Date . .

SDSC 2005 Assignment 1

SID: 56641800 Name: Du Junye.

1. $f(S_1, S_2) = 1 - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$

(i) as S_1, S_2 are two sets,

$\therefore S_1 \cap S_2 = S_2 \cap S_1 \quad S_1 \cup S_2 = S_2 \cup S_1$

$\therefore \left| \frac{S_1 \cap S_2}{S_1 \cup S_2} \right| = \left| \frac{S_2 \cap S_1}{S_2 \cup S_1} \right|$

$\therefore f(S_1, S_2) = f(S_2, S_1)$

$\because S_1 \cup S_2 = S_1^c \cap S_2 + S_2^c \cap S_1 + S_1 \cap S_2$.

$\because S_1^c \cap S_2$, $S_2^c \cap S_1$, $S_1 \cap S_2$ are disjoint

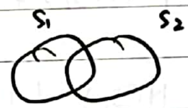
$\therefore |S_1 \cup S_2| = |S_1^c \cap S_2| + |S_2^c \cap S_1| + |S_1 \cap S_2|$

$\because |S_1^c \cap S_2| \geq 0$. $|S_2^c \cap S_1| \geq 0$

$\therefore \left| \frac{S_1 \cap S_2}{S_1 \cup S_2} \right| \leq 1$

$\therefore f(S_1, S_2) \geq 0$

$\Rightarrow f(S_1, S_2) = f(S_2, S_1) \geq 0$



(ii) First prove sufficiency:

if $f(S_1, S_2) = 0$

$f(S_1, S_2) = 1 - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} = 0 \quad \therefore \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} = 1$

$\therefore |S_1 \cap S_2| = |S_1 \cup S_2| = |S_1^c \cap S_2| + |S_2^c \cap S_1| + |S_1 \cap S_2|$ (from Question i)

$\therefore |S_1^c \cap S_2| + |S_2^c \cap S_1| = 0 \quad \because |S_1^c \cap S_2| \geq 0 \quad |S_2^c \cap S_1| \geq 0$

$\therefore |S_1^c \cap S_2| = 0 \quad |S_2^c \cap S_1| = 0$

$\therefore S_2 \subseteq S_1$, $S_1 \subseteq S_2$

$\therefore S_1 = S_2$

Prove necessity

if $S_1 = S_2$. $S_1^c \cap S_2 = S_2^c \cap S_1 = 0 \quad \therefore S_2 \cup S_1 = S_1 \cap S_2$

$\therefore f(S_1, S_2) = 1 - \frac{|S_1 \cap S_2|}{|S_1 \cap S_2|} = 1 - 1 = 0$

$\Downarrow$

$f(S_1, S_2) = 0$ if and only if $S_1 = S_2$

(iii) $f(S_1, S_3) \le f(S_1, S_2) + f(S_2, S_3)$ for any S_1, S_2, S_3.

① If there exists empty set, suppose $S_1 = \emptyset$

$\therefore f(S_1, S_2) = 1 - 0 = 1 \qquad f(S_1, S_3) = 1$

$\because f(S_2, S_3) \ge 0 \qquad \therefore$ inequality hosts.

② If there doesn't exit empty set.

$S_1 S_2 S_3 \ne \emptyset$

$\because$ the Jaccard distance derives from the Min Hashing.

we just let $H(X) = \operatorname{argmin}_{i \in X} \pi(i)$ ($\pi(i)$ is a random permutation)

$\because Js(X, Y) = P(H(X) = H(Y))$

$\therefore f(S_1, S_2) = 1 - Js(S_1, S_2) = P(H(S_1) \ne H(S_2))$

$f(S_2, S_3) = P(H(S_2) \ne H(S_3)) \qquad f(S_1, S_3) = P(H(S_1) \ne H(S_3))$

$\because P(H(S_1) = H(S_3)) \ge P[H(S_1) = H(S_2) \cap H(S_3) = H(S_2)]$

$\Downarrow$ De Morgan's Law

$\therefore P(H(S_1) \ne H(S_3)) \le P[H(S_1) \ne H(S_2) \cup H(S_3) \ne H(S_2)]$

$\Downarrow$

$\therefore P[H(S_1) \ne H(S_3)] \le P[H(S_1) \ne H(S_2)] + P[H(S_3) \ne H(S_2)]$.

