Scalable Data Mining (Autumn 2023)

Assignment 3: ANN Search (Total Marks: 100)

Introduction:

Approximate Nearest Neighbor (ANN) search is a fundamental problem in computer science and data analysis. In this assignment, you will explore the implementation of ANN search using three different algorithms: **Locality-Sensitive Hashing (LSH)**, **FAISS**, and **Hierarchical Navigable Small World (HNSW)**. These algorithms are crucial for efficient similarity search in large datasets and have applications in image retrieval, recommendation systems, and more.

Dataset Description:

The **MNIST dataset** is a collection of 28x28 pixel grayscale images of handwritten digits (0-9). It contains 60,000 training and 10,000 testing images, making it an ideal dataset for various machine-learning tasks. For this assignment, we will focus on using MNIST for ANN search.

Implementation Tasks:

1. Locality-Sensitive Hashing (LSH) Implementation:

LSH is a method for approximate nearest neighbor search. It works by hashing data points in such a way that similar points map to the same "buckets" with high probability. Your task is to implement LSH for ANN search with the MNIST dataset.

2. FAISS Implementation:

FAISS is a powerful library designed for efficient similarity search and clustering of large datasets. In this part, you will use the FAISS library to implement an ANN search on the MNIST dataset.

We recommend referring to the official FAISS documentation and examples for guidance in your implementation.

3. Hierarchical Navigable Small World (HNSW) Implementation:

HNSW is a more recent and highly efficient approach for ANN search, especially for high-dimensional data. Implement HNSW for ANN search with the MNIST dataset.

Results and Comparison:

For each implemented algorithm (LSH, FAISS, and HNSW), evaluate their performance on the MNIST dataset. Measure **recall** (the fraction of true neighbors found), and **precision** (the fraction of found neighbors that are true neighbors).

Report:

In your report, make sure to include the following sections:

Methodology:

- Describe the implementation details for LSH, FAISS, and HNSW.

Results:

- Present the performance metrics for each algorithm.
- Include tables and figures as necessary.

Comparison:

- Analyze and compare the results
- Discuss the strengths and weaknesses of each approach

Conclusion:

Summarize the findings, highlight key takeaways, and suggest real-world applications for ANN search using these algorithms.

Submission Guidelines:

You should submit the following in zipped format (Rollnumber_AssignmentNo.zip):

- **Report** with all the contents as mentioned. **(60 marks)**
- **Python codes:** Code in .py and .ipynb are acceptable. **(40 marks)**