Computer Science COMP-4250 - Winter 2021

Assignment 2

Due: End of Sunday, March 28, 2021

****

Note: Submit your assignment as a **single zip file** on Blackboard. The zip file should be named as Assignment1_*StudentId*.zip, in which replace *StudentId* with your university student number.

****

Problem 1. (9 points)

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.

(a) ***max(x,y) = the larger of x and y***.

(b) ***diff(x,y) = |x−y|*** (the absolute magnitude of the difference between ***x*** and ***y***).

(c) ***sum(x,y) = x+y***.

Problem 2. (9 points)

Find the L_1, L_2, and L_∞ distances between the points ***(5,6,7)*** and ***(8,2,4)***. Note that L_n is the norm distance in n-dimensional Euclidean space. In General, L_r-norm is the distance measure ***d*** defined by:

$$d([x_1, x_2, \ldots, x_n],\ [y_1, y_2, \ldots, y_n]) = (\sum_{i=1}^{n} |x_i - y_i|^r)^{1/r}$$

Problem 3. (5 points)

Prove that if ***i*** and ***j*** are any positive integers, and ***i<j***, then the L_i-norm between any two points is greater than the L_j-norm between those same two points.

Problem 4. (9 points)

Find the **edit distances** (using only insertions and deletions) between the following pairs of strings.

(a) ***abcdef*** and ***bdaefc***.

(b) ***abccdabc*** and ***acbdcab***.

(a) ***abcdef*** and ***baedfc***.

Problem 5. (5 points)

Perform a hierarchical clustering of the one-dimensional set of points ***1,4,9,16,25,36,49,64,81***, assuming clusters are represented by their centroid (average), and at each step the clusters with the closest centroids are merged.

Problem 6. (8 points) (Please use the book for this problem)

Exercise 12.2.1: Modify the training set of Fig. 12.6 so that example b also includes the word "nigeria" (yet remains a negative example – perhaps someone telling about their trip to Nigeria). Find a weight vector that separates the positive and negative examples, using:

(a) The basic training method of Section 12.2.1.

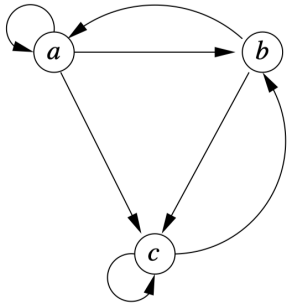
(b) The Winnow method of Section 12.2.3.

(c) The basic method with a variable threshold, as suggested in Section 12.2.4.

(d) The Winnow method with a variable threshold, as suggested in Section 12.2.4.

Problem 7. (10 points)

Considering the following Web graph:

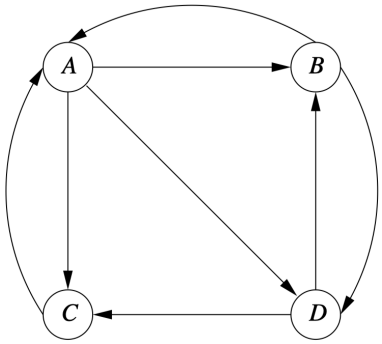


(a) Compute the PageRank of each page without teleporting.

(b) Compute the PageRank of each page with teleporting, assuming $\beta=0.8$.

Problem 8. (10 points)

Considering the Web graph below, assuming only **B** is a trusted page:

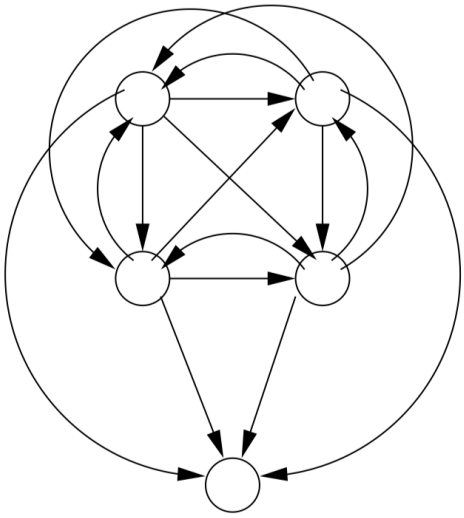


(a) Compute the TrustRank of each page.

(b) Compute the spam mass of each page.

Problem 9. (15 points)

Consider the following Web graph:



This Web has a clique (set of nodes with all possible arcs from one to another) of n nodes and a single additional node that is the successor of each of the n nodes in the clique, for the case $n=4$. Determine the PageRank of each page, considering $\beta=0.8$.

Problem 10. (10 points)

Considering DGIM approach, there are several ways that the bit-stream below could be partitioned into buckets. Find all of them.

1001101001101101011011011001

Problem 11. (10 points)

Describe what happens to the buckets if three more 1's enter the window represented by the stream below. You may assume none of the 1's shown leave the window.