Assignment 6
XML and NoSQL

Applied Data Science

Instructions

In this assignment, you will use the Cricket dataset subsets cricket-xml-1.xml, cricket-xml-2.xml, cricket-xml-3.xml and cricket-xml-4.xml . You will use Python and BaseX.

Submit your assignment as a .pdf file with your name and email in the header, single-spaced and font no-larger than 12pt. **Put all your results in the PDF file and include your code and scripts in an appendix or upload as additional files**.

This assignment will be graded on the correctness of your responses, your explanation of the process used to get the results, and overall quality.

Questions

Question 1

Use Python's XML and CSV packages to read the cricket-xml-1 dataset and rewrite this file as a CSV. Your output CSV file should indicate the match id, the team names, and the amount of runs scored in that match by each team. Describe briefly how you accomplished this (naming the most relevant functions). Upload the output file along with your submission and include your code in either an appendix or an additional file.

Question 2

Use Python's XML and JSON packages to read the cricket-xml-2 dataset and rewrite this file as a JSON. Your output JSON file should indicate the match id, the team names, and the amount of runs scored in that match by each team. Describe briefly how you accomplished this (naming the most relevant functions). Upload the output file along with your submission and include your code in either an appendix or an additional file.

Question 3

Use BaseX to answer the following questions about cricket-xml-1 and cricket-xml-2. Include the text of your queries in an appendix or as an additional file.

1. For how many balls did England bat in cricket-xml-1?
2. How many wickets did India get in cricket-xml-2?
3. How many runs did Australia score against England in cricket-xml-1?
4. How many runs did Pakistan score in cricket-xml-1 and cricket-xml-2 combined? (be sure to remove duplicate rows.)