INFS 4205/7205: Assignment 1

Prepared by Yufei Tao and Junhao Gan

March 19, 2017

This is a coding assignment, which can be attempted in groups of size at most 5. The objective is to implement the R-tree. Each submission will be graded based on correctness and efficiency. The rest of the document explains the details.

How Your Submission Will Be Tested. You will be given a dataset which contains 2D points. The dataset will be provided in a text file of the following format:

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

Specifically, the first line gives the number of points in the dataset. Then, every subsequent line gives a point's id, x-, and y-coordinates.

Your program should build an R-tree in memory from the dataset. Then, we will measure its query efficiency as follows.

- First, your program should display the time of reading the entire dataset once. This time serves as the "sequential-scan benchmark" to be compared with the cost of your query algorithms that leverage the R-tree.
- [Range Query Testing] You will be given a set of 100 range queries in a text file whose format is:

  ```
  x_1 x'_1 y_1 y'_1
  x_2 x'_2 y_2 y'_2
  ...
  x_100 x'_100 y_100 y'_100
  ```

  That is, each line specifies a query whose rectangle is $[x, x'] \times [y, y']$. You should output:

  - to a disk file the *number* of points returned by each query—note: we need only the number of points retrieved, instead of the details of those points.
  - the total running time of answering all the 100 queries, and the average time of each query (i.e., divide the total running time by 100).

- [Nearest Neighbor (NN) Query Testing] You will be given a set of 100 NN queries in a text file whose format is:

  ```
  x_1 y_1
  x_2 y_2
  ...
  x_100 y_100
  ```

  That is, each line specifies a query point with coordinates x and y. You should output:

  - to a disk file the *id*(s) of the point(s) returned by each query—note: if multiple data points happen to have the smallest distance to the query, then all of them should be output.
  - the total running time of answering all the 100 queries, and the average time of each query (i.e., divide the total running time by 100).

Grading Scheme.

- Range Queries: 50%, including
  - Correctness: 25%. If your program correctly answers m (out of 100) queries, you get $25 \cdot (m/100)$ marks for this part.
  - Efficiency: 25%. If the average query time is at least 10 times faster than sequential scan, you get 25 marks for this part. If at least 5 times faster (but less than 10 times), you get 10 marks. If less than 5 times faster, no marks.
- NN Queries: 50%, including
  - Correctness: 25%. Criteria same as for range queries.
  - Efficiency: 25%. Criteria same as for range queries.
- Contribution Requirements: Every team member is required to *declare* the percentage of work that s/he has done. The tutor will carry out a one-to-one interview with each team member to assess whether the percentage is reasonable. If the program receives a score of s (from the previous two bullets) and the team has a size of t, then a team member with p percent contributions receives a final score of

$$s \cdot \min\left\{\frac{p}{1/t}, 1\right\}.$$

- As mentioned, every group can have a size at most 5. No special credits will be given if the group has a smaller size (this assignment encourages team work).

Programming Language. C++ (including variants like C, C#, ...), Java, Python, or any other language approved by the instructor.

Submitting Your Source Code. The code must be uploaded to Blackboard by 11:59pm 29 May. The students must be available for meetings with the tutor from May 30 to June 2.

Zero Tolerance for Cheating. You are required to implement the R-tree from *scratch*. This means that you can use only the standard libraries provided in the programming language of your choice (e.g., for C++, STL is considered as a standard library). Contact the tutor if you have doubts about whether a library is “standard”.

All submissions will be checked for plagiarism. Any confirmed case will receive a 0 mark for the assignment outright, and be reported to the school for disciplinary actions.