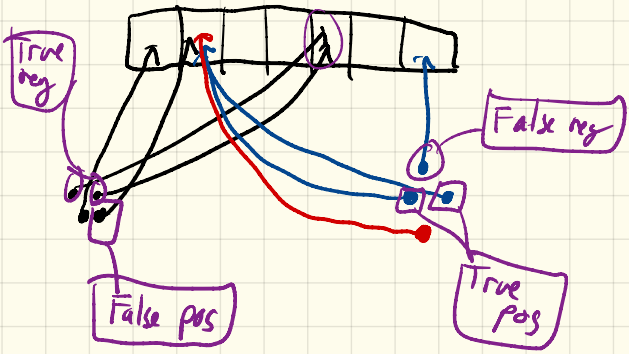
LSH

11/12



Document Sim

Jaccard s.m

given 2 sets A & B

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Shingling:

$k \in \mathbb{Z}^+$ k-shingle document (usually $k=9$ is a good rule of thumb)

document

xyzabaa

{xyz, yza, zab, aba, baa}

Presere sim between sets

characteristic matrix: sets $S_1, \dots, S_n$ as cols
elts $a_1, \dots, a_m$ as rows
$m \times n$ matrix st.
$\forall (i,j)$ w/ $a_j \in S_i$ $M(i,j) = 1$
else 0

4 sets $\{a, b, c, d, e\}$

$S_1 = \{a, d\}$
$S_2 = \{c\}$
$S_3 = \{b, d, e\}$
$S_4 = \{a, c, d\}$

	S_1	S_2	S_3	S_4
a	1	0	0	1
b	0	0	1	0
c	0	1	0	1
d	1	0	1	1
e	0	0	1	0

characteristic matrix

(In principal) idea of Minhashing

Pick a permutation

$$P = (b, e, a, d, c)$$

let $h_P(S_i)$ be the first row w/ a 1 in col S_i

Element	S_1	S_2	S_3	S_4
b	0	0	1	0
e	0	0	1	0
a	1	0	0	1
d	1	0	1	1
c	0	1	0	1

Min hashing & Jaccard Sim

Thm Prob that min hash for random permutation of rows produces the same min hash for two sets $\cong$ Jaccard sim between the the two sets

Consider 2 sets S_1 S_2

3 types of events:

Type X: both have a 1 in the row
Type Y: one has a 1 other has a 0
Type Z: both have a 0

$\Rightarrow$ most rows are type Z
type X and the type Y
↑ end up in num of (and denom) Jaccard sim
↑ end up in denom

$$J(S_1, S_2) = \frac{X}{X+Y}$$

X ← # of type X events
Y ← number of type Y events

Pick k permutations (100s – 1000s)

$\{P_1, \ldots, P_k\}$

Compute the min hash signature for each document

$$S_1 = [h_{P_1}(S_1), h_{P_2}(S_1), h_{P_3}(S_1) \ldots, h_{P_k}(S_1)]$$