Department of Computer Science and Engineering
The Hong Kong University of Science and Technology

CSIT6000P Spatial and Multimedia Databases

2023 Spring

Version 1.0

Assignment 1 [*Total: 30 marks*]

Due date: 11:59pm Saturday 1 April 2023
HKUST Canvas online submission only.

You are given a simplified real-world database *D* of objects represented as polygons, such as buildings, shopping malls, and gardens. Each record contains an ID, name, type, and geometry (a polygon is represented by a sequence of points in the format of their longitude and latitude values). In this assignment, you are asked to conduct an experimental study on the performance of spatial indexing methods on the given dataset and report your findings. This assignment only considers a simplified scenario where all the data are already loaded into memory. That is, no disk-based operations need to be considered.

Task 1 [*5 marks*] (**MBR calculation**)

(1) [*1 mark*] Write a program to compute the spatial extent of dataset *D*. The spatial extent of *D* is the MBR of all polygons in *D*. Report the MBR of *D*.

(2) [*2 marks*] Create dataset *D'* which adds an MBR column for the polygon in each record in *D*. Output the spreadsheet file for the new dataset *D'*.

(3) [*2 marks*] Let *n* be the resolution for recursive decomposition of the space as defined by the spatial extent of *D*. What are the sizes (in cm by cm) of the smallest Peano cells for *n* = 16, 23 and 28 respectively? Show your calculation steps. Please also discuss which resolution value is suitable for *D*.

Task 2 [*10 marks*] (**z-value indexing**)

(1) [*6 mark*] Write a program to generate the base-5 z-value for each polygon, for *n* = 16, 23 and 28 respectively. We use only one z-value for each object based on its MBR. Add three columns of z-values to *D'* for these three different resolution levels. Output the spreadsheet file with the new columns.

(2) [*4 marks*] For each object, compare the sizes of the Peano cells for the same object using the above 3 resolution numbers and analyze your findings. Remember that the Peano cells for an object should be as tight as possible. Your discussions should reveal insights about choosing the proper resolution level for an application. If you see any issues with using just one z-value for one object, discuss possible solutions.

Task 3 [*10 marks*] (**window query processing**) A window query with a given query rectangle represented as $Q = \{(x_{low}, y_{low}), (x_{high}, y_{high})\}$ returns the number of objects inside Q.

(1) [*7 marks*] Write a program to perform window queries using two approaches: (i) by exhaustively checking every object in the dataset; and (ii) by using z-values you generate in Task 2 for the above three *n* values.

(2) [*3 mark*] Use 20 randomly generated window queries of different sizes at different locations to search using the programs you developed above, to report (i) the number of

objects inside each query window, (ii) the number of objects searched for each query for using no z-values (i.e., exhaustive search) and using z-values obtained using different resolution numbers.

Task 4 [*5 marks*] (**analysis and reporting**) You are required to write a report with no more than 6 pages (using this document as the template). Your goal in writing this report is to help the reader understand your design, your code, your experiments, and your findings. The algorithms must be clearly documented in plain language (if you prefer to use pseudocode, please ensure it is readable with proper comments). Note that the marks for this task will be allocated based on your report structure, clarity, and readability, while the assessment of the content in the report concerning each task above will be combined with the assessment of the corresponding tasks.

(1) [*3 marks*] To document your design, with necessary explanations or any notes to make your assignment understandable.
(2) [*2 mark*] To include the outputs and discussions for Tasks 1-3.

Notes:

1. In this assignment, you can use any programming language of your choice. No programming support will be provided in this course. No DBMS is needed. You will load the entire dataset into memory and perform all operations required in memory.
2. The query results (i.e., the number of points inside a query window) should be identical for the same query using different indexing structures, but the number of points compared can be different. This fact can be used to verify the correctness of your code.
3. You may be required to demonstrate and explain your programs in front of the TA. If there is such a need, you will be contacted by the TA to arrange a time and a way that is convenient for both you and the TA.
4. You are required to do this assignment independently, including developing all the code and doing the experiments. You should not copy the code from the Internet, any other sources, or from your classmates.

Submission guideline:

1. Late submission: unless approved by the lecturer or the TA in writing, every delay from one minute to 12-hours will incur a 25% deduction of your total marks for this assignment. That is, a delay of 2 days will lead to 0 marks for this assignment.

2. Submitted materials: should be compressed as a .zip file with student id as the file name

- Project report (up to 6 pages) in PDF format.
- Source code and a Readme file. Please document how we can run your code as well as how to install necessary packages, if any, in the Readme file. There is no need to include the dataset in your submission.
- Make sure your report and code contain your name and student ID.

3. Submission channel: on Canvas.

Warning: *This is an individual assignment. Collusion can be easily detected by software tools. Plagiarism will not be tolerated at HKUST. Please refer to Student Conduct and Academic Integrity regulations. If you are unclear about what level of discussions and help you can get for this assignment, please talk to the lecturer or the TA.*