Teamwork Assignment

Due April 12 at 11:59 PM Starts Mar 6, 2024 12:01 AM

100 % 4 of 4 topics complete

PeerScholar Teamwork Assignment

External Learning Tool

Due April 12 at 11:59 PM Starts Mar 6, 2024 12:01 AM

Introduction

This is a teamwork assignment where teams will work together on processing and analyzing electronic product data represented in XML documents. The assignment aims to apply concepts of XML data handling, information retrieval (IR) techniques, and data mining to extract meaningful insights. Teams will explore using concepts such as XPath, XQuery, TF-IDF, and association rule learning. Each team consists of three students, each team member is assigned a specific role.

Dataset

There are two datasets customized to the distinct tasks of each team role, both of which are provided as attachments:

Electronics Database (electronicsdb.xml): A collection of XML documents detailing electronic products. This semi-structured dataset is intended for the XML Data Engineer and Information Retrieval Specialist tasks, focusing on XML data handling and information retrieval techniques.
Grocery Transactions Dataset: This dataset contains transaction data from a grocery store, designed for the Data Mining Analyst's tasks.

Each dataset is essential for the specific analytical tasks assigned, ensuring all team members have the necessary resources to complete their contributions effectively.

Team Roles

XML Data Engineer

- Focus on parsing, querying, and transforming XML data using XPath and XQuery.
- Extract and structure data for further analysis.

Information Retrieval Specialist

- Apply IR techniques like TF-IDF and document similarity measures.
- Enhance product searchability and relevance based on specifications.

Data Mining Analyst

- Employ data mining techniques to discover patterns and correlations among products.
- Key role in identifying trends and insights for business decisions.

Tasks

XML Data Engineer Tasks

- Import the 'electronicsdb.xml' using the BaseX application, preparing it for analysis.
- Retrieve all products in the 'Smartphones' category or supplied by SupplierID 50.
- Find products sharing the same supplier as 'Alpha Smartphone'.
- Identify 'Cameras' category products supplied by at least one supplier.
- List all products alongside their category and supplier names.
- Compute the total stock quantity for each product.

Information Retrieval Specialist Tasks

- Extract electronic product data from electronicsdb.xml into a CSV file named 'Products.csv'. Ensure accurate conversion of XML elements into CSV columns for

ProductID, ProductName, Specifications, CategoryID, and SupplierID.

- Apply text pre-processing techniques to cleanse the 'Specification' attribute. Techniques include normalization (e.g., lowercasing), removing stop words, stemming/lemmatization, or eliminating punctuation or numbers, aiming to standardize the text data for analysis.

- Create a TermDocumentMatrix to analyze the 'Specification' text data. Define the granularity of terms (unigrams or bigrams) and the weighting measure (TF or TF-IDF), justifying the choice based on the analysis objectives.

- Display the TDM in a matrix format and report its dimensions. Provide insights into the dataset's sparsity.

- Apply a similarity measure, such as cosine similarity, to identify products with specifications most similar to those of the 'Kappa Smart Watch'. Interpret the similarity scores to rank products by relevance.

Data Mining Analyst

- Use a 0.2 minimum support value to apply the apriori algorithm on the attached grocery store transaction dataset, detailing the process and identifying frequent itemsets.

- Generate rules with at least 0.5 confidence for itemsets of three items, explaining the significance of these findings.

- Discuss why specific rules exhibit strong associations based on confidence and support metrics, using examples from the analysis.

Submission

Each team must submit three separate files through the peerScholar platform, each corresponding to the distinct roles within the team. These files must be named to reflect both the role and the name of the student who completed the tasks. The required format for each submission is either DOC, PDF, HTML or ZIP archive named as follows as an example:

- XML_Data_Engineer_Firstname_Lastname.DOC
- Information_Retrieval_Specialist_Firstname_Lastname.HTML
- Data_Mining_Analyst_Firstname_Lastname.DOC

For the XML Data Engineer

Your submission should include:

- A detailed document outlining the implementation of XPath expressions and XQuery (FLWOR) scripts.
- Screenshots of the scripts and their outputs for each task to show your work.
- An explanation of the approach that you followed for each question.

