COMP6210: Assignment 2

Semester 2, 2024
School of Computing
Macquarie University

Marks: 40 marks (40% of total grade)

Due Date: 11:55 PM, 02 November 2024 (Saturday, End of Week 13)

What to Submit: Source code, project report, slides, and presentation video (or the URL linking to the video)

Where to Submit: Electronic submission via iLearn

This is a group assignment focused on R-tree based algorithms. Submissions will be graded based on code correctness, the completeness of the report, quality of slides and presentation.

Group Formation:

- Each group should consist of **3 members**.
- Within each group, one member should be appointed as the coordinator, responsible for communicating with tutors and submitting the final solution.
- Each of the Individual Tasks should be undertaken by **ONE** group member.

Tasks:

- Individual Task 1: Nearest Neighbor Search
- Individual Task 2: Skyline Search
- Individual Task 3: Project Report
- Group Task: Presentation

Grading: Group Marks (15 marks) + Individual Marks (25 marks)

IMPORTANT: The use of GPT, copying from lecture notes, or any form of plagiarism is strictly prohibited. Violations will be reported to the faculty disciplinary team.

Below are the detailed requirements:

Task 1: Nearest Neighbor Search (refer to weeks 8 & 9 & 10 lecture notes)

- **Datasets:** There are three datasets: Restaurant, Shop, and Parking Datasets. Each dataset consists of 2D points, stored in a text file with the following format:

```
id_1 x_1 y_1
id_2 x_2 y_2
...
id_n x_n y_n
```

Each line includes a unique ID for a point and its geographical coordinates, longitude and latitude. For example, an entry in the Shop dataset, such as "`id_1=1, x_1=33.85, y_1=151.21`" precisely indicates the location of a shop, with "x" representing longitude and "y" representing latitude.

- **Queries:** We have 200 users interested in finding the nearest facilities. Their locations are provided in a text file formatted identically to the datasets:

  ```
  id_1 x_1 y_1
  id_2 x_2 y_2
  ...
  id_200 x_200 y_200
  ```

  For example, `id_1=1, x_1=31.45, y_1=150.44` indicates a user's location.

- **Program Design:**
  - Select **ONE** dataset (Restaurant, Shop, or Parking).
  - Find the nearest facility (restaurant, shop, or parking lot) for each query using the following algorithms:
    1. **Sequential Scan Based Method:** Calculate the distance between a query point to every point in the selected dataset to find the nearest neighbor.
    2. **Best First (BF) Algorithm:** Construct an R-tree for the selected dataset. Then, apply the BF algorithm using the R-tree to find the nearest neighbor for each query point.
    3. **BF with Divide-and-Conquer:** Firstly, divide the dataset into two subspaces (based on X dimension or Y dimension), then construct an R-tree for each subspace. Use the BF algorithm to find the nearest point to the query in each subspace. Finally, compare the distance between the nearest points delivered from each subspace to determine the final nearest neighbor in the entire dataset.

- **Output:** For **each** algorithm (Sequential Scan Based, BF Algorithm, and BF with Divide-and-Conquer), display and output the following information in **a single txt file**:
  - The ID, x, and y coordinates of the nearest neighbor identified for each query point (e.g., "`id=56, x=34.15, y=149.21 for query 1`").
  - The **total running time** for processing all 200 queries and the **average time per query** (i.e., divide the total running time by 200).

Task 2: Skyline Search (refer to weeks 8 & 9 & 10 lecture notes)

- **Dataset:** There are three datasets—city1, city2, and city3—each representing homes in the city. These datasets contain 2D points and are stored in text files formatted as follows:

  ```
  id_1 x_1 y_1
  id_2 x_2 y_2
  ...
  id_n x_n y_n
  ```

  Each line in the datasets represents a home, with "x" indicating the cost of the home (for example, $500,000) and "y" representing its size (for example, 200 square meters). The goal is to apply the Skyline Search algorithm to find homes that provide the optimal balance of size and cost. This method helps to filter out properties that are either too costly or too small in comparison to other available options.

- **Program Design:**
  - Select **ONE** dataset (city1, city2, or city3).
  - Implement the following algorithms to assist users in choosing the ideal home based on two key criteria: cost and size.
    1. **Sequential Scan Based Method:** Identify the skyline by sequentially evaluating whether each node is dominated by any other nodes.
    2. **Branch and Bound Skyline (BBS) Algorithm:** Construct an R-tree for the selected dataset. Implement the BBS algorithm with the R-tree to identify the skyline.
    3. **BBS with Divide-and-Conquer:** Firstly, divide the dataset into two subspaces (based on X dimension or Y dimension), then construct an R-tree for each subspace. Implement the BBS algorithm to identify the skyline in each subspace. Finally, obtain the skyline for the entire space through 1D dominance screening method.

- **Output:** For **each** algorithm (Sequential Scan, BBs Algorithm, and BBs with Divide-and-Conquer), display and output the following information into **a single txt file**:
  - The skyline results for the selected dataset, i.e., sequentially output each point's ID, x-value, and y-value.
  - The **execution time** taken to find the skyline for the selected dataset.

Task 3: Project Report (refer to the uploaded template)

The project report should include the following elements:

- A breakdown of the tasks and responsibilities allocated to each member of the group.
- An **in-depth description** of the primary functions within your source code (Tasks 1 and 2).
- A clear specification of the requirements for running your code, including the required operating system environment, the location of input files, and any essential input parameters, etc.
- Based on the two given sample datasets and queries, explain the BF algorithm based NN search and BBS based skyline search **in detail** respectively. You need to **draw the necessary diagrams**, e.g., MBRs and R-Tree, and illustrate the process **step by step** with your detailed analysis.

Group Task: Presentation

Each group is required to prepare a presentation that offers a comprehensive overview of the project.

- You can use the template we provided on iLearn to prepare for the slides.
- **Each** of the group members needs to give the presentation for the corresponding Individual Task.

- The presentation should include an introduction to the project, an in-depth analysis of your design and understanding of **Task 1** and **Task 2**, and a discussion of **the two case studies** in your report.
- Each group is required to prepare a video recording of the presentation that includes a visible appearance of each speaker. The ideal length for this presentation is between **25 to 30 minutes**.

Programming Environment:

We highly recommend using **Python** with only the **standard libraries** provided by the programming languages, rather than relying on existing R-Tree libraries. If you choose to use a different programming language, you must provide detailed instructions on how to configure the programming environment and how to execute the program.

Submission:

The **coordinator** of each group should submit **a single zip file** named **"FirstName_LastName_Assignment2.zip"** via iLearn. The submission must include the following items:

- **Source Code and Output txt files for Task 1 and Task 2:** Make sure the code can be run in the standard general programming environment.
- **Project Report**.
- **Presentation Slides and Video:** Please submit both the slides and the video. We recommend uploading the video to YouTube or Google Drive and providing the URL at the end of your project report.

Late Submission:

Refer to the policy in the Unit Guide.

Zero Tolerance for Cheating:

If you incorporate libraries and/or content from external sources into your submission, you must properly cite these sources and provide the corresponding references. All submissions will undergo plagiarism checks. Any confirmed cases of plagiarism will be reported to the faculty for disciplinary action.