# *BANG* Library

## Contents



## 1. Dataset Preparation

Download the dataset files in .bin format from big-ann-benchmarks (https://github.com/harsha-simhadri/big-ann-benchmarks/blob/main/neurips21/t3/README.md).

Example: For the SIFT10M dataset, download the dataset using:

```
python create_dataset.py --dataset bigann-10M
```

The base dataset and query files are ready.

Note: We generate the groundtruth manually and do not use the one provided with the base dataset.

## 2. Graph (Index) Construction

Using the base dataset downloaded previously, we generate the Graph Index and Compressed Vectors using DiskANN/Vamana (https://github.com/microsoft/DiskANN/blob/main/workflows/SSD_index.md ). The compression factor can be controlled by the '-B' parameter. The higher the value, lower is the compression. Set this parameter appropriately based on the GPU memory that can be allocated to store compressed vectors. In the below e.g. supplying B as 1 (i.e. 1 GiB), would result in compressing the 128-byte base vector into 107 (1GiB/10M) bytes per vector. The no. of chunks (m) is 107.

Build the Vamana Graph Index (from the apps directory) using:

```
./build_disk_index  --data_type  uint8 --dist_fn l2 --data_path <
your dataset path e.g. /mnt/ssd_volume/big-ann-
benchmarks/data/bigann/base.1B.u8bin.crop_nb_10000000> --
index_path_prefix sift10m_index -R 64 -L 200 -B 1 -M 48
```

After a successful index build, the following files are generated:

1. sift10m_index_pq_compressed.bin : Compressed vectors corresponding to the base dataset
2. sift10m_index_pq_pivots.bin: This file has three sections of data
    a. Pivots of the PQ clusters created on the subspaces using k-means clustering.

   b. The centroid of the entire dataset. Used for zero-centring the dataset in BANG Search
   c. PQ Chunk offsets information that indicates the start and end offset of each vector subspace (i.e. chunk)
3. sift10m_index_disk.index: The generated graph index (Full Precision vectors along with the adjacency list)

The sift10m_index_disk.index is not compact, in the sense: there are holes in the disk binary layout. Hence, we remove these holes and make the binary representation contiguous so that it can be efficiently loaded into RAM. The BANG requires some metadata (e.g. medoid, degree bound) for the graph index in a separate file. To accomplish the above, run a Python script provided in the *BANG* repo (https://github.com/karthik86248/BANG-Billion-Scale-ANN/blob/main/BANG_Base/bang_preprocess.py ).

```
python bang_preprocess.py <your path to disk.index file e.g.
/mnt/ssd_volume/diskANN-
latest/DiskANN/build/apps/sift10m_index_disk.index> <provide path to
the o/p .bin file e.g. /mnt/ssd_volume/diskANN-
latest/DiskANN/build/apps/sift10m_index_disk.bin>  128 1 64
```

After the above step, two files are generated: sift10m_index_disk.bin and sift10m_index_disk_metadata.bin:

As mentioned earlier, we compute the groundtruth using DiskANN. We compute the groundtruth using:

```
./compute_groundtruth --data_type uint8 --dist_fn l2 --base_file
<your dataset path e.g. /mnt/ssd_volume/big-ann-
benchmarks/data/bigann/base.1B.u8bin.crop_nb_10000000> --query_file
<your query file path /mnt/ssd_volume/big-ann-
benchmarks/data/bigann/query.public.10K.u8bin> --K 10 --gt_file
<specify the path to generated groundtruth file e.g.
/mnt/ssd_volume/diskANN-
latest/DiskANN/build/apps/sift10m_groundtruth.bin>
```

Now, we are ready to start the *BANG* Search.

# 3. BANG Search.

Download the code from **BANG** Repo : https://github.com/karthik86248/BANG-Billion-Scale-ANN

Navigate to *BANG_Base* directory. Build the code using:

```
mkdir build && cd build && cmake .. && make
```

For example, on the SIFT10M dataset with 10K queries, run the search for 10-recall@10 using:

```
./bang_search <your path to DiskANN files e.g.
/mnt/ssd_volume/diskANN-latest/DiskANN/build/apps/sift10m_index> <
your query file path e.g. /mnt/ssd_volume/big-ann-
benchmarks/data/bigann/query.public.10K.u8bin> <your path to
groundtruth file /mnt/ssd_volume/diskANN-
latest/DiskANN/build/apps/sift10m_groundtruth.bin> 10000 10 uint8 l2
```

Provide various values for worklist length when prompted via the console. The values could be in the range 10 to 152 (assuming recall parameter used is 10).

# 4. Stock Graph Index files

For the SIFT10K dataset (http://corpus-texmex.irisa.fr/ ), pre-built DiskANN Graph Index files and required PQ Compressed files are packaged at the following GitHub location : https://github.com/karthik86248/BANG-Billion-Scale-ANN/blob/main/sift10kfiles.tar.gz

Extract the contents of the tarball. Run the below command from the sift10kfiles folder and provide the location of the respective files as below to *BANG* search:

```
./bang_search ./sift10k_index ./siftsmall_query.bin
./sift10k_groundtruth.bin  100 10 float l2
```