Au-Yeung Fung Yin

57269800

Q1

1. (20 points) Let $JS(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$ be the Jaccard similarity between two sets S_1 and S_2. Prove that $f(S_1, S_2) = 1 - JS(S_1, S_2)$ is a distance measure, that is, $f(\cdot)$ satisfies the following properties

(i) $f(S_1, S_2) = f(S_2, S_1) \geq 0$ (5 points)

(ii) $f(S_1, S_2) = 0$ if and only if $S_1 = S_2$ (5 points)

(iii) $f(S_1, S_3) \leq f(S_1, S_2) + f(S_2, S_3)$, *for any* S_1, S_2, S_3. (10 points)

(i) $f(S_1, S_2) = 1 - JS(S_1, S_2)$

$$= 1 - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

$$= 1 - \frac{|S_2 \cap S_1|}{|S_2 \cup S_1|}$$

$$= 1 - JS(S_2, S_1) = f(S_2, S_1)$$

Since $|S_2 \cap S_1| \leq |S_1 \cap S_2|$, $JS(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \leq 1$

$f(S_2, S_1) = 1 - JS(S_1, S_2) \geq 0$

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

(ii) When $S_1 = S_2$, $|S_1 \cap S_2| = |S_1 \cup S_2| = 1$

$JS(S_1, S_2) = 1$

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

Since $f(S_1, S_2) = 0$

$JS(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} = 1$

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

(iii)

$f(S_1,S_3) = 1 - JS(S_1,S_3)$

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

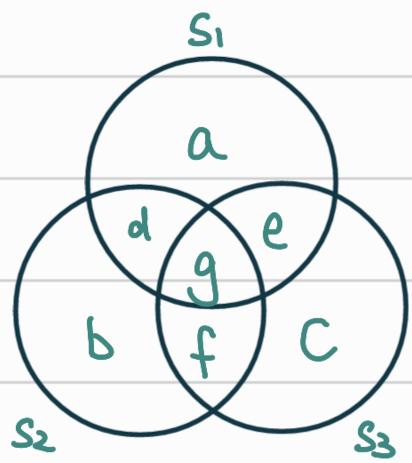
$= (1 - JS(S_1,S_2)) + (1 - JS(S_2,S_3))$

$= 2 - JS(S_1,S_2) - JS(S_2,S_3)$

$f(S_1,S_3) \leq f(S_1,S_2) + f(S_2,S_3)$

$1 - JS(S_1,S_3) \leq 2 - JS(S_1,S_2) - JS(S_2,S_3)$

$JS(S_1,S_3) \geq JS(S_1,S_2) + JS(S_2,S_3) - 1$



Assume $JS(S_1,S_3) \leq JS(S_1,S_2) + JS(S_2,S_3) - 1$

$$\frac{|S_1 \cap S_3|}{|S_1 \cup S_3|} \leq \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} + \frac{|S_2 \cap S_3|}{|S_2 \cup S_3|} - 1$$

$$\frac{e+g}{a+c+d+e+f+g} \leq \frac{d+g}{a+b+d+e+f+g} + \frac{g+f}{b+c+d+e+f+g} - 1$$

$$\frac{e+g}{a+b+c+d+e+f+g} \leq \frac{a+c+d+2g+f}{a+b+c+d+e+f+g} - 1$$

$e+g \leq g - b - e$

$e \leq -b - e \Rightarrow$ contradiction

$$\frac{e+g}{a+b+c+d+e+f+g} \leq \frac{e+g}{a+c+d+e+f+g}$$

$$\frac{d+g}{a+b+d+e+f+g} \leq \frac{c+d+g}{a+b+c+d+e+f+g}$$

$$\frac{g+f}{b+c+d+e+f+g} \leq \frac{a+g+f}{a+b+c+d+e+f+g}$$

$\because \frac{x}{y} \leq \frac{x+z}{y+z}$

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

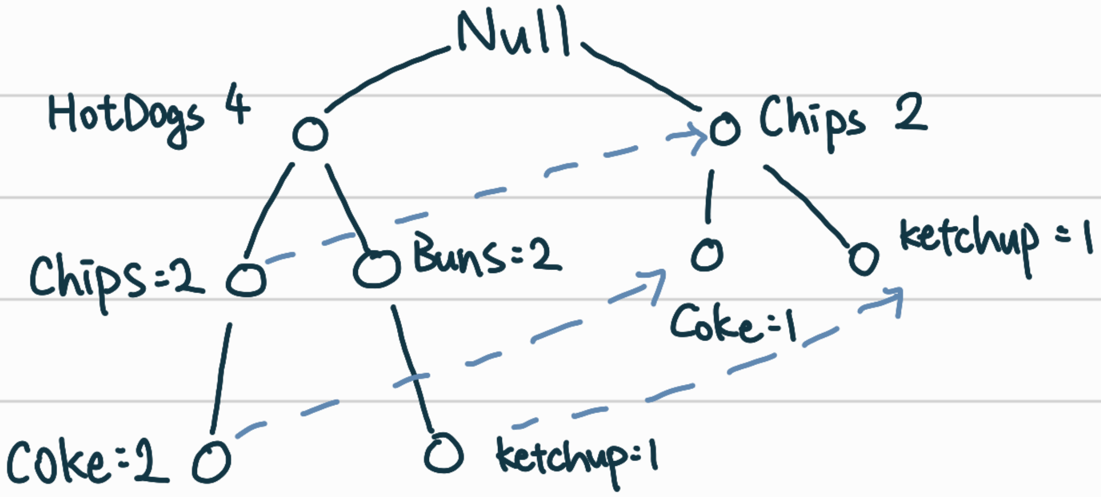
Q2

2. (15 points) Build an FP-tree for the following transaction database. Sort items in support descending order. Draw the FP-tree.

Transaction ID	Items
1	HotDogs, Buns, Ketchup
2	HotDogs, Buns
3	HotDogs, Coke, Chips
4	Chips, Coke
5	Chips, Ketchup
6	HotDogs, Coke, Chips

Orders : HotDogs : 4 , Chips : 4 , Coke : 3 , Buns : 2 , Ketchup : 2

FP-tree :



3. (15 points) Consider computing an LSH using $k = 160$ hash functions. We want to find all object pairs which have Jaccard similarity at least $t = 0.85$. Suppose we use the (r, b)-way AND-OR construction, which means that a pair of documents with similarity s is considered as a candidate pair with probability $1 - (1 - s^r)^b$. Choose the best r and b. Justify why your choice is the best.

$k = b \times r = 160$ possible combination: (1,160) (2,80) (4,40)

(5,32) (8,20) (10,16)

$f(s) = 1 - (1 - s^r)^b$

False positive rate: minimize $\int_0^{0.85} 1 - (1 - s^r)^b \, ds$

False negative rate = minimize $\int_{0.85}^{1} 1 - (1 - s^r)^b s$

$\int 1 - (1 - s^r)^b \, ds = \int 1 \, ds - \int (1 - s^r)^b \, ds$

$= s + r \frac{(1 - s^r)^{b+1}}{b+1} + C$

Let $u = 1 - s^r$

$du = -r \, ds$

$-r \int u^b \, du$

$= -r \left(\frac{u^{b+1}}{b+1} \right) du$

$= -r \frac{(1 - s^r)^{b+1}}{b+1}$

After testing the possible value of r and b,

I found $r = 10$, $b = 16$ could minimize the sum of

False positive and False negative rate