Laura Colbran, Jimin Yoo

Professor Dave Musicant

COSC 324

January 21, 2015

**Locality Sensitive Hashing Report**

In this assignment, we compared two approaches for obtaining the average Jaccard similarity among k number of nearest neighbors (or, in this case, k number of similar documents)—simple Brute Force (BF) and Locality Sensitive Hashing (LSH). We changed variables such as data size (number of documents), k (number of nearest neighbors), r (number of rows in a band), and number of permutations, and compared the Jaccard similarity obtained by the two approaches.

Overall, we found that the gap between the Jaccard similarties of the two approaches was small, keeping the data size and k constant. However, the gap between the two approaches’ Jaccard similarities got bigger when r, the number of rows in a band, is large or when the number of permutations was small.

**Figure 1. Data size (number of documents to read) and averaged Jaccard similarity**

Both brute force approach and LSH approach’s Jaccard similarities remain similar throughout the changning data size. The average Jaccard similarity gets bigger as the data size increases, which makes sense, since the larger the number documents we read, the larger the probability that we can find documents that are more similar to our target document.

**Figure 2. K (number of nearest neighbors and average Jaccard similarity)**

Both approaches’ Jaccard similarities remain similar. The avearge similarity gets smaller as k gets larger. This is probably because we have to start including not-so-similar documents (or at least, less similar than the ones closest to the target document) as we try to find larger number of neighbors.

**Figure 3. Average jaccard similarity and number of permutations**

Since all that was being altered here was the number of permutations put in the signature matrix, the brute force method was unchanged. LSH got closer to the brute force method as there were more permutations. This makes sense, given the theory that if one used every possible permutation, it would equal the Jaccard similarity exactly.

**Figure 4. Average Jaccard similarity and number of rows in a band**

Again, the LSH method was what was being altered, so the brute force was unchanged. As you increase the number of rows per band, LSH gets worse at estimating the overall similarity. Again, this makes sense. Larger bands means it’s less likely that a document would qualify as a candidate pair, which means the knn has fewer options to choose from.

Appendix

Figure 1. Data size (number of documents to read in) and running time

Figure 2. K (number of nearest neighbors) and running time