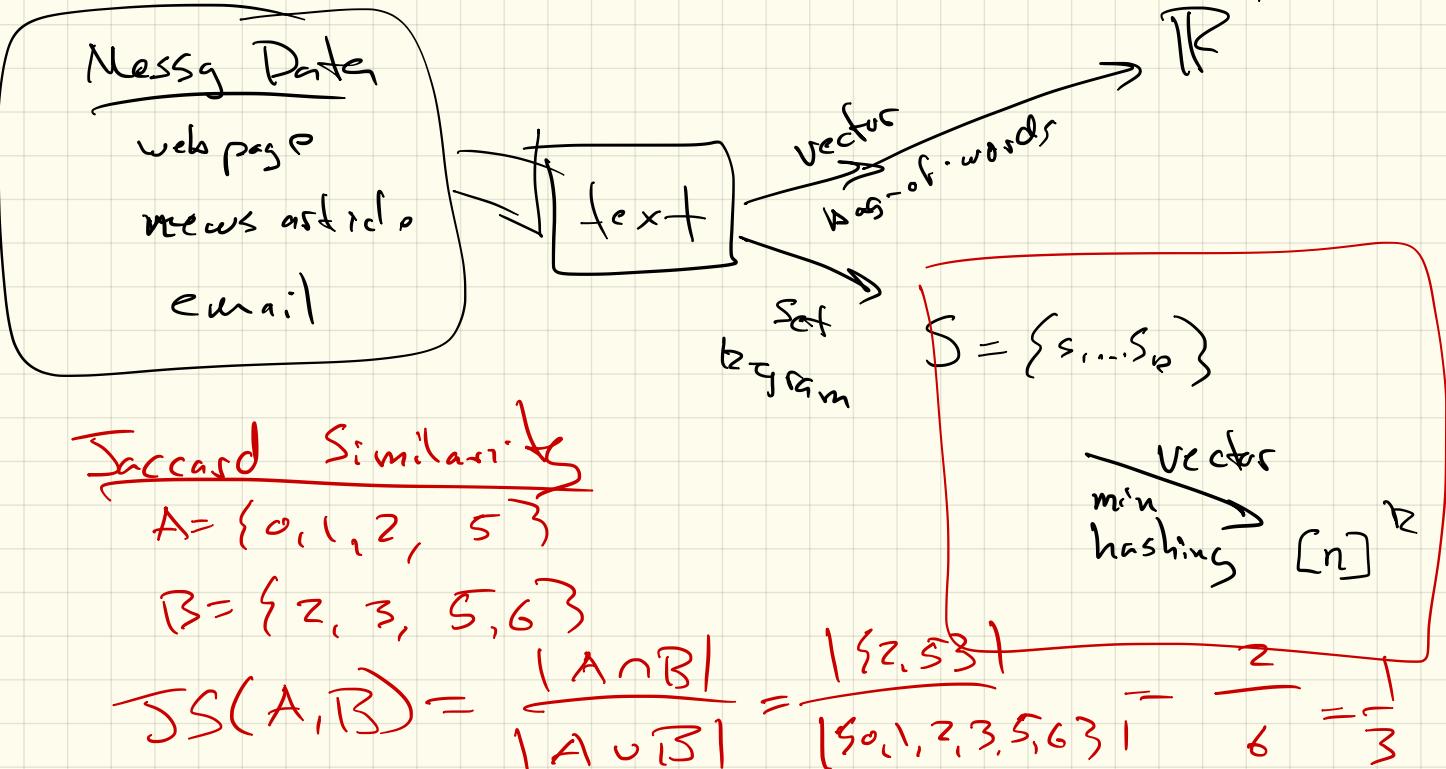


# Min Hashing



Family Hash Functions  $\mathcal{F}$

all  $h \in \mathcal{F}$  are permutations

$$S_1 = \{1, 2, 5\}$$

$$S_2 = \{3\}$$

$$S_3 = \{2, 3, 4, 6\}$$

$$S_4 = \{1, 4, 6\}$$

Sets  $S_i \subset [n]$

Randomly draw  $h_T \in \mathcal{F}$

then  $h_T$  deterministic

$h_{T,i} : [n] \rightarrow [n]$

$\hookrightarrow$  domain

Important:  
 $h_{T,i}$  has order

$$h_{T,i}(1) \rightarrow 7$$

$$h_{T,i}(2) \rightarrow 3$$

$$h_{T,i}(5) \rightarrow 4$$

$$g_i(\xi) = \min_{s \in S_i} h_{T,i}(s)$$

$$\text{e.g. } g_1(\xi_1) = 3$$

$\tau_i$  = permutation

domain	1	2	3	4	5	6	7	8	9	10	$[n] = [10]$
$h_{T,i}$	1	7	3	6	5	2	8	9	4	10	$\rightarrow 1$
$h_{T,2}$	4	3	2	6	9	10	1	9	8	5	$\rightarrow 2$

Way to go from

Set  $S_i \subset [n]$

$g_1(S_i) \in [n]$

$\times \mathbb{R}$

$$g_1(S_i) \rightarrow v_1$$

$$g_2(S_i) \rightarrow v_2$$

:

