Blake Calvin
Homework 1

Question 1

a) The item set {A,B} is in baskets 1, 3, 4, and 6. So the absolute support of the item set {A, B} is 4.

b) To find relative support…

relative support = absolute support / total baskets
relative support = 4 / 6

relative support = 0.66 or 66%

c) To find confidence of A => B, first find all baskets A are in.
A is in every basket, so A is contained in 6baskets.

Then we find all baskets B is that also has A in it.
B is in 4 baskets that A is also in, so 4 baskets.

So the confidence is 4/6
0.66 or 66%

Question 2

a) Position of {i,j} where i<j is the formula below..

Position = (i-1)(n-i/2) + j – i | i = 7, j = 8, n = 20

= (7-1)(20-(7/2)) + 8-7
= (6)(20-3.5) + 1
= (6)(16.5) + 1
= 99 + 1
= 100

b) The tabular method would be preferred. This is because the tabular method is better than the triangular method when at most 33% of all pairs have a nonzero count. And since the question told us that we have knowledge that only 10% of the total pairs have a nonzero count, then since 10% < 33%, we can safely say that tabular method is better.

Question 3

a) Since all the items possible in the baskets are {1, 2, 3, 4, 5, 6}, we need to find the support for each of those items and all their pair combinations

Starting with their individual supports, we just count each time one of the items is in a basket and check every basket. So doing that gives us the following data.

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

Then we find all supports for every possible pair of 1, 2, 3, 4, 5, and 6

Item Pair	**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}	3
{3,6}	2
{4,5}	3
{4,6}	3
{5,6}	2

b) Hash function = i * j mod 11

Item Pair	Function	Bucket
{1,2}	1 * 2 mod 11 =	2
{1,3}	1 * 3 mod 11 =	3
{1,4}	1 * 4 mod 11 =	4
{1,5}	1 * 5 mod 11 =	5
{1,6}	1 * 6 mod 11 =	6
{2,3}	2 * 3 mod 11 =	6
{2,4}	2 * 4 mod 11 =	8
{2,5}	T2 * 5 mod 11 =	10
{2,6}	2 * 6 mod 11 =	1
{3,4}	3 * 4 mod 11 =	1
{3,5}	3 * 5 mod 11 =	4
{3,6}	3 * 6 mod 11 =	7
{4,5}	4 * 5 mod 11 =	9
{4,6}	4 * 6 mod 11 =	2
{5,6}	5 * 6 mod 11 =	8

c) A bucket is considered frequent if the bucket count is at least the support threshold. Since there are 11 buckets, and none of them count higher than 2 (as shown in the table from question 3a), and since the support threshold is 4, then we know that none of the buckets are frequent.

d) For a pair to be in pass 2, both members in the pair must be frequent items, and the pair must be hashed to a frequent bucket.
We know that every item in the itemset is frequent, however none of the buckets are frequent, so this doesn't satisfy the second condition meaning that none of the pairs move to pass 2.

Question 4

The paper "Winnowing: Local Algorithms for Document Fingerprinting" is a paper that addresses the ongoing issue and conflicts with document fingerprinting. The issue this paper and, the winnowing algorithm later to be addressed is trying to resolve, is that there are ongoing issues with document copying. These issues can be students plagiarizing, multiple versions of the same document being produced, web sites are being mirrored, etc. This causes problems for businesses or document owners if there is either misinformation being spread, or someone isn't getting their share of what they earned since their document is copied elsewhere (There are many other issues, but this is just a couple). This ongoing problem exists because detecting partial similarities between documents can be difficult to do well and optimized. Which introduces the idea of document fingerprinting.

This document fingerprinting is helpful because it helps tell us how similar documents are to each other based off their number of shared fingerprints. And in this paper, the whole premise is centralized around the idea that a new method of document fingerprinting is being developed which is called Winnowing. This winnowing algorithm is efficient and guarantees matches of certain length strings are detected each time. This works because the algorithm is done by separating the text in the documents into smaller sections of text and utilizes hashing functions by hashing the smaller text and using the hash values to compare similarities between documents.

This is beneficial because since the authors are trying to solve the issue of partial similarities between documents rather than the whole document itself, dividing up the documents into smaller text hashes let us compare micro versions of these documents to compare their fingerprints. The paper even describes some of the experiments they underwent using the Winnowing algorithm and it proved to be more efficient than other forms of documents fingerprinting, there was even some results as high as to prove 82% of the fingerprints using the Winnowing algorithm was only chosen once.

This whole paper is helpful to understand more about data scraping and learning about similarities of all forms of documents. The winnowing algorithm is an interesting process to observe and has shown effectiveness over many other algorithms.