Data Mining: Learning from Large Data Sets - Fall Semester 2015

mmarti@student.ethz.ch
trubeli@student.ethz.ch

October 15, 2015

Approximate near-duplicate search using Locality Sensitive Hashing

In this project we used linear hashing to approximate the similarity between videos, represented by a list of shingles. The first step in our solution was to produce a signature matrix. This matrix is obtained by using a min hash algorithm on each list of shingles. For every i^{th} shingle in a video we pick two random numbers a_i and b_i which are coprime. The procedure for computing the signature matrix is described as follow:

Algorithm 1 Min Hash Algorithm

```
1: procedure MinHash(N, K)                    ▷ K hash fonction applied on N shingles
2:     Initialize:
       w_l ← ∞, l = 1, ..., k
3:     for i = 1 to n do
4:         for j = 1 to k do
5:             if h_j(n_i) < w_j then
6:                 w_j ← h_j(n_i)
```