Locality Sensitive Hashing (LSH)

Note Title 1/27/2016

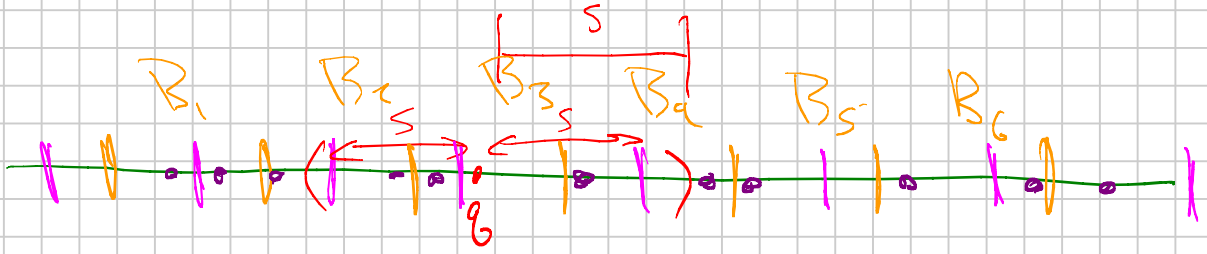
Doc → Sets → Vectors

$$E[S(V_1, V_2)] = JS(S_1, S_2)$$

Many Documents $\# = n = 1$ million

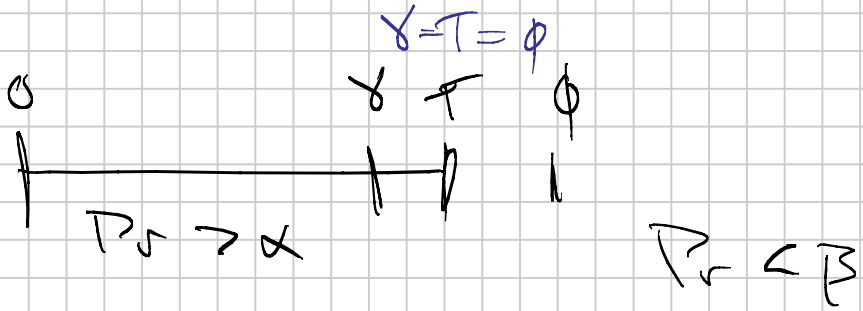
(Q1) What items are similar? n^2

(Q2) Given query, which in set are similar? n

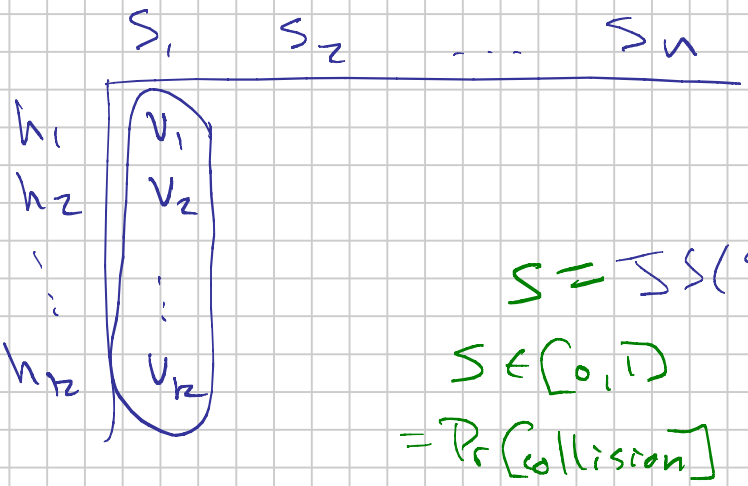


$h \in \mathcal{H}$ $(\gamma, \phi, \alpha, \beta)$ - similar

- $Pr[h(a) = h(b)] \geq \alpha$ if $d(a,b) \leq \gamma$
- $Pr[h(a) = h(b)] \leq \beta$ if $d(a,b) \geq \phi$



$h_i : [n] \to [n \log n]$



$S = \sum S(s_1, s_2) = E\left[\frac{1}{k} \sum_i (h_i(s_1) = h_i(s_2))\right]$

$S \in [0,1]$

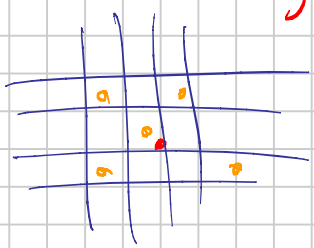
$= Pr[\text{collision}]$

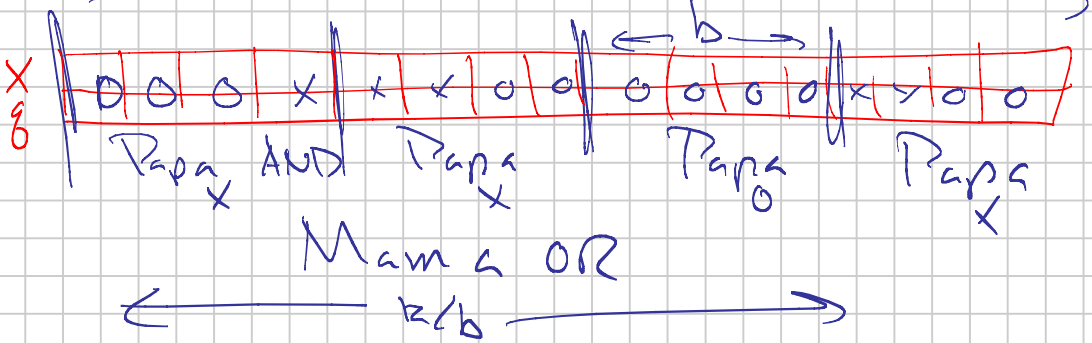
- similar if all hash fxn collide
  - too harsh → Papa bear
- similar if any hash fxn collide
  - too nice → Momma bear

