L3S Research Center
Large Scale Data Mining, SS 2016
Dr. Avishek Anand
Solution to Assignment 2, due: **28 April 2016**

Problem 1. Suppose we have a universal set U of n elements, and we choose two subsets S and T at random, each with m of the n elements. What is the expected value of the Jaccard similarity of S and T ?

Solution:
The Jaccard Similarity of 2 sets is defined as: $\frac{S \cap T}{S \cup T}$
The probability of that S and T intersect (at least one element) is given by: $P(S \cap T) = \frac{m}{n} * \frac{m}{n}$
Let $P(S \cap T) = \frac{m}{n} * \frac{m}{n} = p$, which implies that the probability of S and T not intersecting ($|S \cap T| = 0$) is $1 - p$.
Now to calculate the expectation of $S \cap T$, we must first find $P(|S \cap T| = i)$, where $1 \leq i \leq m$

$$P(|S \cap T| = i) = p^i (1-p)^{m-i} \tag{1}$$

Therefore the expectation of $S \cap T$ is

$$\begin{aligned} E(S \cap T) &= \sum_{i=0}^{m} i\big[P(|S \cap T| = i)\big] \\ &= \sum_{i=0}^{m} i\big[p^i (1-p)^{m-i}\big] \\ &= \sum_{i=0}^{m} i\big[\big(\frac{m}{n}\big)^{2i} (1 - \big(\frac{m}{n}\big)^2)^{m-i}\big] \end{aligned} \tag{2}$$

Similarly take $S \cup T$,

$$E(S \cup T) = 2m - E(S \cap T) \tag{3}$$

The estimated Jaccard similarity is thus calculated as

$$\frac{E(S \cap T)}{E(S \cup T)} = \frac{\sum_{i=0}^{m} i\big[\big(\frac{m}{n}\big)^{2i} (1 - \big(\frac{m}{n}\big)^2)^{m-i}\big]}{2m - E(S \cap T)} \tag{4}$$

Problem 2.

1. If we use the stop-word-based shingles, and we take the stop words to be all the words of three or fewer letters, then what are the shingles in the sentence **'Even if we hash them to four bytes each, the space needed to store a set is still roughly four times the space taken by the document.'**
2. What is the largest number of $k-$shingles a document of n bytes can have? You may assume that the size of the alphabet is large enough that the number of possible strings of length k is at least as n.

Solution: 1: First eiminate all stop words. Let the shingling be done on a word level. Then set k (for example k=3). Shingles = $\{Even\,hash\,them,\ hash\,them\,four, ...\}$
2: $n - k + 1$

Problem 3. Suppose we want to use a MapReduce framework to compute minhash signatures. If the matrix is stored in chunks that correspond to some columns, then it is quite easy to exploit

parallelism. Each Map task gets some of the columns and all the hash functions, and computes the minhash signatures of its given columns. However, suppose the matrix were chunked by rows, so that a Map task is given the hash functions and a set of rows to work on. Design Map and Reduce functions to exploit MapReduce with data in this form.

Solution:

Let $\mathcal{R}$ be the set of all rows, $\mathcal{S}$ be the set of all documents (columns) being compared, and H be the set of all hash functions. Each mapper takes as input:
$< R, \mathcal{S}, H >$ where $R \subset \mathcal{R}$. Each mapper gets a unique R. The mapper computes the signature for R rows and outputs key value pairs for all $S \in \mathcal{S}$ and $h \in H$ (h corresponds to a single hash function). The value of the key $< S, h >$ is the signature value corresponding to that row and column.
There is a reducer for each element in the characteristic matrix since the mapper phase produces a key,value pair for each cell. In a single reducer we have the signature values (dentoed by the set V) for all $R \subset \mathcal{R}$ of a single document(coulmn) S for a single hash funtion h. The output of the reducer is then the $min(V)$ which gives us the first row where we found a 1 in the characteristic matrix. Thus the number of reducers is $|\mathcal{S}| * |H|$ and the number of mappers is equal to the number of subsets of rows or at most $|\mathcal{R}|$.

Problem 4. Approximate the S-curve $1 - (1 - s^r)^b$ when s^r is very small.
Hint: $(1 + a)^b = e^{ab}$, where a is very small.

Solution:

$1-(1-s^r)^b$, is the probability that signatures agree in all rows of at least one band, and therefore become a candidate pair.
Let $a = s^r$. The S-curve is then given by:

$$\begin{aligned} 1 - (1 - s^r)^b &= 1 - (1 - a)^b \\ &= 1 - e^{-ab} \end{aligned} \tag{5}$$

When a is very small we get:

$$lim_{a \to 0}\ 1 - e^{-ab} = 1 - 1 = 0 \tag{6}$$

Since a is very small e^{-a} is close to 1. This means that $1 - (1 - s^r)^b$ tends to 0 when s^r is very small. Thus the probability of generating candidate pairs is very low if the probability of a signature match in r rows is also very low.

