# COMP 9102 Data Management and Information Retrieval Assignment 2 Indexing and Similarity Search for Dense Multidimensional Vectors

Tianle WANG
Univ No. 3030096596
The University of Hong Kong

*Date: October 28, 2022*

## Preparation

### 0.1 Environment Requirement

```
MacOS Monterey == 12.6 (OS that can run python3 command)
python == 3.8
matplotlib == 3.6.1
numpy == 1.23.3
```

### 0.2 Compile

```
python range_query.py --numpivots 10 --eps 0.2
python knn_query.py --numpivots 10 --k 5
python evaluation.py
```

## 1 Range Query

### 1.1 Naive

```
[5739, 4186, 1095]
average distance comp per query (Naive) = 10000
total time Naive = 5.03141975402832
```

### 1.2 Pivot-based

```
[5739, 4186, 1095]
average distance comp per query (Pivot-based) = 16.71
total time Pivot-based = 2.8375251293182373
```

### 1.3 iDistance-based

```
[5739, 4186, 1095]
average distance comp per query (iDistance) = 2079.085
total time iDistance = 2.779356002807617
```

## 2 KNN Query

### 2.1 Naive

```
average distance comp per query (Naive) = 10000
total time Naive = 5.208529949188232
```

### 2.2 Pivot-based

```
average distance comp per query (Pivot-based) = 3156.76
total time Pivot-based = 6.8921568393707275
```

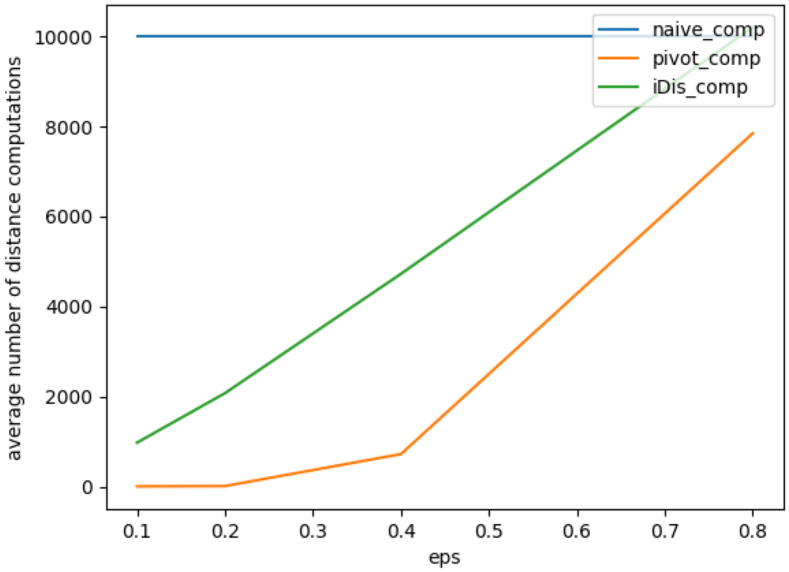
### 2.3 iDistance-based

```
average distance comp per query (iDistance) = 9906.48
total time iDistance = 7.634191036224365
```

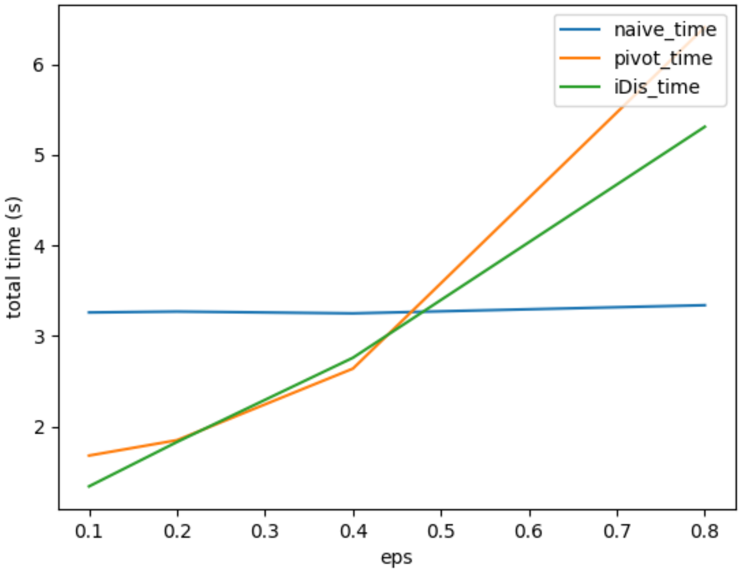
## 3 Evaluation

### 3.1 Query Range

Methods comparation based on number of computation:

