Ryerson University

CIND110
Data Organization for Data Analysts

Assignment 3

Starts: Wednesday, November 6, 2021, 8:00 PM

Due: Thursday, December 4, 2021, 11:59 PM

This assignment counts for 10% of the final grade

1 Section-XML

Write the XPath and Xquery scripts: [Total Points: 40]

INSTRUCTIONS

- Download the **Bookstore.xml** file given along with the assignment file and create a database in your **BaseX** environment.
- Write the correct XPath expressions for questions no. 1 through 5 and XQuery scripts for questions 6 through 11.
- You need to show the script and the screenshot of the output corresponding to your code to obtain full marks.

---

1. [**2 Pts.**] Write an XPath expression to find all authors along with their corresponding books.
2. [**2 Pts.**] Write an XPath expression to find the prices of all books and their genre.
3. [**2 Pts.**] Write an XPath expression to find the title, price and the description in text of the last book book in the catalog.
4. [**2 Pts.**] Write an XPath expression to find the authors and titles of the books which cost more than 40 dollars, along with the respective prices.
5. [**2 Pts.**] Write an XPath expression to find the authors and prices of the books belonging to Computer genre.

---

6. [**5 Pts.**] Write an XQuery (FLWOR) script to find all the authors without their path descriptions.
7. [**5 Pts.**] Write an XQuery (FLWOR) script to find the titles of the books arranged in ascending order of price, of which the price are more than 30 dollars.
8. [**5 Pts.**] Write an XQuery (FLWOR) script to provide only the descriptions of the books, which cost less than 5 dollars.
9. [**5 Pts.**] Write an XQuery (FLWOR) script which gives the various genre along with the text of the title of the books in them.
10. [**5 Pts.**] Write an XQuery (FLWOR) script which gives the description text showing that the book belongs to Fantasy genre.
11. [**5 Pts.**] Write an XQuery (FLWOR) script which gives the list of authors whose books cost less that 30 dollars and provides the titles of the books otherwise.

2 Section-IR

Answer the following questions with detailed calculations: [Total Points: 45]

INSTRUCTIONS

- The attached script reads 100 transcripts exported from **https://www.ted.com/talks**, and creates a Term Document Matrix to store the frequencies of words/terms and the references to the documents that contain them.
- Use RStudio for this assignment. Edit the file 'Sec-2-IR_Q-rmd.Rmd' and insert your R code, then click the Knit button to generate an (HTML, Word or PDF) document that includes both the content and the output of the embedded R code chunks.
- When you are done with your answers and before submitting, save the file with the following naming convention: your Lastname.Firstname. Submit the source (in RMD format) and the output (either in PDF, WORD, or HTML format) files. Failing to submit both will be subject to mark deduction.

Questions

1. **[15 Pts.]** Apply three different text pre-processing techniques to cleanse the data before creating the Term Document Matrix. For example, you might trim the suffix and prefix of the original words by applying a Stemming algorithm. In addition, you might remove the most commonly used words in the language, which seldom contribute to the meaning of the sentence, by applying a Stopword Removal algorithm.
2. **[10 Pts.]** Create a unigram TermDocumentMatrix (TDM) and inspect for 10 rows.
3. **[10 Pts.]** Represent TDM in a matrix format and display its dimension.
4. **[10 Pts.]** Generate a wordcloud of the most occured 100 words across all transcripts.

3 Section-Data Mining

Find out the best Association Rules : [Total Points: 15]

INSTRUCTIONS

- This assignment is hand calculation work. Except for a calculator, nothing else is expected to be required. You must do calculations manually and report all the steps 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.
- Avoid submitting a picture of your hand-written notes and do not submit any coding in R, Python or Weka and upload those in your submission for this Section.
- 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

Apply the APRIORI algorithm to answer the following questions

1. **[10 Pts.]** Show the Total itemset, sets of items for C1 and L1 and sets of items for C2 and L2 and/or higher sets till the program terminates.
   Minimum Support = 0.3
2. **[5 Pts.]** Describe the Association Rules obtained from the calculation which have **confidence of 75%** or higher for an itemset.