Zeid Al-Ameedi
Cpts 315 Homework 1
02-03-2020

Analytical Part

Q1

a) 4
b) (4/6) = 0.66
c) 66%

Q2

a) Given that n=20; For pair [7, 8] the position = (i - 1) * (n - (i / 2)) + j - i = (7 - 1) * (20 - (7 / 2)) + (8 - 7) = (6 * 16.5) + (8 - 7) = 100
b) If there are fewer elements, you should proceed with the triangular method because of efficiency.

Q3

a)

Support(1)	4
Support(2)	6
Support(3)	8
Support(4)	8
Support(5)	6
Support(6)	4
Support({1,2})	2
Support({1.3})	3
Support({1,4})	2
Support({1,5})	1
Support({1,6})	0
Support({2,3})	3
Support({2,4})	4
Support({2,5})	2
Support({2,6})	1
Support({3,4})	4
Support({3,5})	3
Support({3,6})	2
Support({4,5})	3
Support({4,6})	3
Support({5,6})	2

b)

{1,3}	3
{1,4}	4
{1,5}	5
{1,6}	6
{2,3}	6
{2,4}	8
{2,5}	10
{2,6}	1
{3,4}	1
{3,5}	4
{3,6}	7
{4,5}	9

{4,6}	2
{5,6}	8

c) To determine if a bucket is frequent in pass 1 outcome, note that the support threshold pairs are hashed in that bucket. Observe that buckets 1, 2, 4, 8 are frequent buckets.

d) Let {I, j} be the pair counted within the 2nd pass. Observe that it must follow through with multiple conditions (according to algorithm in my notes)
1. I and j are frequent items
2. Pair {I, j} are hashed to a frequent bucket.
Notice that all the items are frequent items so condition 1 is correctly assumed to be satisfied. Condition two (by observing the table) is only satisfied by the following pairs
{1,2} {1,4} {2,4} {2,6} {3,4} {3,5} {4,6} {5,6}

NOTE: Those are the pairs counted within the 2nd pass.

Q4

Detecting copies has become super essential in the world we live in today. This document introduces the class of local document fingerprinting algorithms. The process captures an essential property of the presented fingerprinting technique which appears to guarantee to detect copies. Most applications will find it is useful to record the fingerprints of the document but also the location of those fingerprints within the document.

The paper dives into a great example where the following is applied. Given a database D of fingerprints (f) generated on some window W. Essentially running through the theory the document is fingerprinted and looked up inside an index of all possible matching fingerprints for each document.

Essentially the paper is saying don't cheat.