Toronto Metropolitan University

CIND-110
Data Organization for Data Analysts

Assignment II
XML Data Management, Information Retrieval and Data Mining

Assignment Context: A local botanical garden has a diverse collection of flowers from all over the world. The designated database administrators have maintained a semi-structured XML database to store and organize information about these flowers. However, with the increasing size and complexity of the data, they have encountered several challenges:

- The database has been used by multiple units in the garden administration, each maintaining its local database copy with some additional information. The goal is to align their local databases with the central database and ensure all units have consistent and up-to-date information about the flower collection.
- A data analyst has been tasked to extract and analyze specific information to gain insights about the flower collection and make informed decisions for further expanding the collection.
- To migrate their local servers to a computational cloud platform, the admins need to convert the semi-structured XML data to a structured CSV format to leverage more advanced data analysis techniques.

Starts: Wednesday July 22, 2023, 11:59 PM
Due: Wednesday, August 12, 2023, 11:59 PM
This assignment counts for 15% of the total grade

Section I

XML Hierarchical Data Model: Total Points: 40
As an integral component of central database management, you are tasked with designing and implementing XPath expressions and XQuery (FLWOR) scripts. These standards are necessary to retrieve information from a semi-structured XML dataset, allowing for efficient access and manipulation of the flower collection data.

Instructions

- Download the **Flowers_XML_Data.xml** file given along with the assignment file and create a database in your **BaseX** environment.
- Write the correct XPath expressions for questions 1 through 4 and XQuery (FLWOR) scripts for questions 5 through 8.
- You need to show the script and the screenshot of the output corresponding to your code to obtain full marks.

Questions

1. **[5 Pts.]** Find all distinct `name` elements.
2. **[5 Pts.]** Filter and select flower elements with a `name` value of `Rose`.
3. **[5 Pts.]** Find the first `variety` of each flower element.
4. **[5 Pts.]** Find `flower` elements that have more than four `variety` elements.
5. **[5 Pts.]** For each flower, construct and display a new ⟨*flowerName*⟩ element in the result, containing the name of that flower.
6. **[5 Pts.]** Write an XQuery expression that returns the text "This XML has [n] flowers", where `n` represents the count of flower elements in the XML.
7. **[5 Pts.]** Create and display all flower elements along with a new flowerInfo element in the result. This flowerInfo should follow the format: ⟨*flowerInfo*⟩ Flower [name] originates from [origin]. ⟨/*flowerInfo*⟩. Hint: You may use the `string-join()` function to concatenate multiple origins if needed.
8. **[5 Pts.]** Compare by listing in the displayed result the number of flowers with more than three varieties to those with three or fewer varieties.

Section II

Information Retrieval (IR) Approaches: Total Points: 30

In preparation for the server migration, your task is to convert the semi-structured XML dataset into a more structured CSV format. Following this, you will start the text analytics phase by cleansing the description attribute within the flower data. This process helps eliminate inconsistencies and prepares the data for further analysis. Clean and organized data are essential in creating accurate vectors and applying similarity functions. These functions will facilitate the ranking of flowers based on the similarity of their descriptions to a specified flower, such as the Rose. This ranking can provide valuable insights, such as identifying flowers with similar characteristics or cultivation needs.

Instructions

- For Question 1, use BaseX application. [Hint: For an example of how to do this, refer to Lab - 9 manual. item number 5 on "Serializing and Parsing XML documents"]

- For Questions 2 and 3. use RStudio application. You are expected to create an R Markdown, RMD, file and insert your R code within it. The following R packages will need to be installed and put into use: 'tm', 'RWeka', 'textstem', 'textclean', and 'text2vec'. After completing your work, use the 'Knit' button to create an HTML, DOCX, or PDF file. This file should contain both your text content and the output from your embedded R code chunks. [Hint: For an example of how to do this, refer to the Lab-10 manual.]

- As part of your submission, please include both the RMD source file and the output file in either HTML, DOCX or PDF format. **Failing to submit both files will result in a deduction of marks.**

Questions

1. **[5 Pts.]** Write an XQuery script to convert the XML dataset (**Flowers_XML_Data.xml**) used in Section I to a relational dataset. Save the file as 'Flowers_CSV_Data.csv' document.

2. **[5 Pts.]** Read the relational dataset, and apply three different text pre-processing techniques to cleanse the description attribute.

3. **[10 Pts.]** Create a unigram TermDocumentMatrix (TDM), then represent it in a matrix format and display its dimension.

4. **[10 Pts.]** Using the vectors obtained in the previous question, apply the cosine similarity function and identify which flower is most similar to `Rose`.

Section III

Data Mining - Applying Association Rules: Total Points: 30

As a data analyst in a retail store, you have been provided with a dataset of ten transactions, each involving various fruits and vegetables. These transactions represent purchases made by different customers in a store. Your task is to apply association rule data mining techniques to this dataset. By analyzing the co-occurrence of items across transactions, you can identify trends and relationships that will help in-store inventory decisions, customer recommendations, and targeted marketing efforts. Notably, while the association rules identify patterns, they do not imply causation. Therefore, the insights derived should be considered as correlations between item sets, not direct cause-and-effect relationships.

Instructions

- You are required to calculate your answers manually. Although a calculator may be used for computations, you must document every step toward your final decisions, including all relevant formulas applied during the process.

- Please use Word, Excel, PDF or simple Text files to prepare and submit your report. Note that submissions in the form of photographs or handwritten notes will not be accepted.

- Also, refrain from using programming languages like R and Python or tools like WEKA for this assignment section. Calculations conducted through these mediums will not be considered for grading.

- To attain full credit for this section, provide a detailed breakdown of every step involved in your calculations. We are interested not only in the correct answer but also in the methodology you employ to obtain the solution.

Questions

111		sugar	coffee	tea	milk	cake	biscuit	chocolate	brownies
112		coffee	bread	milk	chocolate	apple	banana	cookies	cashew
113		tea	sugar	orange	banana	cake	brownies	nuts	biscuit
114		coffee	sugar	orange	banana	cashew	milk	nuts	chocolate
115		coffee	sugar	bread	biscuit	nuts	milk	apple	chocolate
116		tea	sugar	coffee	cashew	nuts	milk	apple	dates
117		dates	sugar	coffee	tea	cashew	juice	orange	chocolate
118		coffee	bread	sugar	tea	cake	cookies	apple	juice
119		tea	coffee	milk	biscuit	brownies	chocolate	juice	sugar
120		sugar	tea	milk	cookies	cake	chocolate	biscuit	orange

1. **[15 Pts.]** Using a minimum support of 65%, apply the Apriori algorithm to this dataset.
2. **[15 Pts.]** List all possible association rules that meet the minimum confidence level of 95% or higher.

End of Assignment