$\Downarrow$

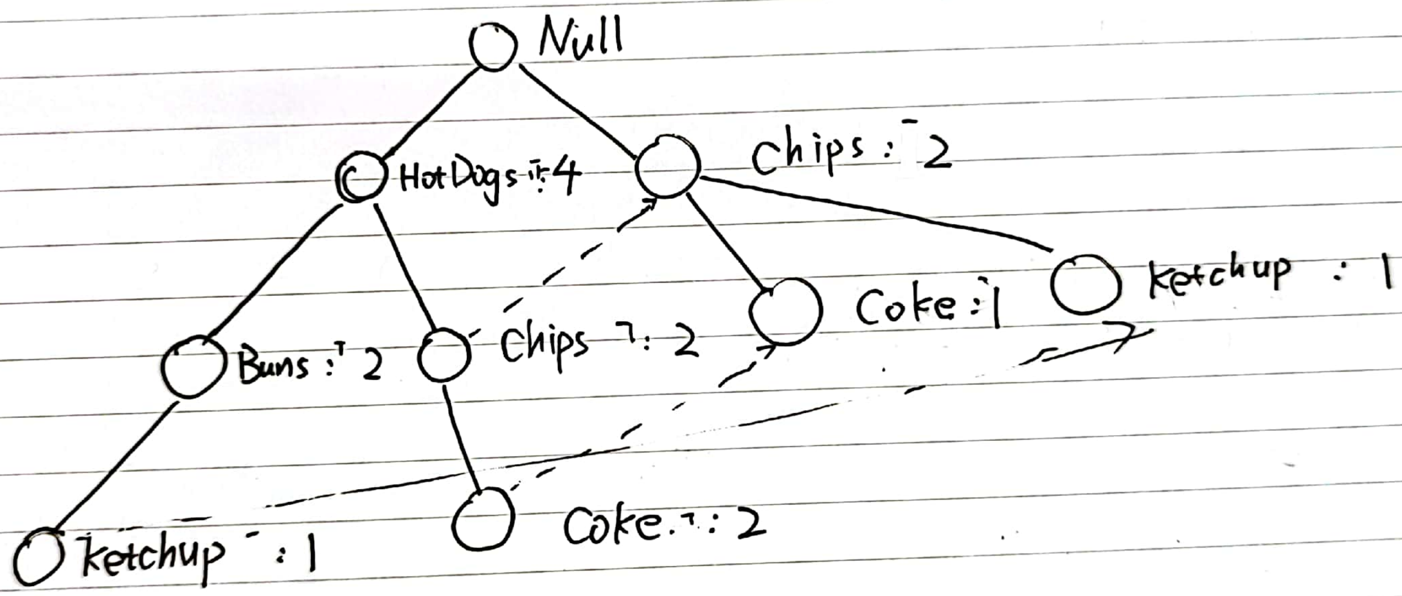
$\therefore f(S_1, S_3) \le f(S_1, S_2) + f(S_3, S_2)$

$= f(S_1, S_2) + f(S_2, S_3)$.

$\therefore$ In conclusion: $f(S_1, S_3) \le f(S_1, S_2) + f(S_2, S_3)$

for any S_1, S_2, S_3.

2. Scan the DB once to find frequent single items
Hot dogs : 4   Buns : 2   Ketchup : 2
Coke : 3   Chips : 4
In Descending order:
Hotdogs : 4 → Chips : 4 → Coke : 3 → Buns : 2 → Ketchup : 2
Resort the DB:
1. Hot Dogs. Buns. Ketchup
2. Hot Dogs. Buns
3. Hot Dogs. Chips. Coke.
4. Chips. Coke
5. Chips. Ketchup
6. Hotdogs. chips. coke.

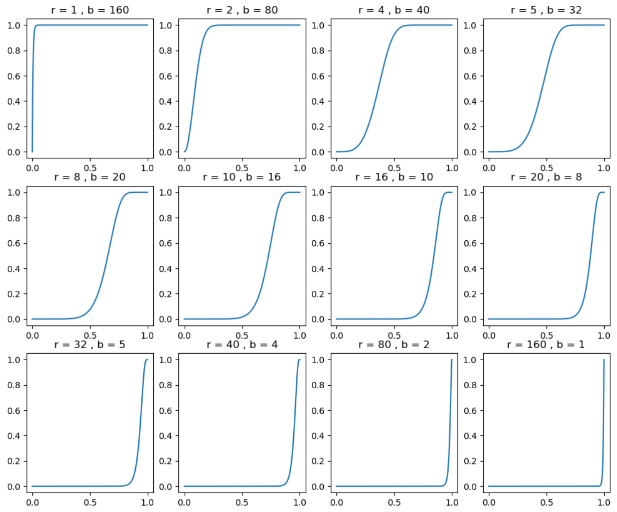


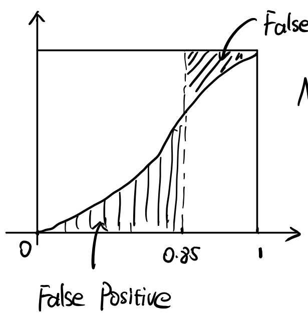
3. Using the (r, b)-way And-Or Construction:

$\Rightarrow r \cdot b = 160.$

$\because f(x) = 1 - (1 - x^r)^b, \quad x \in [0, 1]$

$f'(x) = br\, x^{r-1}(1 - x^r)^{b-1} = 160\, x^{r-1}(1 - x^r)^{b-1}$





$M(x) =$ Area of False Positive + Area of False Negative

$$= \int_0^{0.85} 1 - (1 - x^r)^b \, dx + \int_{0.85}^{1} (1 - x^r)^b \, dx$$

```
The Area of parameter r=1 b=160 is 0.84379
The Area of parameter r=2 b=80 is 0.75138
The Area of parameter r=4 b=40 is 0.49098
The Area of parameter r=5 b=32 is 0.39262
The Area of parameter r=8 b=20 is 0.20470
The Area of parameter r=10 b=16 is 0.13249
The Area of parameter r=16 b=10 is 0.04818
The Area of parameter r=20 b=8 is 0.04801
The Area of parameter r=32 b=5 is 0.08320
The Area of parameter r=40 b=4 is 0.09992
The Area of parameter r=80 b=2 is 0.13152
The Area of parameter r=160 b=1 is 0.14379
```

From the table.

we could find that $r=20$. $b=8$ has the minimum area.

so $r=20$. $b=8$ is the best.

4.

(1)

0.0001: total: 18694

{733, 8133, 7998, 1740, 90}

0.0002: total 8219

{592. 4553. 2734. 334. 6}

0.0003: total 5077

{524. 3081. 1342. 129, 1}

0.0004: total 3556

{470. 2237. 800. 49}

0.0005: total 2695

{437. 1727 505. 26}

(2). (3)

the acceleration tequiques and comparison are included in the source code. which is an ipynb file.