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

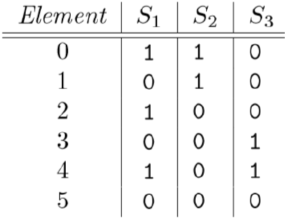
GAO Ming

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 $\{2,4,6\}$.
2. Prove that if the Jaccard similarity of two columns is 0, then minhashing always gives a correct estimate of the Jaccard similarity.
3. a. Compute the Jaccard similarity of each of the pairs of columns.
   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

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.
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 $X_i = \begin{cases} 1, & \text{if } h_i(S_1) = h_i(S_2); \\ 0, & \text{otherwise.} \end{cases}$