Algorithm Foundations of Data Science and Engineering
Welcome Tutorial :-)
Tutorial 2

GAO Ming

DaSE @ ECNU

7 Mar., 2019

Tutorial 2

1. Compute the Jaccard similarities of each pair of the following three sets: $\{1,2,3,4\}$, $\{2,3,5,7\}$, and $\{2,4,6\}$.
   Let $A = \{1,2,3,4\}$, $B = \{2,3,5,7\}$, and $C = \{2,4,6\}$.
   $sim(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{1}{3}$; $sim(A,C) = \frac{|A \cap C|}{|A \cup C|} = \frac{2}{5}$; $sim(B,C) = \frac{|B \cap C|}{|B \cup C|} = \frac{1}{6}$
2. Prove that if the Jaccard similarity of two columns is 0, then minhashing always gives a correct estimate of the Jaccard similarity.
   Let X be a doc(set of shingles), $y \in X$ is a shingle.
   Let y be $s.t.\pi(y) = min(\pi(C_1 \cup C_2))$, then $\pi(y) = min(\pi(C_1))$ or $\pi(y) = min(\pi(C_2))$
   Thus, the prob. that both are true is the prob. $y \in C_1 \cap C_2$.
   Final, we have $P(min(\pi(C_1)) = min(\pi(C_2))) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} = sim(C_1, C_2)$
   So, if $sim(C_1, C_2) = 0$, then $P(min(\pi(C_1)) = min(\pi(C_2))) = 0$

Tutorial 2

3. a. Compute the Jaccard similarity of each of the pairs of columns.

$sim(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} = \frac{1}{4}$

$sim(S_1, S_3) = \frac{|S_1 \cap S_3|}{|S_1 \cup S_3|} = \frac{1}{4}$

$sim(S_2, S_3) = \frac{|S_2 \cap S_3|}{|S_2 \cup S_3|} = \frac{0}{4} = 0;$

b. Compute the minhash signature for each column if we use the following three hash functions: $h_1(x) = 7x + 1 \bmod 6$; $h_2(x) = 11x + 2 \bmod 6$; $h_3(x) = 5x + 2 \bmod 6$.

Element	S_1	S_2	S_3
0	1	1	0
1	0	1	0
2	1	0	0
3	0	0	1
4	1	0	1
5	0	0	0

Tutorial 2

3. b.

Table: Element location after hash.

h1	h2	h3
1	2	2
2	1	1
3	0	0
4	5	5
5	4	4
0	3	3

Table: Minhash Signature

hash	s_1	s_2	s_3
h_1	1	1	4
h_2	0	1	4
h_3	0	1	4

Tutorial 2

4. For LSH, please to determine the similarity threshold t, i.e., the value of similarity t at which the probability of becoming a candidate is 1/2, which can be a function of b and r.
   Prob. that all rows in band equal $= t^r$
   Prob. that some row in band unequal $= 1 - t^r$
   Prob. that no band identical $= (1 - t^r)^b$
   Prob. that at least 1 band identical $= 1 - (1 - t^r)^b$

Tutorial 2

5. Let two sets S_1 and S_2 be presented in the form of binary vectors, $\{h_1, \cdots, h_k\}$ be k random permutations, and $h_i(S)$ record the first 1 in each column after permutation. Please prove that $\widehat{JS}(S_1, S_2) = \frac{1}{k}\sum_{i=1}^{k} X_i$ is within ε error with probability at leat $1-\delta$ if $k = \frac{2\ln(1/\delta)}{\varepsilon^2}$, where $JS(A, B) = \frac{|A \cap B|}{|A \cup B|}$, and
   $X_i = \begin{cases} 1, & \text{if } h_i(S_1) = h_i(S_2); \\ 0, & \text{otherwise.} \end{cases}$
   We want to prove that $P(|\hat{JS} - JS| > \varepsilon JS) < \delta$
   The left-hand side is equal to
   $P(\hat{JS} > (1+\varepsilon)JS + P(\hat{JS} < (1-\varepsilon)JS) < 2exp(-\mu\varepsilon^2/4), \quad \mu = kp$
   So, $2exp(-kp\varepsilon^2/4) = \delta$
   $k = \frac{4ln(1/\delta)}{p\varepsilon^2}$