Homework 3 Solution

Shixin Tian
Large Data ComS 535/435

October 27, 2017

1 Task 1

1.1 (a) Binary Term-Frequency Vector

As there are 9 terms in the corpus, we need vectors of length 9.

$T_1 = [1, 0, 0, 1, 0, 1, 1, 1, 0]$.

$T_2 = [0, 1, 1, 1, 0, 0, 1, 0, 1]$.

1.2 (b) Jaccard Similary

$Jac(D_1, D_2) = \frac{T_1 \cdot T_2}{L_2(T_1)^2 + L_2(T_2)^2 - T_1 \cdot T_2} = \frac{2}{5+5-2} = \frac{1}{4}$.

1.3 (c) Cosine Similary

$Cos(D_1, D_2) = \frac{T_1 \cdot T_2}{L_2(T_1) * L_2(T_2)} = \frac{2}{5}$.

2 Task 2

Let T_1 and T_2 be the two binary term-frequency vectors of D_1 and D_2. Then $C = \frac{T_1 \cdot T_2}{L_2(T_1) * L_2(T_2)}$ and $J = \frac{T_1 \cdot T_2}{L_2(T_1)^2 + L_2(T_2)^2 - T_1 \cdot T_2}$

Let x denote $L_2(T_1)^2$, y denote $L_2(T_2)^2$, z denote $T_1 * T_2$.

Since T_1 and T_2 are binary term-frequency vectors, we have $a_i * b_i \leq {a_i}^2$ and $a_i * b_i \leq {b_i}^2$ ($a_i, b_i \in \{0, 1\}$). Thus, we have $z = \sum_{i=1}^{m} a_i * b_i \leq \sum_{i=1}^{m} {a_i}^2 = x$ and $z \leq y$.

2.1 (a) $C^2 \leq J$

$$
\begin{aligned}
C^2 - J &= \left(\frac{T_1 \cdot T_2}{L_2(T_1) * L_2(T_2)}\right)^2 - \frac{T_1 \cdot T_2}{L_2(T_1)^2 + L_2(T_2)^2 - T_1 \cdot T_2} \\
&= \frac{z^2}{x * y} - \frac{z}{x + y - z} \\
&= \frac{z(x+y-z) - xy}{x * y * (x+y-z)/z} \\
&= -\frac{(x-z)(y-z)}{x * y * (x+y-z)/z}
\end{aligned}
$$

Since $x + y - z \geq 0$, $x - z \geq 0$ and $y - z \geq 0$, we have $C^2 - J \leq 0$. Proved.

2.2 (b) $J \leq \frac{C}{2-C}$

$$\begin{aligned} J - \frac{C}{2-C} &= \frac{z}{x+y-z} - \frac{\frac{z}{\sqrt{x*y}}}{2-\frac{z}{\sqrt{x*y}}} \\ &= \frac{z}{x+y-z} - \frac{z}{2\sqrt{x*y}-z} \\ &= \frac{2\sqrt{x*y}-x-y}{(x+y-z)*(2\sqrt{x*y}-z)/z} \\ &= -\frac{(\sqrt{x}-\sqrt{y})^2}{(x+y-z)*(2\sqrt{x*y}-z)/z} \\ &\leq 0 \end{aligned}$$

Proved.

3 Task 3

	D_1	D_2	D_3	D_4
$\prod_1 : (2x+1) \mod 5$	0	1	2	0
$\prod_2 : (3x+4) \mod 5$	0	2	1	0
$\prod_3 : (x+3) \mod 5$	0	3	1	0

4 Task 4

Since h is a one-one function from $\{1, 2...n\}$ to $\{1, 2...n, n+1\}$, there must be a number which is not mapped by any number in $\{1, 2...n\}$. Hence there are two cases 1) $\exists h^{-1}(1)$ i.e.,($\exists k \in \{2, 3...n, n+1\} s.t., \forall i \in \{1, 2...n\} h(i) \neq k$) and 2) $\nexists h^{-1}(1)$. In the first case,$min(h(D_1))$ or $min(h(D_2))$ equals to 1; in the second case, $min(h(D_1))$ and $min(h(D_2))$ don't equal to 1 and at least one of them would equal to 2.

As we mentioned before, there would always be number which is not mapped by any number. Suppose this number is y. Now we calculate $Pr(h(i) = j, \nexists h^{-1}(y))$.

$$\begin{aligned} Pr[h(i) = j, \nexists h^{-1}(y)] &= Pr[h(1) \neq j \bigcap h(1) \neq y...h(i) = j, h(i+1) \neq y...h(n) \neq y] \\ &= \frac{n-1}{n+1} * \frac{n-2}{n} * ... * \frac{n+1-i}{n+3-i} * \frac{1}{n+2-i} * \frac{n-i}{n+1-i} ... * \frac{1}{2} \\ &= \frac{1}{(n+1)n} \end{aligned}$$

$$\begin{aligned} Pr[min(h(D_1)) = min(h(D_2)) = 1] &= Pr[\exists x \in D_1 \bigcup D_2, h(x) = 1, \exists h^{-1}(1)] \\ &= \sum_{k=2}^{n+1} Pr[\exists x \in D_1 \bigcup D_2, h(x) = 1, \nexists h^{-1}(k))] \\ &= \sum_{k=2}^{n+1} |D_1 \bigcap D_2| Pr[h(x) = 1, \nexists h^{-1}(k)] \\ &= \sum_{k=2}^{n+1} \frac{|D_1 \bigcap D_2|}{(n+1)n} \\ &= \frac{|D_1 \bigcap D_2|}{n+1} \end{aligned}$$

$$\begin{aligned} Pr[min(h(D_1)) = min(h(D_2)) = 2] &= Pr[\exists x \in D_1 \bigcup D_2, h(x) = 2, \nexists h^{-1}(1)] \\ &= Pr[\exists x \in D_1 \bigcup D_2, h(x) = 2, \nexists h^{-1}(1)] \\ &= |D_1 \bigcap D_2| Pr(h(x) = 1), \nexists h^{-1}(k))) \\ &= \frac{|D_1 \bigcap D_2|}{(n+1)n} \end{aligned}$$

Thus, we have $Pr[min(h(D_1)) = min(h(D_2)] = Pr[min(h(D_1)) = min(h(D_2)) = 1] + Pr[min(h(D_1)) = min(h(D_2)) = 2] = \frac{|D_1 \bigcap D_2|}{n}$

5 Task 5

A simple solution would be 'randomly' pick up a number from the vector,i.e., $h(U) = u_i$ where $i \in \{1, 2...M\}$ is randomly picked.

$$\begin{aligned} Pr_{h \in H}(h(U) = h(V)) &= Pr(u_i = v_i) \\ &= \frac{|i|u_i = v_i|}{M} (since\ i\ is\ randomly\ picked) \end{aligned}$$

If we want to evaluate this similarity, we can calculate the hash values of two vectors and then the similarity would be approximated by the number of equal pairs divided by M.

Recall that Min-hash is a Jaccard-similarity preserving hash family for documents. However, Min-hash only works for documents (sets of terms) but not for vectors. So another way to measure this similarity is to map a vector to a document (such that the orders of elements in the vector do not matter). One naive way is to map a vector $U =< u1, u2, ..um >$ into a set $D = \{(u1, 1), (u2, 2), ...(um, m)\}$. Note that Jaccard similarity is not the similarity we are interested. You need to do some transformation to get it.