→ Baby bear

Try until 1st collision
↳ then count

Banding





k/b

$r = k/b$

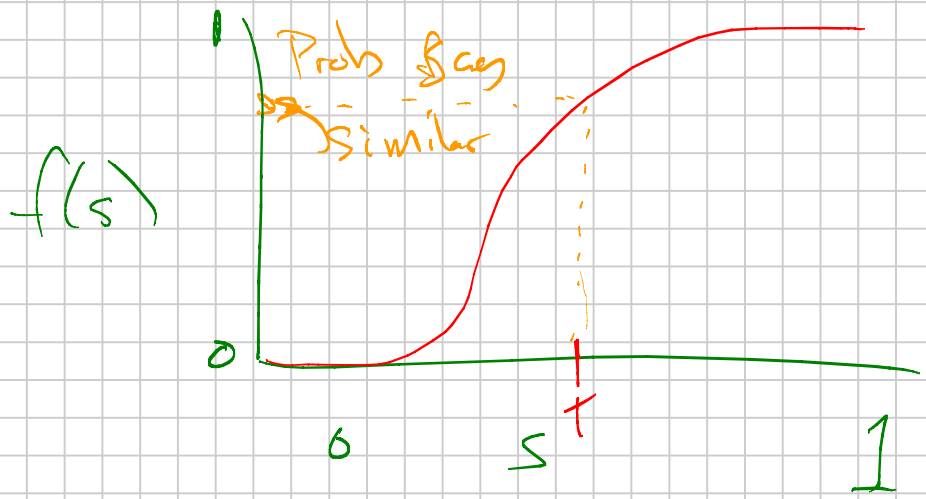
S = Prob collision

S^b = Prob collision in 1 band

$(1 - S^b)$ = Prob no coll in 1 band

$(1 - S^b)^r$ = Prob no coll in r bands

$f(s) = 1 - (1 - s^b)^r$ = Prob of coll in any of r bands



Budget k hash fxns

↳ # hash in band b

↳ # bands r $\quad r \cdot b \leq k$

Similarity threshold T

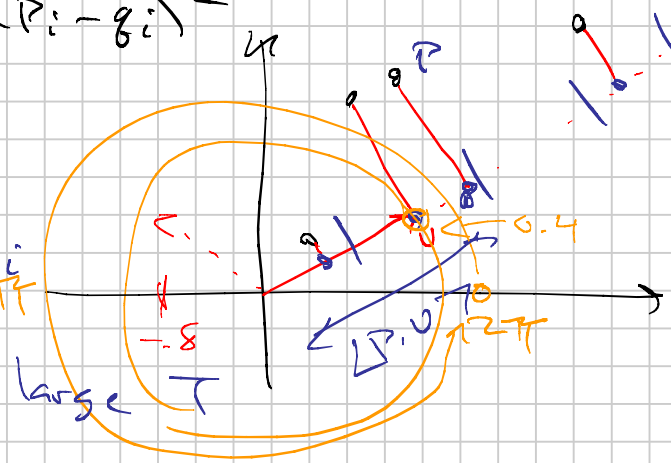
$b \approx -\log_T(k) \qquad r = k/b$

$$d_E(p, q) = \|p - q\| = \sqrt{\sum_{i=1}^{d} (p_i - q_i)^2}$$

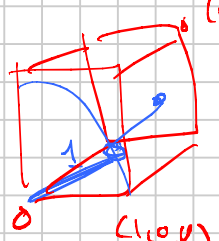
$p \in \mathbb{R}^d$

$h_v \in \{ \pi_v(p) = \langle p, v \rangle = \sum_{i=1}^{d} p_i \cdot v_i \}$

$h_{v,s}(p) = \lfloor \langle p, v \rangle + s \bmod (T\gamma) \rfloor$

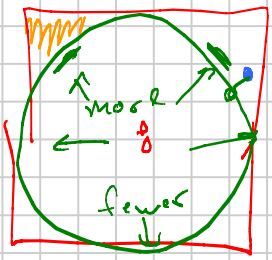


How to pick $v \leftarrow$ rand unit vector in $\mathbb{R}^d$



$r \in \mathbb{R}^d$

$r_i \leftarrow Unif(0,1)$

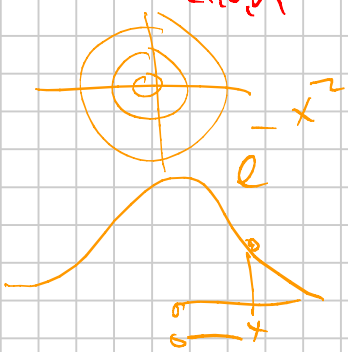


$a \in Unif[0,1]$

$b \in$ Rand-bit in $\{0,1\}$

$g \in \mathbb{R}^d$

each i $g_i \sim N(0,1)$



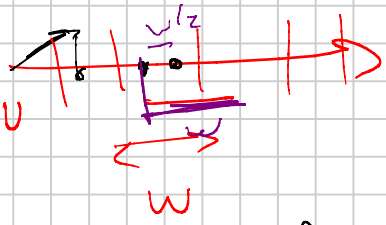
Box-Mueller Transf

$u_1, u_2 \sim Unif[0,1]$

$g_1 = \sqrt{-2\ln(u_1)}\cos(2\pi u_2)$

$g_2 = \sqrt{-2\ln(u_1)}\sin(2\pi u_2)$

close case



$\langle p, v \rangle - \langle q, v \rangle$

$=$

$\langle p - q, v \rangle \le \|p - q\|$

$\|p - q\| > 2w$

far case