:

$$g_{t^*}(S_i) \rightarrow v_{t^*}$$

$$v = (v_1, v_2, \dots, v_{t^*}) \in [n]^{t^*}$$

$$v(S) = (3, 1, 5, \dots, x)$$

$$\underline{v(S)} = (6, 2, 6, \dots, x)$$

$$\text{Ansatz } \sum_{i=1}^{t^*} S_i = \frac{1}{t^*} \sum_{i=1}^{t^*} \begin{cases} 1 & \text{if } v_i(S) = v_i(S) \\ 0 & \text{o.w.} \end{cases}$$

$$S_1 = \{1, 2, 5\}$$

$$S_2 = \{3\}$$

$$S_3 = \{2, 3, 4, 6\}$$

$$S_4 = \{1, 4, 6\}$$

For any two sets  $S_1, S_2$

$h_{S_1}, h_{S_2}, \dots, h_{S_k} \stackrel{\text{iid}}{\sim}$

$$E[\overline{SS}(S_1, S_2)] = \overline{SS}(S_1, S_2)$$

---

$$\begin{aligned} E[\overline{SS}(S_1, S_2)] &= E\left[\frac{1}{k^2} \sum_{i=1}^k \prod_{j=1}^k \mathbb{1}(g_i(S_1) = g_j(S_2))\right] \\ &= \frac{1}{k^2} \sum_{i=1}^k E\left[\prod_{j=1}^k \mathbb{1}(g_i(S_1) = g_j(S_2))\right] \end{aligned}$$

$$\underline{P_r[g_i(S_1) = g_i(S_2)] = \overline{SS}(S_1, S_2)}$$

---

Decompose  $[n] \rightarrow A, B, C$

$\overline{SS} = \frac{|A|}{|A| + |B|}$

A objects hashed to by  $s \in S_1$  and  $s \in S_2$

B objects hashed to by  $s \in S_1$  or  $s \in S_2$ , not both

C objects hashed to key  $x \in S_1 \cup S_2$

Fast Min Hash

$$\hat{g}_i \approx g_i : (\text{set } \subset [n]) \rightarrow [n]$$

choose (random) hash function

$$f_i : [n] \rightarrow [m]$$

$$m > n$$

$$v_i = \infty$$

$H_c$

Domain

$$\text{for } j=1 \text{ to } k \quad S = \{x_1, x_2, \dots, x_j\}$$

for  $i=1$  to  $k$

$$\text{if } (f_{i,j}(x_j) \in v_i)$$

$$v_i \leftarrow h_i(x_j)$$

Return  $V = (v_1, v_2, \dots, v_k)$

Domain

$$\Sigma = [n]$$

every set

$$S \subset \Sigma$$

$$x \in S$$

$$S \in 2^{\Sigma}$$

$\epsilon$ -H

How large should  $\epsilon$  be?

$$X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \mu \quad X_i \in [0, 1]$$

$$A = \frac{1}{n} \sum_{i=1}^n X_i$$

$$E[A] = E[X_i] = \mu$$

$$\Pr[|A - \mu| > \epsilon] \leq 2 \exp(-2\epsilon^2 n) \leq \frac{\delta}{0.01}$$

$\epsilon = 0.05$

$$\delta = 2 \exp(-2(0.05)^2 n)$$

$$\ln\left(\frac{\delta}{2}\right) = -2\left(\frac{\epsilon}{\mu}\right)^2 n$$

$$\ln\left(\frac{\delta}{2}\right) = -2\left(\frac{\epsilon}{\mu}\right)^2 n \Rightarrow n = \frac{400}{2} \ln(200)$$

$$(2 = 200 \ln(200)) = 0.60$$

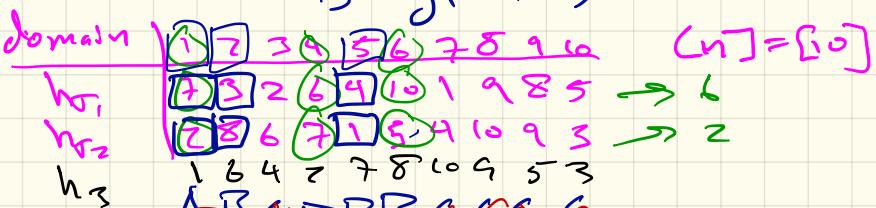
$$\begin{aligned}
 h_{\sigma_1}(1) &\rightarrow 7 \\
 h_{\sigma_1}(2) &\rightarrow 3 \\
 h_{\sigma_1}(5) &\rightarrow 4
 \end{aligned}$$

$h_{\sigma_2}$

$\sigma_i$  = permutation

$$g_i(s) = \min_{s \in S} h_{\sigma_i}(s)$$

$$\text{e.g. } g_i(s_i) = 3$$



$$\begin{aligned}
 s, \quad & h_1 & h_2 \\
 & 7(3)4 & 28(1)
 \end{aligned}$$

$$\begin{aligned}
 s_2, \quad & h_1 & h_2 \\
 & 7(6,10) & (2)75
 \end{aligned}$$

0

8

1