For the Information Retrieval Specialist

Your submission should include:

- The output (in HTML or PDF format) of an RMD source file showcases your work on text pre-processing, TermDocumentMatrix creation, and cosine similarity analysis.
- Documentation within the RMD file that explains your methodology, the choice of text pre-processing techniques, and the rationale behind your approach to similarity analysis.

For the Data Mining Analyst

Your submission must include:

- A report in Word, Excel, PDF, or simple Text files detailing the application of the apriori algorithm, association rule analysis, and any additional data mining tasks assigned.
- Within the report, document each step taken in your analysis, including manual calculations, applying relevant formulae, and the logic behind each decision.
- Explanations that detail the interpretation of your findings, especially the significance of association rules identified and the rationale behind the strength of these rules.
- Ensure no programming code is included in this section; focus on manual calculation and logical reasoning.

Additional Instructions for All Roles

- Ensure your submissions are well-organized, with clear labels for each work section to facilitate easy navigation and review.
- Include an introductory section in your documents outlining your role in the project, the objectives of your tasks, and a brief overview of the methodologies employed.
- Discuss any assumptions made during your analysis and the strategies used to address them.
- For computational tasks, detail your process, even if specific steps are straightforward or seemingly minor.
- Missing components, such as required screenshots for the XML Data Engineer and the RMD output files for the Information Retrieval Specialist, lack of detailed breakdowns, or using automated tools for the Data Mining Analyst will reduce marks.
- Aim for clarity, thoroughness, and accuracy in your submissions to reflect the depth of your understanding and the quality of your work on this project.

Due Dates

In this assignment, you are required to submit your work as a group & peer assess other groups' submissions individually. The assignment has three phases:

- Create phase: Each team submits their work for peer review. **From Mar 06 to Apr 3**

- Assess phase: Each group member gives feedback to at least two groups. **From Apr 4 to Apr 8**

- Reflect phase: Each team sees the feedback from their peers and reflects on it by refining and resubmitting their work. From **Apr 9 to Apr 12**

During the assignment's reflection phase, each student must evaluate their peers' contributions within the group. Notably, students have been assigned to the groups randomly and in phases where collaborative efforts are expected; if a team member opts not to participate or fails to attend scheduled meetings consistently, the remaining two members are suggested to select any two roles to focus on and complete the corresponding tasks. The third role, which the absent or non-collaborating member would have covered, will not be required to be completed by the remaining group members. This adjustment ensures that the project can progress despite collaboration challenges, allowing active team members to concentrate on the required tasks effectively.

Here is a link to more information on navigating the assignment phases using the peerScholar platform.

Best of luck
--Teaching Team

Electronics Database ✓

XML Document

Grocery Transaction Dataset

Image

Assignment 2 (TeamWork - 15% of the final grade)

Assignment

Due April 12 at 11:59 PM

Available on Mar 6, 2024 11:59 PM. **Access restricted before availability starts.**

This is a teamwork assignment where teams will work together on processing and analyzing electronic product data represented in XML documents. The assignment aims to apply concepts of XML data handling, information retrieval (IR) techniques, and data mining to extract meaningful insights. Teams will explore using concepts such as XPath, XQuery, TF-IDF, and association rule learning. Each team consists of three students, each team member is assigned a specific role.

The dataset consists of a collection of XML documents, each representing data for an electronic product. These documents include ProductID, ProductName, Specifications, CategoryID, and SupplierID.

In this assignment, you are required to submit your work as a group & peer assess other groups' submissions individually. The assignment has three phases:

- Create phase: Each team submits their work for peer review. **From Mar 06 to Apr 3**

- Assess phase: Each group member gives feedback to at least two groups. **From Apr 4 to Apr 8**

- Reflect phase: Each team sees the feedback from their peers and reflects on it by refining and resubmitting their work. **From Apr 9 to Apr 12**

You can access the assignment details by clicking on the 'Teamwork Assignment' section under 'Table of Contents' on the course shell and then selecting '**PeerScholar Teamwork Assignment.**'