Problem 5. For the (r, b) pair:

1. $r = 6$ and $b = 20$
2. $r = 3$ and $b = 10$
3. $r = 35$ and $b = 60$

compute the threshold, that is, the value of s for which the value of $1 - (1 - s^r)^b$ is exactly 1/2. How does this value compare with the estimate of $(1/b)^{1/r}$ that was suggested in the lecture?

Solution:

1: Estimating the value of s can be done by:

$$\frac{1}{b}^{\frac{1}{r}} = \frac{1}{20}^{\frac{1}{6}} \tag{7}$$

Now we find the value of s using the formula $1-(1-s^r)^b$, which is given to be $1/2$

$$1-(1-s^3)^{10} = 1/2 \tag{8}$$

Solving for s we get: s=0.569
estimated value = 0.607

2: s= 0.406, estimate = 0.464
3: s= 0.880, estimate = 0.89
As the number of rows and bands increase, the more accurate the estimate is, but the computation involved also increases.

Problem 6. On the space of nonnegative integers, which of the following functions are distance measures? If so, prove it; if not, prove that it fails to satisfy one or more of the axioms.

1. max(x,y) = the larger of x and y.
2. diff(x,y) = $|x-y|$ (the absolute magnitude of the difference between x and y).
3. sum(x,y) =$x+y$.
4. Jaccard distance
5. shortest path between a pair of nodes in an weighted (weights are non-negative) undirected graph.

Solution:

1. Not a disance measure. The axiom $d(x,y)=0$ if $x=y$ does not hold since $d(x,y)=x$ or y but not 0.
2. It is a distance measure. Triangle inequlaity also holds.

   Proof by Contradiction Assume $|x-y| > |x-z|+|y-z|$ holds, where $z \geq 0$.
   if $z=0$ then the assumption is invalid.
   Now consider $z = x+1$,

$$\begin{aligned} |x-z|+|y-z| &= |x-(x+1)|+|y-(x+1)| \\ &= 1+|y-x-1| \\ &= 1+|x-y+1| \end{aligned} \tag{9}$$

   but,

$$|x-y| < 1+|x-y+1| \tag{10}$$

   which is contraray to our assumption again.

   This assumption is false for all n where, $z = x+n$ and $n>0$.

3. Not a distnace measure since $d(x, y) = 2x$ and not 0 when $x = y$.

4. Distance Metric.
Jaccard Distance $J(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$
Non-negative: When X=Y, $|X \cap Y| = |X \cup Y|$ and $J(X, Y) = 0$.
When X and Y are disjoint then $|X \cap Y| < |X \cup Y|$ and $0 < J(X, Y) < 1$
Symmetry: since union and intersection are symmetric
Identity: When X=Y, $|X \cap Y| = |X \cup Y|$ and $J(X, Y) = 0$.
Triangle Inequality:

$$1 - \frac{|X \cap Y|}{|X \cup Y|} \leq 1 - \frac{|X \cap Z|}{|X \cup Z|} + 1 - \frac{|Z \cap Y|}{|Z \cup Y|} \tag{11}$$

Let $Z = \emptyset$, then

$$\begin{aligned} J(X, Y) &\leq 1 - \frac{|X \cap Z|}{|X \cup Z|} + 1 - \frac{|Z \cap Y|}{|Z \cup Y|} \\ &< 2 \end{aligned} \tag{12}$$

and $J(X, Y) \leq 1$. For a set $Z = Y$ then

$$J(X, Y) = 1 - \frac{|X \cap Z|}{|X \cup Z|} \tag{13}$$

$$J(X, Y) \leq J(X, Y) + 1 - \frac{|Z \cap Y|}{|Z \cup Y|} \tag{14}$$

Even if $|Y \cap Z| = 0$ the triangle inequality holds.

5. It is a metric.
Let the shortest distance between two points x and y, on a graph is denoted by $d(x, y)$
Non-negativity: Any path will have a weight greater than 0. Hence the shortest path will also be non-negative.
Identity: Shortest path to the same point from a given point is always 0.
Symmetry: Since the graph is undirected $d(x, y) = d(y, x)$.
Triangle inequality: Proof by contradiction
Assume $d(x, y) > d(x, z) + d(y, z)$ holds.
If z is a point on the shortest path between x and y then

$$d(x, y) = d(x, z) + d(y, z) \tag{15}$$

which is contrary to our assumption. Now for a point z not lying on the shortest path then

$$d(x, y) > d(x, z) + d(y, z) \tag{16}$$

does not hold, which means that $d(x, y)$ is not the shortest path between x and y anymore. This violates our definition of $d(x, y)$ and hence contradicts our assumption.