PSG College of Technology, Coimbatore -641 004

Department of Applied Mathematics and Computational Sciences

8th Semester MSc TCS

18XT87 Data Mining Lab

Problem Sheet – 4

1. For the following vectors x and y, calculate the indicated similarity or distance measures.
   (a) x=(1, 1, 1, 1) , y= (2, 2, 2, 2) cosine, correlation, Euclidean
   (b) x=(0, 1, 0, 1) , y= (1, 0, 1, 0) cosine, correlation, Euclidean, Jaccard
   (c) x=(0, -1, 0, 1) , y= (1, 0, -1, 0) cosine, correlation, Euclidean
   (d) x=(1, 1, 0, 1, 0, 1) , y= (1, 1, 1, 0, 0, 1) cosine, correlation, Jaccard
   (e) x=(2, -1, 0, 2, 0, -3) , y= (-1, 1, -1, 0, 0, -1) cosine, correlation

2. This exercise compares and contrasts some similarity and distance measures.
   For binary data, the Hamming distance is the number of bits that are different between two binary vectors. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance and the Jaccard similarity between the following two binary vectors.

   X=0101010001
   Y=0100011000