Toronto Metropolitan University

CIND 110, FALL 2022
Data Organization for Data Analysts

Assignment 2

Starts: Wednesday, November 2, 2022

Due: Wednesday, December 7, 2022

This assignment counts for 15% of the final grade. It covers 3 of the modules after the midterm, and there is one question from each module. Section I questions are from XML (module 9), Section II is from Information Retrieval (module 10), and Section III questions are from Data Mining (module 11). Please follow the instructions listed before the questions and submit your answers accordingly.

Section I

XML Hierarchical Data Model: Total Points: 40

Instructions

- Download the **Section-I_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.

---

1. [**5 Pts.**] List the title of all books published in December.
2. [**5 Pts.**] List the title and prices of Romance or Fantasy books.
3. [**5 Pts.**] List the title, price and the description of the last book in the dataset.
4. [**5 Pts.**] List the authors of the books which cost more than 5 dollars and were published in 2001.

---

5. [**5 Pts.**] Find the book title with the highest price.
6. [**5 Pts.**] Find the number of Fantasy books.
7. [**5 Pts.**] Find the number of authors who published more than two books
8. [**5 Pts.**] Compare by listing the number of books with a price greater than 5 dollars to the number of books with a price less than 5 dollars.

Section II

Information Retrieval (IR) Approaches: Total Points: 35

Instructions

- Use BaseX for Question 1.

  [Hint: Refer to Lab - 9 manual item no. 5, Serializing and Parsing XML documents]

- Use RStudio for Questions 2 and 3. Create an RMD file and insert your R code.

  The following R packages need to be installed and used: tm, RWeka, textstem, textclean and text2vec. Use the Knit button to generate an HTML, DOCX or PDF file that includes both the content and the output of the embedded R code chunks.

  [Hint: Refer to Lab-10 manual example]

- Submit the source file in RMD format and the output file either in HTML, DOCX or PDF format. Failing to submit both will be subject to a mark deduction.

Questions

1. **[10 Pts.]** Write an XQuery script to convert the XML dataset (**Section-I_XML-Data.xml**) used in Section I to a relational dataset. Save the file as 'booksCSV.csv' document.

2. **[10 Pts.]** Read the relational dataset, apply three different text pre-processing techniques to cleanse the description attribute, and then create a Term Document Matrix.

3. **[15 Pts.]** Create a unigram TermDocumentMatrix (TDM), then represent it in a matrix format and display its dimension.

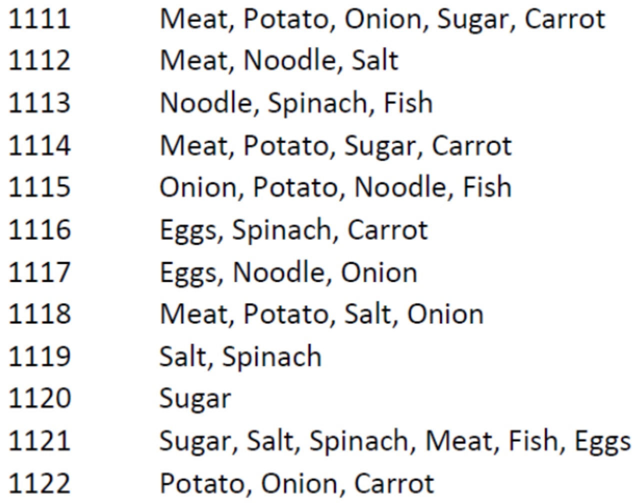
Section III

Data Mining: Total Points: 25

Instructions

- This assignment is hand calculation work. Except a calculator, nothing else is expected to be required. You must have to do calculations manually and report all the steps that you have followed to reach the decisions (including formulas). However, you are suggested to use digital editing Software, such as WORD, EXCEL, PDF or Simple TEXT files, to submit your results/report.
- Please do not submit (1) a picture of your hand-written notes and (2) any coding in R, Python or WEKA. Calculations done using such program tools will not be graded.
- You are expected to show the details of all steps in your calculation to score full marks.

One of the major techniques in data mining involves the discovery of association rules. These rules correlate the presence of a set of items with another range of values for another set of variables. The database in this context is regarded as a collection of transactions, each involving a set of items, as shown below.



1111	Meat, Potato, Onion, Sugar, Carrot
1112	Meat, Noodle, Salt
1113	Noodle, Spinach, Fish
1114	Meat, Potato, Sugar, Carrot
1115	Onion, Potato, Noodle, Fish
1116	Eggs, Spinach, Carrot
1117	Eggs, Noodle, Onion
1118	Meat, Potato, Salt, Onion
1119	Salt, Spinach
1120	Sugar
1121	Sugar, Salt, Spinach, Meat, Fish, Eggs
1122	Potato, Onion, Carrot

Questions

1. **[15 Pts.]** Apply the Apriori algorithm on this dataset with minimum support = 0.3
2. **[10 Pts.]** Describe the Association Rules obtained from the calculation which have a confidence of 75% or higher for an itemset.

End of Assignment