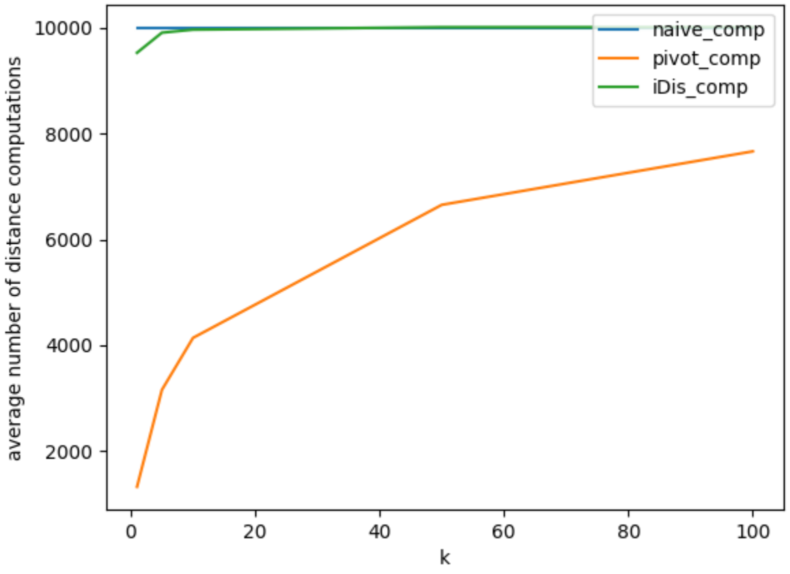


Methods comparation based on total time (s):

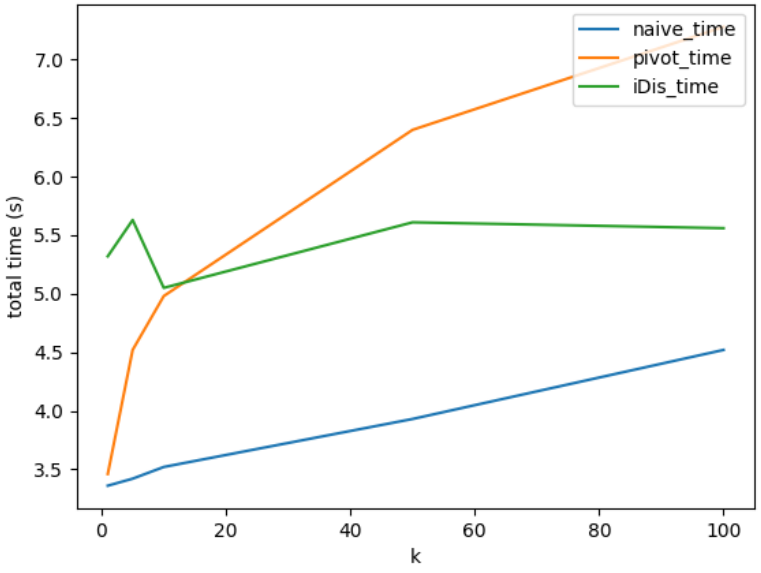


## 3.2 KNN Range

Methods comparation based on number of computation:



Methods comparation based on total time (s):



## 4 Conclusion

1. As the distance $\epsilon$ increases, both the number of computations and the total time for the Pivot-based and iDistance-based methods increase.
2. The number of computations and the total time of the Naive method are independent of increasing

distance $\epsilon$.

3. As the number of neighbors $k$ increases, both the number of computations and the total time of the Pivot-based and iDistance-based methods increase.
4. The number of computations of the Naive method is independent of increasing number of neighbors $k$, and the total time of that is positively correlated with increasing number of neighbors $k$.