Min Hashing

$A = \{0, 1, 2, 3, 6\}$

$B = \{1, 2, 4, 6, 8\}$

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

$$= \frac{|\{1, 2, 6\}|}{|\{0, 1, 2, 3, 4, 6, 8\}|} = \frac{3}{7}$$

Data set of Sets $\{A_1, A_2, \ldots A_n\}$

$n = 1$ million

Doc → set → vector

$D_i \xrightarrow[\text{k-gram}]{} A_i \xrightarrow[\text{min hash}]{} v_i \in \mathbb{R}^k$

length k

Property

as $k \to$ larger

$JS(A_i, A_j) \approx \hat{JS}(v_i, v_j)$ closer

Matrix / Vector Set Representation

Exact

Represent set A_i as bit vector $b_i \in \{0,1\}^n$

$n = 6$

$$A_1 = \{1, 2, 5\}$$
$$A_2 = \{3\}$$
$$A_3 = \{2, 3, 4, 6\}$$
$$A_4 = \{1, 4, 6\}$$

$$\begin{array}{c|cccc} & b_1 & b_2 & b_3 & b_4 \\ \hline 1 & 1 & 0 & 0 & 1 \\ 2 & 1 & 0 & 1 & 0 \\ 3 & 0 & 1 & 1 & 0 \\ 4 & 0 & 0 & 1 & 1 \\ 5 & 1 & 0 & 0 & 0 \\ 6 & 0 & 0 & 1 & 1 \end{array}$$

Min Hashing

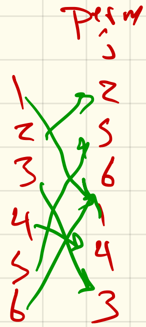
1. Randomly Reorder (permute) the rows

2. For each set / column find the top / first 1 bit $m(A_i) =$ top 1 bit

3. Repeat Step 1 & 2 $\underline{\underline{k}}$ times ≈ 160

$$v_i = \begin{bmatrix} m_1(A_i) \\ m_2(A_i) \\ \vdots \\ m_k(A_i) \end{bmatrix}$$

Step 1, 2
- Reorder
- top 1 bit



b_1	b_2	b_3	b_4
1	0	1	0
1	0	0	0
0	0	1	1
1	0	0	1
0	0	1	1
0	1	1	0

$m_j(b_i) = 2, 3, 2, 6$

$\hat{JS}(i, i') = 1$ if $m(i) = m(i')$

0 otherwise

$E[\hat{JS}(i, i')] = JS(A_i, A_{i'})$

$$\Pr[m(i) = m(i')] = E[\hat{JS}(i, i')] = JS(A_i, A_{i'})$$

T_x = x rows w/ 1 both columns

T_y = y rows w/ 1 in exactly 1 column

T_z = z rows w/ 0 both columns

$$JS(A_i, A_{i'}) = \frac{x}{x+y} = \Pr[m(i) = m(i')]$$

	b_i	$b_{i'}$
T_y	1	0
T_y	0	1
T_x	1	1
T_z	0	0
T_y	0	1

$$\Pr[m(i) = m(i')] = 1$$

iff top row type x

ignoring type z

How big should k be?

\# permutations.

$\boxed{\text{Chernoff-Hoeffding}}$

k R.V. iid $X_1, X_2, \dots X_k$ $\quad E[X_i] = \mu$

$$M = \frac{1}{k} \sum_{i=1}^{k} X_i \qquad E[M] = E[X_i]$$

$X_i \in \{0, 1\}$

$$\Pr\big[|M - E[M]| > \varepsilon\big] \le \overbrace{2\exp(-2\varepsilon^2 k)}^{\delta = \text{prob. of failure}}$$

$= 0.1$

error tolerance 0.05

$$0.1 = 2e^{-2(0.05)^2 k}$$

$$\ln(0.05) = -2\left(\frac{1}{400}\right) k$$

$$k = 200 \cdot \ln\left(\frac{1}{0.05}\right) \approx 600$$

Fast Min Hash Signatures

Set A_i to vector $v_i \in \mathbb{Z}^k$

k hash functions $h_j : [n] \to [n']$

$h_j \in \mathcal{H}$

don't need to know

instead

$h : \Sigma^3 \to [n']$

alphabet of char

Algo: Init $v_i(j) = \infty$ for all $j \in [1 \ldots k]$

for $x \in A_i$ do
  for $j = 1$ to k
    if $(h_j(x) < v_i(j))$
      $v_i(j) \leftarrow h_j(x)$

one pass over data

$$\hat{JS}_k(v_i, v_{i'}) = \frac{1}{k} \sum_{j=1}^{k} \begin{cases} 1 & \text{if } v_i(j) = v_{i'}(j) \\ 0 & \text{o.w.} \end{cases}$$

Example Fast Min Hash

Turns set $A_i = \{1, 2, 4\}$ into vector v_i

use 3 hash functions $h_1, h_2, h_3 \in \mathcal{H}$

h_1		h_2		h_3	
$1 \to 7$		$1 \to 2$		$1 \to 6$	
$2 \to 4$		$2 \to 10$		$2 \to 4$	
$3 \to 10$		$3 \to 3$		$3 \to 8$	
$4 \to 1$		$4 \to 9$		$4 \to 5$	
$5 \to 7$		$5 \to 5$		$5 \to 9$	

use the same hash functions for each mapping $A_i \to v_i$

Pass through A_i

Step

0: $v_i = (\infty, \infty, \infty)$

$x = 1$

h_1	h_2	h_3
7	2	6

1: $v_i = (7, 2, 6)$

$x = 2$

h_1	h_2	h_3
4	10	4

$4 < 7$, $4 < 6$

2: $v_i = (4, 2, 4)$

$x = 4$

h_1	h_2	h_3
1	9	5

$1 < 4$

3: $\boxed{v_i = (1, 2, 4)}$ ← the min hash signature of A_i