School of Computer Science and Engineering (SCOPE)
Winter Semester 2024-25

BCS206L – FOUNDATIONS OF DATA SCIENCE
Digital Assignment - 1

Course Mode	**Theory only**
Slot and Class Number	**Thoery : C2+TC2** (VL2024250501999)
Faculty In-Charge	**Prof. RA K SARAVANAGURU**
No. of Students	**70**
Date and Time	**9-February-2025 (Sunday)**

Main Focus :

1. **Data Collection from any two Internet sources through Web Scraping**
2. **Data Processing**
3. **Storing Data**
4. **Querying**

Web Scraping, Data Structuring, and Graph Database Querying

Objective:

The goal of this project is to collect data from two different sources, convert unstructured data into a structured format, and store it in a graph database for querying.

Guidelines:

1. **Data Collection (Web Scraping and API extraction)**

   i. Choose **two different but related sources** (websites, APIs, PDFs, or text files).
   ii. Extract at least three relevant attributes from each source.
   iii. If scraping a website, use BeautifulSoup, Scrapy (Python) or rvest (R).
   iv. If using an API, extract data using requests (Python) or httr (R).
   v. Convert extracted data into JSON or a DataFrame (Pandas in Python or Tidyverse in R).

2. **Data Processing (Unstructured to Structured Conversion)**

   i. Clean and transform raw data into a structured format (handling missing values, standardizing fields).
   ii. Merge data from both sources into a single structured dataset.

3. **Storing Data in a Graph Database**

   i. Choose a graph database such as Neo4j.
   ii. Define an appropriate graph schema (nodes, edges, relationships).
   iii. Insert structured data into the database using Py2neo (Python) or RNeo4j (R).

4. Querying the Graph Database

i. Write at least three queries using Cypher (Neo4j) to retrieve insights.
Example queries:
ii. Find all items (books, players, movies, etc.) and their relationships.
iii. Retrieve data for a specific entity based on an attribute.
iv. Analyze relationships (e.g., find authors who wrote multiple books).

Deliverables:

Python or R scripts for scraping, processing, storing, and querying.
Database schema design (graph structure with nodes and edges).
Query results (screenshots or CSV output).
A short report (7-9 pages) explaining the process and insights from the data.

Evaluation Criteria:

Comprehensiveness: Depth of data collection and the scope of wrangling tasks performed.
Accuracy: Correctness of the processed dataset and handling of errors.
Creativity: Innovation in feature enrichment and presentation.
Documentation: Clarity, organization, and completeness of the report
Technical Soundness: Use of appropriate tools and techniques for the tasks.