Introduction to locality sensitive hashing

Andee Kaplan

Duke University
Department of Statistical Science
andrea.kaplan@duke.edu

February 9, 2018

Slides available at http://bit.ly/cimat-lsh

Goal and outline

Finding similar items

Jaccard similarity

Shingling

Your turn (shingling and Jaccard similarity by hand)

Useful packages/functions in R

Example data

Your turn (shinging and Jaccard similarity with R)

Hashing

Why do it?

Similarity preserving summaries of sets

Characteristic matrix

Minhashing

LSH (avoid pairwise comparisons)

Banding and buckets

Your turn (banding in R)

Putting it all together

“Easy” LSH in R

Evaluation