction with Numpy * Numpy one dimensional Array * Numpy two dimensional Array * Numpy Array Operations |  |
| 6 | * Descriptive Statistics * Data Manipulation * Data Wrangling |  |
| 7 | * Working with Pandas * Descriptive Statistics with Pandas * Group by with Python * Data Manipulation with Pandas |  |
| 8 | * Data Wrangling with Pandas * Discussion regarding exam |  |
|  |  |  |
| 9 | * Introduction to Matplotlib * Basic Plotting with Matplotlib * Line Plots * Area Plots * Histograms |  |
| 10 | * Bar Charts * Pie Charts * Box Plots * Scatter Plots * Word Cloud |  |
| 11 | * What is Spark and what is its purpose? * Components of the Spark unified stack * Resilient Distributed Dataset (RDD) * Scala and Python overview |  |
| 12 | * Understand how to create parallelized collections and external datasets * Work with Resilient Distributed Dataset (RDD) operations * Utilize shared variables and key-value pairs |  |
| 13 | * Describe and run some Spark examples * Pass functions to Spark * Create and run a Spark standalone application |  |
| 14 & 15 | * Understand and use the various Spark libraries |  |
|  |  |  |