Toronto Metropolitan University

CIND 110
Data Organization for Data Analysts

Semi-Structured Data Processing Teamwork Assignment

Introduction

This assignment involves teamwork to process and analyze data represented in XML and JSON formats. The goal is to apply data handling, querying techniques, and information retrieval (IR) concepts. Students will explore XPath, XQuery, TF-IDF, and JSON CRUD operations. Teams will consist of three to four members and each student will be assigned a specific role: XML Data Engineer, JSON Data Analyst, or Information Retrieval Specialist.

Dataset

The dataset contains landmark information in XML and JSON formats:

- `landmarksdb.xml`: An XML dataset with details about landmarks and their attributes.
- **JSON Files:**
  - `countries.json`: Contains information about countries, including their names, populations, and official languages.
  - `heritage_status.json`: Describes various heritage statuses assigned to landmarks.
  - `visitor_groups.json`: Lists visitor groups, their sizes, and visit purposes by country.
  - `landmark_group_visits.json`: Records visit logs for different landmarks by specific visitor groups.
  - `landmarksdb.json`: A semi-structured dataset with countries, landmarks, heritage statuses, and visit data.

Team Roles and Responsibilities

Each role focuses on different aspects of data handling and retrieval:

1. XML Data Engineer

- Use XPATH or XQuery/FLWOR to:
  1. List all landmarks in France and the USA, their location details, and their country name.
  2. List the Total Number of Landmarks in Each Country
  3. Identify All Landmarks Established After 1900 That Have a Heritage Status
  4. Identify Unique Pairs of Landmarks Located in the Same Country and City
  5. Identify countries with landmarks that include the terms ”citadel”,”fort” or ”castle” in their description
- Submit code, results, and screenshots of queries and outputs.

2. JSON Data Analyst

- Use MongoDB queries to:
  1. List all landmarks in France and the USA, their location details, and their country name.
  2. List the Total Number of Landmarks in Each Country
  3. Identify All Landmarks Established After 1900 That Have a Heritage Status
  4. Identify Unique Pairs of Landmarks Located in the Same Country and City
  5. Identify countries with landmarks that include the terms ”citadel”,”fort” or ”castle” in their description
- Submit code, results, and screenshots of queries and outputs.

3. Information Retrieval Specialist

- First, extract the `LandmarkName` and `Description` attributes from the provided XML dataset into a CSV file using the BaseX application.
- Use RStudio to:
  1. Read the extracted CSV dataset and apply at least three text pre-processing techniques (e.g., lowercasing, stop-word removal, lemmatization).
  2. Create a TF-IDF matrix from the cleaned descriptions and analyze its sparsity.
  3. Use cosine similarity to find landmarks similar to the Big Ben landmark.
- Submit an R Markdown file with code, results, and explanations.

Submission Guidelines

Each team member must submit their part through the peerScholar platform:

- `XML_Data_Engineer_StudentName.DOCX`
- `JSON_Data_Analyst_StudentName.DOCX`
- `IR_Specialist_StudentName.HTML`

Reports must include:

- **Introduction**: Role description and objectives.
- **Task Execution**: Steps, code, and screenshots.
- **Assumptions and Challenges**: Mention difficulties and solutions.
- **Conclusion**: Summarize insights and findings.

Assignment Workflow

The assignment follows three phases on peerScholar:

- **Create Phase**:
  Submit initial work between `Wed, Nov 06, 12:01 AM` and `Thu, Nov 28, 11:59 PM`.
- **Assess Phase**:
  Review assigned submissions between `Fri, Nov 29, 12:01 AM` and `Tue, Dec 03, 11:59 PM`.
- **Reflect Phase**:
  Resubmit refined work between `Wed, Dec 04, 12:01 AM` and `Sun, Dec 08, 11:59 PM`.

Collaboration Guidelines

Collaboration is required during the `Create` and `Reflect` phases. In the `Reflect` Phase, besides working on submitting a refined work, students will evaluate their peers' contributions within the group. On the other hand, during the `Assess` Phase, each student will work independently to assess the work of two other groups.

To ensure progress, each group is encouraged to assign one specific role to each member at the start of the project. Groups may also choose to establish a hierarchy, where one member serves as the lead and the other two act as peer reviewers. A member who leads one task or role may be a reviewer for other tasks. Notably, the lead is primarily responsible for completing assigned tasks, while the reviewers provide feedback, contribute ideas, and assist in completing the assigned tasks.

If a member becomes unresponsive, the remaining members should prioritize two key roles to complete the assignment. Tasks initially assigned to the absent member will no longer be required, and the non-participation of a team member will not impact the group's overall grade.

You can begin the assignment by navigating to the Teamwork Assignment section under the Table of Contents on the course shell and selecting PeerScholar Teamwork Assignment. For more details, refer to the student guide: https://www.torontomu.ca/courses/toolbox/peerscholar/student-guide

This is the end of the assignment