Min Hashing

Messy Data
- web page
- news article
- email

$\text{Messy Data} \rightarrow \text{text}$

$\text{text} \xrightarrow[\text{bag-of-words}]{\text{vector}} \mathbb{R}^d$

$\text{text} \xrightarrow[\text{k-gram}]{\text{set}} S = \{s_1, \dots s_k\} \xrightarrow[\text{min hashing}]{\text{vector}} [n]^k$

Jaccard Similarity

$A = \{0, 1, 2, 5\}$

$B = \{2, 3, 5, 6\}$

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|\{2,5\}|}{|\{0,1,2,3,5,6\}|} = \frac{2}{6} = \frac{1}{3}$$

Family Hash Functions $\mathcal{H}$

all h_σ, where σ perm $[n] \to [n]$

Randomly draw $h_{\sigma_1} \in \mathcal{H}$

then h_{σ_i} deterministic

$D \leftarrow$ domain

$h_{\sigma_i} : [n] \to \underbrace{[n]}_{\text{important: has } \underline{\text{order}}}$

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

$S_2 = \{3\}$

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

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

Sets $S_i \subset [n]$

	h_{σ_2}
$h_{\sigma_1}(1) \to 7$	2
$h_{\sigma_1}(2) \to 3$	8
$h_{\sigma_1}(5) \to 4$	1

$$g_1(S) = \min_{s \in S} h_{\sigma_1}(s)$$

e.g. $g_1(S_1) = 3$

σ_i = permutation

domain	1	2	3	4	5	6	7	8	9	10	
h_{σ_1}	7	3	2	6	4	10	1	9	8	5	$\to$ 6
h_{σ_2}	2	8	6	7	1	5	4	10	9	3	$\to$ 2

$[n] = [10]$

$\boxed{\begin{array}{l}\text{Way to go from} \\ \text{set } S_i \subset [n] \\ g_1(S_i) \in [n]\end{array}}$ $\times k$

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

$S_2 = \{3\}$

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

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

$$\left.\begin{array}{l} g_1(S_i) \to v_1 \\ g_2(S_i) \to v_2 \\ \vdots \\ g_k(S_i) \to v_k \end{array}\right) \to v = (v_1, v_2, \dots v_k) \in [n]^k$$

$v(S_1) = (3, 1, , 6 \dots , x)$

$v(S_4) = (6, 2, 6 \dots x)$

$\quad\quad\quad\quad 0 \quad 0 \quad 1$

$$\text{Apr } \tilde{JS}(S_1, S_4) = \frac{1}{k} \sum_{i=1}^{k} \begin{cases} 1 & \text{if } v_i(S_1) = v_i(S_4) \\ 0 & \text{o.w.} \end{cases}$$

$= \frac{1}{3}$

For any two sets S_1, S_2

$h_{\sigma_1}, h_{\sigma_2}, \ldots h_{\sigma_k} \sim_{iid} \mathcal{H}$

$$E\left[\hat{JS}(S_1, S_2)\right] = JS(S_1, S_2)$$

$$E\left[\hat{JS}(S_1, S_2)\right] = E\left[\frac{1}{k}\sum_{i=1}^{k} \mathbb{1}\left(g_i(S_1) = g_i(S_2)\right)\right]$$

$$= \frac{1}{k}\sum_{i=1}^{k} E\left[\mathbb{1}\left(g_i(S_1) = g_i(S_2)\right)\right]$$

$$\Pr\left[g_i(S_1) = g_i(S_2)\right] = JS(S_1, S_2)$$

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

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 hashe to by $x \notin S_1 \cup S_2$

$$JS = \frac{|A|}{|A| + |B|}$$

Fast Min Hash

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

choose (random) hash function

$f_i : [n] \to [m] \quad m > n$

$v_i = \infty \quad \forall i$

for $j = 1$ to ℓ $\quad \boxed{S = \{x_1, x_2, \dots x_\ell\}}$ Domain

for $i = 1$ to k

if $(f_i(x_j) < v_i)$

$v_i \leftarrow h_i(x_j)$

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

Domain

$\Omega = [n]$

every set

$S \subset \Omega$

$x \in S$

$S \in 2^{\Omega}$

C-H

How large should k be?

$X_1, X_2, \ldots X_k \underset{iid}{\sim} \mu \qquad X_i \in [0,1]$

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

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

$$\Pr[|A-\mu| > \varepsilon] \leq 2\exp(-2\varepsilon^2 k) \leq \delta_{0.01}$$

$\varepsilon = 0.05$

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

$$\ln\left(\frac{\delta}{2}\right) = -2\left(\frac{1}{20}\right)^2 k$$

$$\ln\left(\frac{2}{\delta}\right) = 2\left(\frac{1}{20}\right)^2 k \implies k = \frac{400}{2}\ln(200)$$

$$k = 200\ln(200) = 1060$$

$h_{\sigma_1}(1) \to 7$ | h_{σ_2}: 2

$h_{\sigma_1}(2) \to 3$ | 8,

$h_{\sigma_1}(5) \to 4$ | 1

$$g_1(S) = \min_{s \in S} h_{\sigma_1}(s)$$

e.g. $g_1(S_1) = 3$

σ_i = permutation

domain	1	2	3	4	5	6	7	8	9	10	
h_{σ_1}	$\boxed{7}$	$\boxed{3}$	2	6	$\boxed{4}$	10	1	9	8	5	$\to$ 6
h_{σ_2}	$\boxed{2}$	$\boxed{8}$	6	7	$\boxed{1}$	5	4	10	9	3	$\to$ 2
h_3	1	6	4	2	7	8	10	9	5	3	
	A	B	~~C~~	B	B	B	~~C~~	~~C~~	~~C~~	~~C~~	

$[n] = [10]$

	h_1	h_2	h_3
S_1	7 3 4	2 8 1	1 ~~6~~ 7
S_2	7 6, 10	2 7 5	1 2 8
	0	8	1