Homework Assignment 2

Problem 1

$$E(sim(S,T)) = \begin{cases} 1 & \text{if} \quad 0 < n \le m; \\ \sum_{k=m}^{n} \frac{2m-k}{k} \frac{\binom{m}{2m-k}\binom{n-m}{k-m}}{\binom{n}{m}} & \text{if} \quad m < n \le 2m-2; \\ \sum_{k=m}^{2m-1} \frac{2m-k}{k} \frac{\binom{m}{2m-k}\binom{n-m}{k-m}}{\binom{n}{m}} & \text{if} \quad n > 2m-2; \end{cases}$$

Problem 2

Question 1

{

'Even hash them ',

'hash them four bytes each, ',

'four bytes each, space needed ',

'space needed store ',

'store still roughly four times ',

'still roughly four times space taken ',

'space taken document ',

'document '

}

Question 2

n-k+1

Problem 3

Map phase:

Map(index_of_band, index_of_doc, shingles_of_every_doc) -> <index_of_band, index_of_doc,

hash_per_band>

Inside ever Mapper hash table for every documents is caculated within

same band. Meanwhile indexes of documents are only be preserved.

Reduce phase:

Reduce(index_of_doc, hash_table_per_band) -> <index_of_doc, hash_table_per_doc>

For every Reducer hash tables per band are grouped and ordered according to document id

---

Problem 4

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

$$=1-(1-s^{\frac{r}{2}})^b(1+s^{\frac{r}{2}})^b$$

$$=1-\lim_{n\to\infty}\prod_{i=1}^{n}e^{\frac{s^r b}{2i}}$$

$$=1-e^{s^r b \lim_{n\to\infty}\sum_{i=1}^{n}\frac{1}{2i}}$$

$$=1-e^{s^r b}$$

Problem 5

1. precise threshold : 0.569

   estimate threshold : 0.607

   relative difference(|pt-et|/pt) : 0.066

2. precise threshold : 0.406

   estimate threshold : 0.464

   relative difference : 0.143

3. precise threshold : 0.880

   estimate threshold : 0.890

   relative differnce : 0.011

The esitimate threshold approches to precise threshold when the value of formular is exactly 1/2, especially when the value of b as well as r are sufficiently great.

Problem 6

1. max(x, y) is not a distance measure. When x!=0, max(x, x)=x!=0 Identity not holds.

2. diff(x, y) is a distance measure. Because

    1. Non-negativity: diff(x, y) = |x - y| >= 0, if and only if x = y the equality holds.
    2. symmetricity: diff(x, y) = |x - y| = |y - x| = diff(y, x)
        1. Identity: diff(x, y) = |x - y| = 0, iff x = y
        2. Trangle inequality: diff(x, y) + diff(y, z) = |x - y| + |y - z| case 1: x = z |x - y| + |y - z| =|x - y| + |y - x|>=0=|x - x| = |x - y| = diff(x, y) case 2: x != z, let x < z unequality={x-y+z-y=z-x+2x-2y>z-x=diff(x, z), if 0<y<x<z {y-x+z-y=z-x=diff(x,z), if x<=y<z {y-z+y-x=z-x-2z+2y>z-x=diff(x, z), if x<z<=y so, it's a metric
3. sum(x, y) = x + y is not a metric. if x!=0 sum(x, x)=2x!=0 Identity not holds.

4. Jaccard distance is a matric. Given three sets A, B and C, Jaccard distance of two sets (e.g. A and B) is defined as J(A, B) = 1-|AnB|/|AUB|

