Sonam Yangtso
Cpts 315
Homework #1

Analytical Part:

1)

a) The Absolute support for the item set {A, B} = 4
b) The Relative support for the item set {A, B} is 4/6 = .67
c)The confidence of Association Rule for A => B is 4/6 = .67

2)

a) N = 20, i = 7, j = 8
Index = (i - 1)(n - i/2) + (j - 1)
= (7 - 1)(20 - 7/2) +(8 -1)
= (6)(16.5)+(7)
= 106

b) We chose the Tabular method over the Triangular matrix if the pairs which have a nonzero count are less or equal to ⅓ of all pairs. Since 10 % is less than ⅓, therefore we choose Tabular method.

3)

Given that Support threshold = 4
Support for each item :

Item	Support
1	4
2	6
3	8
4	8
5	6
6	4

Support for each pair of Items:

Item Pairs	Support
(1,2)	2
(1,3)	3
(1,4)	2
(1,5)	1
(1,6)	0
(2,3)	3
(2,4)	4
(2,5)	2
(2,6)	1
(3,4)	4
(3,5)	4
(3,6)	2
(4,5)	3
(4,6)	3
(5,6)	2

b) Pairs will be hashed to bucket based on hash function (i * j mod 11)

Bucket number	pair	count
0		
1	(2,6),(3,4)	1+4=5
2	(1,2),(4,6)	2+3=5
3	(1,3)	3

4	(1,4),(3,5)	2+4=6
5	(1,5)	1
6	(2,3)	3
7	(3,6)	2
8	(2,4),(5,6)	2+2=4
9	(4,5)	3
10	(2,5)	2

c)

Bucket number	Pair	Count
1	(2,6),(3,4)	5
2	(1,2),(4,6)	5
4	(1,4),(3,5)	6
8	(2,4),(5,6)	4

A Bucket is frequent if the number of pairs hashed to each bucket is greater or equal to the given support threshold. From the above table, I can conclude that Bucket number 1,2,4 and 8 are frequent buckets.

d) Pairs (i,j) are counted to the second pass of the PCY algorithm if the following conditions were satisfied.

1) If both i and j are frequent item
2) The pair {i,j} hashed to a bucket which is a frequent

Therefore,(2,6), (3,4),(1,2),(4,6),(1,4), (3,5) ,(5,6)and (2,4) pairs will count on the second pass of the PCY algorithm.

4) In this report he has discussed fingerprinting, a copyright protection technique and methods with their advantages and disadvantages. Plagiarism occurs everywhere in this world from the students to the business world. It's not fair getting someone's hard work credit by someone else by copying from them. Therefore, we need a copy detection technique to detect any level of plagiarism and any illegal copying documents or papers is called fingerprinting.In this research report , the author was trying to explain various types of fingerprinting detection algorithms. All cpy detection algorithms should have the following three properties. White space insensitive, Noise suppression and position independent.

K-gram is one of the techniques used for detecting partial copies.They used the substring of length k.Rabin-Karp Algorithm is apparently the earliest version of fingerprinting based on k-grams and it is for fast substring matching algorithm. This algorithm is mainly used in genetic detection.Since it works on finding occurence of particular string of length k in a much longer string.This algorithm works on the principle of hashing and it has some disadvantages working on hashing especially with relatively smaller characters.

All to all matching techniques was the first scheme to apply fingerprinting to collections of documents and it was developed by Manber, who discovered Karp-Rabin string matching and applied it to detecting similar files in file systems.Above all three algorithm work based on k-gram technique. This technique is easy to implement but it doesn't guarantee you will find matches between documents are detected.This technique fails to detect the matches if there is any gap between the strings.

Instead of using k-grams technique, the strings to fingerprint can be chosen by looking for sentences or paragraphs, or by choosing fixed-length strings.

In this paper, the author mainly focused on explaining the winnowing algorithm. Winnowing is a fingerprinting algorithm for documents. The Winnowing selects fingerprints from hashes of k-grams, a contiguous substring of length k.Given a set of documents, we want the find substring matches between them that satisfy two properties:

1. If there is a substring match at least as long as the guarantee threshold, t, then this match is detected, and
2. We do not detect any matches shorter than the noise threshold.

In each window select the minimum hash value. If there is more than one hash with the minimum value, select the rightmost occurrence. Now save all selected hashes as the fingerprints of the document.In many applications it is useful to record not only the fingerprints of a document, but also the position of the fingerprints in the document. Author talked about the local algorithm and how effective it was from the winnowing algorithm. At the end of the paper he showed his experiment on winnowing on real data and showed us how effective it was on detecting fingerprinting.