Assignment - 3
Approximate Nearest Neighbors

Harika Nalam
SJSU ID - 015939963

To Implement ANN algorithms ona dataset.

ANN algorithms:

1. Exhaustive search
2. Locality Sensitive Hashing
3. Product Quantization
4. Tree and Graph
5. Hierarchical Navigable Small World Graph

The dataset SIFT 1M is used to implement ANN algorithms.

About dataset:

The dataset consists of titles of png files to indicate the column positions of the SIFT features. Each SIFT feature is a 128D column, and the corresponding patch is saved as an image of png format. This dataset is used to evaluate the approximate nearest neighbors algorithms.

It comprises three subsets of vectors.

1. Base vector
2. Query vector
3. Learning vector

The base vector is used to train the model and the query vector is used to search the indices of the nearest neighbors.

Time taken by the ANN algorithm to search the query vector

Algorithms	Time in sec	Comments about Algo
Exhaustive Search	0.26	Algo has high search quality but it is not the best when compared to the time taken to search. vectors are stored the same without any modification/change
Locality Sensitive Hashing	3.97	Vectors are encoded by mapping the data points to the buckets. The memory used to store the indexes is reduced
Product Quantization	30.04	It doesn’t provide the optimal solution. PQ with IVF improves the search speed
Tree and Graph(Annoy)	18.14	It gives a model with good accuracy but at the cost of performance (time). It builds trees and

		each tree is constructed using random splits of data.
HNSW using faiss library	8.87	It is an algorithm with high performance. It has great search quality with good search time.
HNSW using nmslib	1608.68	It is an optimal solution but it takes long time to build and search the indices.

Comparing all the ANN algorithms based on the time Exhaustive search takes less time out of all the other algorithms but this algorithm consumes more memory as it stores the vectors and indexes without modification.

When compared to the quality of the algorithm or the accuracy of the model HNSW, Tree and Graph (annoy) are the best algorithms.

For the algorithm with great accuracy and good search time, HNSW is the best algorithm.