    1. Non-negativity: 1-|AnB|/|AUB|>=0, iff A=B the equality holds.
    2. Symmetricity: J(A, B)=1-|AnB|/|AUB|=1-|BnA|/|AUB|=J(B, A)
    3. Identity: J(A, A)=1-|AnA|/|AUA|=1-|A|/|A|=0
    4. Trangle inequality: see solution below:

$$a + A \setminus B \setminus C, b + B \setminus C \setminus A, c + C \setminus A \setminus B$$

$$\alpha + A \cap C \setminus B, \beta + a \cap B \cap C, \gamma + A \cap B \setminus C, \delta + B \cap C \setminus A$$

$$J(A, B) + \frac{a + b + \alpha + \delta}{\alpha + \beta + \gamma + \delta + a + b}$$

$$J(B, C) + \frac{b + c + \alpha + \gamma}{\alpha + \beta + \gamma + \delta + b + c}$$

$$J(A, C) + \frac{a + c + \gamma + \delta}{\alpha + \beta + \gamma + \delta + a + c}$$

$$J(A, B) + J(B, C) \geq J(A, C)$$

$$\Leftrightarrow \frac{a + b + \alpha + \delta}{\alpha + \beta + \gamma + \delta + a + b} + \frac{b + c + \alpha + \gamma}{\alpha + \beta + \gamma + \delta + b + c} \geq \frac{a + c + \gamma + \delta}{\alpha + \beta + \gamma + \delta + a + c}$$

$$\Leftrightarrow \frac{a+\delta}{\alpha+\beta+\gamma+\delta+a+b}+\frac{c+\gamma}{\alpha+\beta+\gamma+\delta+b+c}+\frac{b+\alpha}{\alpha+\beta+\gamma+\delta+a+b}+$$
$$\frac{b+\alpha}{\alpha+\beta+\gamma+\delta+b+c} \geq \frac{a+c+\gamma+\delta}{\alpha+\beta+\gamma+\delta+a+c}$$
$$\Leftarrow \frac{a+c+\gamma+\delta}{\alpha+\beta+\gamma+\delta+a+b+c}+\frac{b+\alpha}{\alpha+\beta+\gamma+\delta+a+b}+\frac{b+\alpha}{\alpha+\beta+\gamma+\delta+b+c} \geq$$
$$\frac{a+b+c+\gamma+\delta}{\alpha+\beta+\gamma+\delta+a+b+c}$$
$$\Leftrightarrow \frac{b+\alpha}{\alpha+\beta+\gamma+\delta+a+b}+\frac{b+\alpha}{\alpha+\beta+\gamma+\delta+b+c} \geq \frac{b}{\alpha+\beta+\gamma+\delta+a+b+c}$$
$$\Leftarrow \frac{b+\alpha}{\alpha+\beta+\gamma+\delta+a+b+c}+\frac{b+\alpha}{a+\beta+\gamma+\delta+b+c} \geq \frac{b}{\alpha+\beta+\gamma+\delta+a+b+c}$$
$$\Leftrightarrow \frac{\alpha}{\alpha+\beta+\gamma+\delta+a+b+c}+\frac{b+\alpha}{\alpha+\beta+\gamma+\delta+b+c} \geq 0$$

5. Shotest path is a distance measure Given a graph G=(V,E,W), where by V is the set of vertex, E is the set of edge which is a sub set of VxV, W is the set of weight, for each edge there is a non-negtive weight.

   Define len(V')=sum(w'∈W) where V'⊆V, w's are with V' corresponded weight.

   Define sp(x, y): ∀x∈V,∀y∈V:(V(x, y)∨(∃z∈V:sp(x, z)∧V(z, y))∧(¬∃z∈V:len(sp(x, z))+len(sp(z, y))<len(sp(x, y))).

   1. Non-negativity: Since lenth of shotest path is defined as sum of weight of subset of edge, and weight edges is non-negative, length of shothest path is non-negative.
   2. Symmertricity: Since the graph G is a undirected graph, length(sp(x,y))=length(xp(y,x))
   3. Identity: Obviously len(sp(x, x))=0 ∀x∈V
   4. Trangle inequality:

      given sp(x, y) (∀x∈V,∀y∈V) is a shortest path between x and y.

      assume ∃z∈V, let len(sp(x,z))+len(sp(z,y))<len(sp(x,y)).

      Then z either on sp(x,y), which meet a contradiction of the defination of length of a path, or not on sp(x,y), which leads to another contradiction that x->z->y is the shortest path, instead of x->y.

      ∴len(sp(x,z))+len(sp(z,y))>=len(sp(x,y))