L5: Min Hashing

Jan 22
2025

Jeff Phillips

Input: Document ≡ string words, or characters.

↳ k-grams

↳ Sets

Sets

$D_1 \rightarrow S_1 \rightarrow v_1$

$D_2 \rightarrow S_2 \rightarrow v_2$

$D_3 \rightarrow S_3 \rightarrow v_3$

$v_j \in \mathbb{R}^{1000}$

MH

Jaccard Sim

$$JS(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

min hashing

↳ (could have)

Vector

$D_1 \rightarrow v_1$

$D_2 \rightarrow v_2$

bag-of-words

$v_j \in \mathbb{R}^{100,000}$

w/ bigram $\mathbb{R}^{(100,000)^2}$

Round-about Min-Hashing

Set → matrix → vector

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

$S_2 = \{3\}$

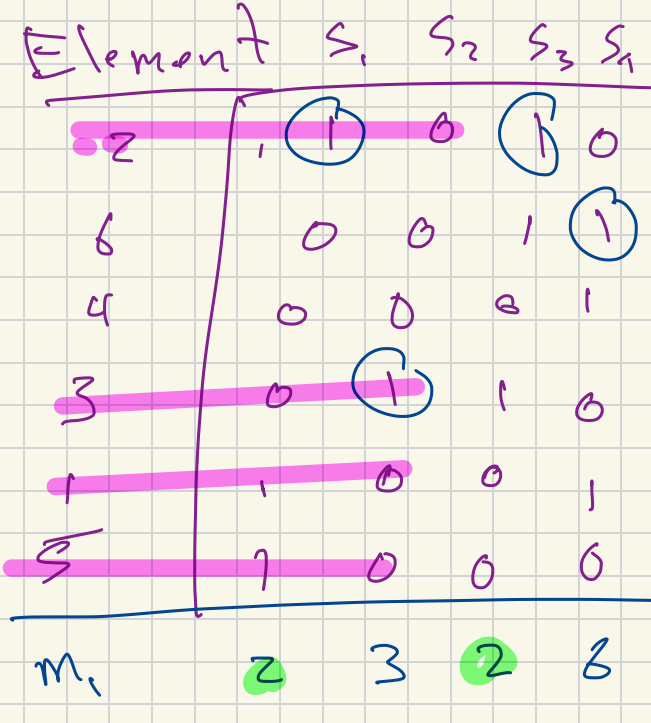
$S_3 = \{2, 3, 4, 6\}$

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

Element	S_1	S_2	S_3	S_4
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

permutation

2, 6, 4, 3, 1, 5



Map Set to vector $\mathbb{R}^k$

k = # iterations
↳ # permutations

$S_j \xrightarrow{\text{permute}}$ $\begin{bmatrix} m_1(S_j) \\ m_2(S_j) \\ m_3(S_j) \\ \vdots \\ m_{12}(S_j) \end{bmatrix} \in \mathbb{R}^k$

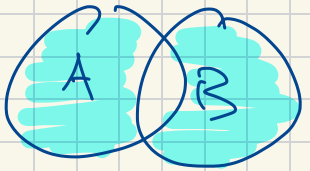
v_j

$m_i(S_j)$ single value representation of S_j using ith permutation.
Value is first row w/ 1

$$\widehat{JS}_k(S_1, S_2) = \frac{1}{k} \sum_{i=1}^{k} \begin{cases} 1 & \text{if } (v_1[i] = v_2[i] \\ 0 & \text{o.w.} \end{cases}$$

$$E_{perm}\left[\hat{JS}_1(S_1, S_2)\right] = JS(S_1, S_2)$$

$$Pr_{perm}\left[m_1(S_1) = m_1(S_2)\right] = JS(S_1, S_2)$$



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

$A \cap B \Rightarrow$ # element in both

$A \Delta B \Rightarrow$ # element in 1 or other not both

which element [m] can be first non-zero in either A or B. $\equiv A \cup B$

which element [m] can give collision $\Rightarrow m_1(A) = m_1(B)$ $\equiv A \cap B$

Chernoff Hoeffding

k R.V.s $X_1 \dots X_k$ i.i.d $X_i \in \{0,1\}$

know $E[X_i] = u$

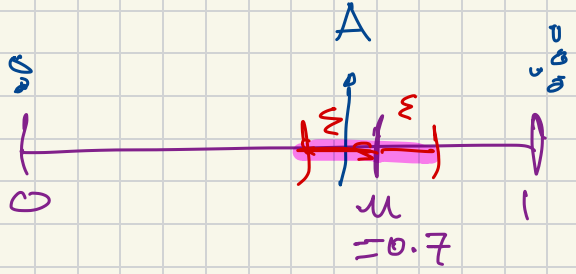
$$A = \frac{1}{k} \sum_{j=1}^{k} X_j \qquad E[A] = u$$

$$Pr[|A-u| > \varepsilon] \leq 2 \exp\left(-\frac{\varepsilon k^2}{2}\right) = \delta$$

δ ε

Probably Approx Correct

PAC



Solve for $k = \frac{2}{\varepsilon^2} \log \frac{2}{\delta}$

$\varepsilon = 0.05$

$\delta = 0.01$

$\Rightarrow k = 800 \cdot 8$

$\Rightarrow 6,400$

Define hash function

$h_i : [|\Sigma|] \to [N]$ (random) (w/ salt $a_1 \dots a_k$, so maps uniformly to $[N]$)

$|\Sigma|$: # things in all sets

$N = 9|\Sigma|^2$

only need comparison operator

---

once chosen $h_i \sim \mathcal{H}$, then deterministic.

Convert Set S to vector $v \in \mathbb{R}^k$

$h_1, h_2, \dots h_k \sim \mathcal{H}$

```
for (x_i ∈ S)
    for (j = 1 to k)
        if (h_j(x_i) < v_j)
            v_j ← h_j(x_i)
```

init $v = (N+1, N+1 \dots N+1) \in \mathbb{R}^k$

h_1 h_2 (v_j) h_k

Fast Min Hash

Next Scenario

$Z =$ { 1 billion documents
1 billion webpages
7 million Wikipedia articles

query document

find element $z \in Z$ so $JS(q, z)$ largest.

quickly w/ scanning all documents

↳ w/ Locality sensitive Hashing

LSH