Name : Pawan Bathe Roll No.:CS18M519

CS6745: Mining Massive DataSets

Assignemnt 2 Report

1. How does the jaccard similarity estimated using minhashing vary when the number of rows is varied?

   Jaccard similarity estimated using minhashing decreases when numbers of rows are increased.

2. How much do the calculated jaccard similarity and the estimated similarity vary?

   Calculated jaccard similarity and the estimated similarity do not vary much, they are nearly equal.

3. How does the performance of LSH compare to brute-force, with respect to average similarity and running time?

   Average similarity of brute force approach is better. Also LHS running time is more than brute force approach. Time increase when number of permutations are increased.

4. How does this vary with dataset size, k, r, and number of rows in signature matrix?

   Running time increases with increase in dataset size, k, r and number of rows in signature matrix. Average similarity decreases with increase n dataset size, k, r and number of rows in signature matrix.