COMP 9102 Data Management and Information Retrieval
Assignment 3
Top-k and Skyline Queries

Tianle WANG
Univ No. 3030096596
The University of Hong Kong

Date: November 21, 2022

Preparation

0.1 Environment Requirement

```
MacOS Monterey == 13.0 (OS that can run python3 command)
python == 3.8
pandas == 1.5.1
```

0.2 Compile

```
python3 baseline.py --category_ids [2,5] --k 10
python3 topk_query.py --category_ids [2,5] --k 10
python3 skyline_query.py --category_ids [2,5]
```

1 Top-k queries

1.1 baseline

```
                  Player   Tm  AST   PTS     SCORE
559    Russell Westbrook  OKC  840  2558  1.927152
211         James Harden  HOU  906  2356  1.921032
551            John Wall  WAS  831  1805  1.622848
270         LeBron James  CLE  646  1954  1.476902
121        Stephen Curry  GSW  523  1999  1.358733
517       Isaiah Thomas  BOS  449  2199  1.355241
```

```
326         Damian Lillard  POR  439  2024  1.275791
550           Kemba Walker  CHO  435  1830  1.195535
513            Jeff Teague  IND  639  1254  1.195525
20   Giannis Antetokounmpo  MIL  434  1832  1.195213
```

1.2 NRA algorithm

```
[559, 211, 551, 270, 121, 517, 326, 550, 513, 20]
[1.9271523178807946, 1.9210320562939796, 1.6228479410135195, 1.476902312271418,
    1.358732591514825, 1.3552409701978125, 1.2757906200864015, 1.1955351086579435,
    1.1955247528853772, 1.1952132167273342]
55
```

2 Skyline queries

2.1 BNL algorithm

```
[211, 559]
```

3 Conclusion

1. The running times of several algorithms are fast.
2. NRA iteratively retrieves objects and their atomic scores from the ranked inputs in a round-robin fashion.
3. NRA accesses objects sequentially from all inputs and updates the upper bounds for all objects seen so far unconditionally.
4. Cost: $O(n)$ per access (the expected distinct number of objects accessed so far is $O(n)$).
5. No input list is pruned until the algorithm terminates.
6. BNL can do multiple passes over the data (for high dimensional data the domination probability is low).