ComS 535x:
Homework #2

Due on Mar 3, 2015

Instructor: Professor Pavankumar Aduri

Chenguang He

Contents

Question 1

Consider the following documents D1 = 1, 4, 6, 7, 8 and D2 = 2, 3, 9, 4, 7

(a) What are the binary term-frequency vectors of D1 and D2?

Answer:

$T = (1, 4, 6, 7, 8, 2, 3, 9)$

$Tf_{11} =< 1, 1, 1, 1, 1, 0, 0, 0 >, Tf_{12} =< 0, 1, 0, 1, 0, 1, 1, 1 >$

(b) What is the Jaccard Similary of D1 and D2 (with respect to binary term-frequency vectors)

Answer:

$$Jac(Tf_{11}, Tf_{12}) = \frac{|Tf_{11} \cap Tf_{12}|}{|Tf_{11} \cup Tf_{12}|} = \frac{Tf_{11} \cdot Tf_{12}}{Tf_{11}^2 + Tf_{12}^2 - Tf_{11} \cdot Tf_{12}} = 0.25$$

(c) What is the cosine similarity of D1 and D2 (with respect to binary term-frequency vectors)

Answer:

$$Cos(Tf_{11}, Tf_{12}) = \frac{Tf_{11} \cdot Tf_{12}}{||Tf_{11}||||Tf_{12}||} = 0.4$$

Question 2

Let D1 and D2 be two documents. Let C be the cosine similarity of the documents with respect to binary term-frequency vectors and J be the jacquard similarity with respect to binary term-frequency vectors. Show that

(a) $C^2 \leq J$
Answer:

Let $\dfrac{C^2}{J} = \dfrac{|D_1 \cap D_2|^2}{|D_1||D_2|} \times \dfrac{|D_1 \cup D_2|}{|D_1 \cap D_2|} = \dfrac{|D_1 \cap D_2| \cdot |D_1 \cup D_2|}{|D_1||D_2|}$, when D1 = D2, $C^2 = J$, otherwise, $|D_1 \cap D_2|$ will go smaller and the inequality will less than 1.

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

Answer:

Let D1 = X and D2 = Y, and i from 1 to n.

Since, J = $\dfrac{\sum X_i Y_i}{\sum X_i^2 + \sum Y_i^2 - \sum X_i Y_i}$ and C = $\dfrac{\sum X_i Y_i}{\sqrt{\sum X_i^2} + \sqrt{\sum Y_i^2}}$
We have $\dfrac{C}{J} = \dfrac{\sum X_i^2 + \sum Y_i^2 - \sum X_i Y_i}{\sqrt{\sum X_i^2} + \sqrt{\sum Y_i^2}} = \sqrt{\dfrac{\sum X_i^2}{\sum Y_i^2}} + \sqrt{\dfrac{\sum Y_i^2}{\sum X_i^2}} - C$
Thus, $J = \dfrac{C}{\sqrt{\dfrac{\sum X_i^2}{\sum Y_i^2}} + \sqrt{\dfrac{\sum Y_i^2}{\sum X_i^2}} - C}$
Because of X and Y are two vectors, we can write the relationship between X and Y as: $|X| = c|Y|$ where $c > 0$. then when $c = 1$, $|X| = |Y|$, we have $J = \dfrac{C}{2 - C}$, we have $J \leq \dfrac{C}{2 - C}$, since $\sqrt{\dfrac{\sum X_i^2}{\sum Y_i^2}} + \sqrt{\dfrac{\sum Y_i^2}{\sum X_i^2}}$ has minimum value, when X = Y. Otherwise, it $J < \dfrac{C}{2 - C}$ when $X \neq Y$

Question 3

3. Let D1 and D2 be two documents such that D1U D2 = $1, , n$**. Show that the Jaccard similarity of D1 and D2 can be computed exactly in time O(n log n**).

Answer:

There are two parts for Jaccard similarity, the union of two sets and the intersection of two sets. When we calculate the union of two set, we can simply count all element in two sets, it takes O(n). When we calculate the intersection of two sets, we simply sort two sets, it takes O(nlgn) times and for each element in set D_1, we simply do binary search in D_2, since D_2 is sorted. It takes O(lgn) to find an element in D_2, and there are n element in D_1 . Therefor entire program takes O(nlgn) time.

Question 4

Suppose we picked the following permutations (2x+1)the MinHash matrix.

