KAKATIYA UNIVERSITY WARANGAL

Under Graduate Courses (Under CBCS AY: 2021-2022 on words)

B.Sc. DATA SCIENCE

II Year: Semester-III

Paper – III: Data Engineering with Python

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

Objective: The main objective of this course is to teach how to extract raw data, clean the data, perform transformations on data, load data and visualize the data

Outcomes:

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

- Handle different types of files and work with text data
- Use regular expression operations
- Use relational databases via SQL
- Use tabular numeric data
- Use the data structures: data series and frames
- Use PyPlot for visualization

Unit – I

Data Science: Data Analysis Sequence, Data Acquisition Pipeline, Report Structure [Reference 1(Chapter 1-Unit1 to Unit 3)]]
Files and Working with Text Data: Types of Files, Creating and Reading Text Data, File Methods to Read and Write Data, Reading and Writing Binary Files, The Pickle Module, Reading and Writing CSV Files, Python os and os. Path Modules. [Reference 2, Chapter 9)]
Working with Text Data: JSON and XML in Python[Reference 2, Section12.2]

Unit – II

Working with Text Data: Processing HTML Files, Processing Texts in Natural Languages [Reference 1(Chapter3 –Unit 13, and Unit16)
Regular Expression Operations: Using Special Characters, Regular Expression Methods, Named Groups in Python Regular Expressions, Regular Expression with *glob* Module [Reference 2-Chapter 10]

Unit – III

Working with Databases: Setting Up a MySQL Database, Using a MySQL Database: Command Line, Using a MySQL Database, Taming Document Stores: MongoDB [Reference 1(Chapter4-Unit17toUnit20)]
Working with Tabular Numeric Data(Numpy with Python): NumPy Arrays Creation Using *array()* Function, Array Attributes, NumPy Arrays Creation with Initial Placeholder Content, Integer Indexing, Array Indexing, Boolean Array Indexing, Slicing and Iterating in Arrays, Basic Arithmetic Operations on NumPy Arrays, Mathematical Functions in NumPy, Changing the Shape of an Array, Stacking and Splitting of Arrays, Broadcasting in Arrays. [Reference 2: Section 12.3)]

Unit – IV

Working with Data Series and Frames: Pandas Data Structures, Reshaping Data, Handling Missing Data, Combining Data, Ordering and Describing Data, Transforming Data, Taming Pandas File I/O [Reference 1 (Chapter 6-Unit 31 to Unit 37)]

Plotting: Basic Plotting with PyPlot, Getting to Know Other Plot Types, Mastering Embellishments, Plotting with Pandas [Reference 1(Chapter8-Unit 41 to Unit 44)]

References:

1. Data Science Essentials in Python: Collect, Organize, Explore, Predict, Value. Dmitry Zinoriev,The Pragmatic Programmers LLC, 2016
2. Introduction to Python Programming**.** Gowrishankar S., Veena A. CRC Press, Taylor & Francis Group, 2019

Suggested Reading

3. Python for Everybody: Exploring Data Using Python 3. Charles R Severance, 2016
4. Python Data Analytics – Data Analysis and Science using Pandas, matplotlib and the Python Programming Language. Fabio Nelli, Apress, 2015
5. Website Scraping with Python. Using BeautifulSoup and Scrapy. GáborLászlóHajba, Apress, 2018
6. Machine Learning with Python Cookbook:.Practical Solutions from Preprocessing to Deep Learning. Chris Albon, O’Reilly 2018

Practical- 3: Data Engineering with Python (Lab)

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

Objective:

The main objective of this laboratory is to put into practice the ETL (extract, transform, load) pipeline which will extract raw data, clean the data, perform transformations on data, load data and visualize the data.

This requires mentoring by TCS.

Libraries

In this course students are expected to extract, transform and load input data that can be text files, CSV files, XML files, JSON, HTML files, SQL databases, NoSQL databases etc.,. For doing this, they should learn the following Python libraries/modules:
pandas, numpy, BeautifulSoup, pymysql, pymongo, nltk, matplotlib

Datasets

For this laboratory, appropriate publicly available datasets, can be studied and used. Example:
MNIST (http://yann.lecun.com/exdb/mnist/),
UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.html),
Kaggle (https://www.kaggle.com/datasets)
Twitter Data

Exercises

1. Write programs to parse text files, CSV, HTML, XML and JSON documents and extract relevant data. After retrieving data check any anomalies in the data, missing values etc.
2. Write programs for reading and writing binary files
3. Write programs for searching, splitting, and replacing strings based on pattern matching using regular expressions
4. Design a relational database for a small application and populate the database. Using SQL do the CRUD (create, read, update and delete) operations.
5. Create a Python MongoDB client using the Python module pymongo. Using a collection object practice functions for inserting, searching, removing, updating, replacing, and aggregating documents, as well as for creating indexes
6. Write programs to create numpy arrays of different shapes and from different sources, reshape and slice arrays, add array indexes, and apply arithmetic, logic, and aggregation functions to some or all array elements
7. Write programs to use the pandas data structures: Frames and series as storage containers and for a variety of data-wrangling operations, such as:
    - Single-level and hierarchical indexing
    - Handling missing data
    - Arithmetic and Boolean operations on entire columns and tables
    - Database-type operations (such as merging and aggregation)
    - Plotting individual columns and whole tables
    - Reading data from files and writing data to files