The topic for my project proposal is Locality Sensitive Hashing (LSH). LSH provides an efficient approach to similarity search. I found the basic idea of the topic to be not too challenging, and was interested by its countless applications (e.g., search engines, dating apps, music recommendations, plagiarism checking, etc.).

With very large datasets, it is impossible to compare values against every other value. The goal of LSH is to reduce the number of comparisons to a smaller subset. This smaller subset should already be grouped such that it contains values that are expectedly similar. The search time within the subset is sublinear, which is much faster than searching the entire set.

To create the subsets, the values are passed through a hash function. In contrast to what we have learned, LSH hash functions want to maximize collisions, given that the hashed values are similar. Similar hash values should indicate similar pre-hash values.

From my research, I found LSH can be implemented in different ways. Two popular approaches are:

1. Shingling, MinHashing, and Banding
2. Random Projection

I plan on doing further research into both approaches. Then, I hope to implement at least one approach, and perhaps both in order to compare results.

I found this repository that has large datasets built for similarity search testing, which I hope to eventually use to evaluate my implementation - https://github.com/brmson/dataset-sts