$(2x+1)\%5 = 0- > 1, 1- > 3, 2- > 0, 3- > 2, 4- > 4$ so $D1 = 0, D2 = 1, D3 = 2, D4 = 0$

$(3x+4)\%5 = 0- > 4, 1- > 2, 2- > 0, 3- > 3, 4- > 1$ so $D1 = 0, D2 = 2, D3 = 1, D4 = 0$

$(x+3)\%5 = 0- > 3, 1- > 4, 2- > 0, 3- > 1, 4- > 2$ so $D1 = 0, D2 = 3, D3 = 1, D4 = 0$

0	1	2	0
0	2	1	0
0	3	1	0

Question 5

Suppose that we toss a biased coin (probability of head 1/4) n times. Give a lower bound the probability that we see at least log n consecutive heads. In locality sensitive hashing, we showed that if two documents are s-similar, then the probability that they are mapped to the same bucket in some hash table is at least $1 - (1 - s^r)^b$. Do you see similarity between the two proofs?

Yes. Assume that, we toss the coin n times. For each time, we map it into $[\frac{n}{logn}]$ blocks. Because the coin is biased, the probability of consecutive $logn$ tosses is $(\frac{1}{4})^{logn}$, and the proablitly of toss in each block of tail is at most $1 - (\frac{1}{4})^{logn}$.

Therefore, in lgn toss, the probability of each block get all head is $1 - (1 - (\frac{1}{4})^{logn})^{\frac{n}{logn}}$, where is similarity with $1 - (1 - s^r)^b$ for s = $\frac{1}{4}$, r= $logn$, b=$\frac{n}{logn}$

Question 6

Prove Claim 3 from Notes II

Define random variable X_i to be the number of ith permutation where MH_a and MH_b matches. So we have, X_i = 1 if $min[\prod_i(D_a)]$ = $min[\prod_i(D_b)]$, X_i = 0. otherwise.

Since, for each X_i, it is independent event. We use chernoff's bound there:

$$Pr[|\frac{x}{k} - Jac(D_a, D_b)| \geq Jac(D_a, D_b) \cdot \delta] \leq 2 \cdot e^{-\dfrac{\delta^2 \cdot k \cdot Jac(D_a, D_b)}{2}}$$

$$\Longrightarrow Pr[|\frac{x}{k} - Jac(D_a, D_b)| \leq Jac(D_a, D_b) \cdot \delta] \geq 1 - 2 \cdot e^{-\dfrac{\delta^2 \cdot k \cdot Jac(D_a, D_b)}{2}}$$

The output of algorithm A is $\frac{l}{k}$. Let $k = c \cdot \dfrac{1}{\varepsilon^2 \cdot log\frac{1}{\delta}}$ and $\delta = \dfrac{\varepsilon}{Jac(D_a, D_b)}$, then we have:

$$Pr[|Output of A - Jac(D_a, D_b)| \leq \varepsilon] = 1 - 2e^{-(\dfrac{\varepsilon}{Jac(D_a, D_b)})^2 \cdot \dfrac{c \cdot \frac{1}{\varepsilon} \cdot log\frac{1}{\varepsilon} \cdot Jac(D_a, D_b}{2}} = 1\text{-}2 \cdot \delta^{(\dfrac{c}{2 \cdot Jac(D_a, D_b})}$$

Because when c increase to large enough, then $2 \cdot \delta^{(\dfrac{c}{2 \cdot Jac(D_a, D_b})} \leq \delta$

We add $(1-)$ at both side, then we have: $1 - 2 \cdot \delta^{(\dfrac{c}{2 \cdot Jac(D_a, D_b})} \geq 1 - \delta$

Therefor $Pr[|Output of A - Jac(D_a, D_b|) \leq \varepsilon] \geq 1 - \delta$

Question 7

Because the hash function is one to one mapping, by the Claim 1 in Note, we have $Let S = 1, ..., m, T = 1, ...m + 1 and Let h be a hash fu$
we want to prove that:

$$Pr[h(i) = j] = \frac{1}{m+1}$$

Proof:

$$Pr[h(i) = j] = \frac{m}{m+1} \times \frac{m-1}{m} \times \frac{m-2}{m-1} .... \times \frac{m-i+1}{m-i+2} \times \frac{1}{m-i+1} = \frac{1}{m+1}$$

Now, by Claim 2, we want to prove that:

$$Pr[min[h(D_a)] = min[h(D_b)]] = Jac(D_a, D_b)$$

Proof:

$$Pr[min[h(D_a)] = min[h(D_b)]] = \frac{D_a \cap D_b}{m+1} = \frac{D_a \cap D_b}{D_a \cup D_b} = Jac(D_a, D_b)$$