Katie Briggs

Activities Data Mining

3.1.1

A={1, 2, 3, 4}, B={2, 3, 5, 7}, and C={2, 4, 6}

SIM (A, B) = 2/6 = 1/3

SIM (A, C) = 2/5

SIM (B, C) = 1/6

3.2.1

“The”, “he “, “e m”, “ mo”, “mos”, “ost”, “st “, “t e”, “ ef”, “eff” .

“The most effective”, “most effective way”, “effective way to”, “way to represent”, “to represent documents”, “represent documents as”, “documents as sets”, “as sets for”, “sets for purpose”, “for purpose of”

3.3.3

Element S1 S2 S3 S4 2x+1 mod 6 3x+2 mod 6 5x+2 mod 6 0 0 1 0 1 1 2 2 1 0 1 0 0 3 5 1 2 1 0 0 1 5 2 0 3 0 0 1 0 1 5 5 4 0 0 1 1 3 2 4 5 1 0 0 0 5 5 3

Minhash Signature Matrix S 1 S 2 S 3 S 4 h1(0) \ 1 \ 1 h2(0) \ 2 \ 2 h3(0) \ 2 \ 2 h1(1) \ 1 \ 1 h2(1) \ 2 \ 2 h3(1) \ 1 \ 2 h1(2) 5 1 \ 1 h2(2) 2 2 \ 2 h3(2) 0 1 \ 0 h1(3) 5 1 1 1 h2(3) 2 2 5 2 h3(3) 0 1 5 0 h1(4) 5 1 1 1 h2(4) 2 2 2 2 h3(4) 0 1 4 0 h1(5) 5 1 1 1 h2(5) 2 2 2 2 h3(5) 0 1 4 0

The final minhash signature matrix is: S1 S2 S3 S4 5 1 1 1 2 2 2 2 0 1 4 0

h3 is a true permutation.

similarities 1-2 1-3 1-4 2-3 2-4 3-4 col/col 0 0 0.25 0 0.25 0.25 sig/sig 0.33 0.33 0.67 0.67 0.67 0.67

The estimated Jaccard similarities are not close to the true.

3.4.1

Values of the S-curve for b=10 and r=3 s 1 − （1 − s 𝑟） 𝑏

0.0100

0.0772

0.2394

0.4839

0.7369

0.9123

0.9850

0.9992

1.0000

Values of the S-curve for b=20 and r=6

0.0000

0.0013

0.0145

0.0788

0.2702

0.6154

0.9182

0.9977

1.0000

Values of the S-curve for b=50 and r=5

0.0005

0.0159

0.1145

0.4023

0.7956

0.9825

0.9999

1.0000

1.0000