Mark Shinozaki
11672355
Cpts 315 – Homework #1
Analytical Part

Q1. Consider the following market-basket data, where each row is a basket and shows the list of items that are part of that basket.

1. {A,B,C}
2. {A,C,D,E}
3. {A,B,F,G,H}
4. {A,B,F,G,H}
5. {A,C,D,P,Q,R,S}
6. {A,B,L,M,N}

(a) What is the absolute support of item set {A,B}?
- It appears that the absolute support of the item set {A,B} is 3, it appears in the baskets 1,3,4 and 6.

(b) What is the relative support of item set {A,B}?
- Since the item set of {A,B} is (3/6) in the total set, this would mean its 50% or the item set {A.B} appears in 50% of the total baskets in the market-basket data.

(c) What is the confidence of association rule A⟶ B
- The confidence of association rule is the # of baskets containing both A and B divided by the # of baskets containing A. The # of baskets containing A and B is 3 and of baskets containing A is 6, so the confidence of association rule is (3/6) or 50%.

Q2. Answer the below questions about storing frequent pairs using triangular matrix and tabular method.

a. Suppose we use a triangular matrix to count pairs and the number of items n = 20. If we store this triangular matrix as a ragged one-dimensional array Count, what is the index where count of pair (7,8) is stored?
- With the pair (7,8), the index where the count of pair (7,8) is stored would be: 21. The breakdown would look like 7 * (7-1)/2+8=21.

b. Suppose you are provided with the prior knowledge that only ten percent of the total pairs will have a non-zero count. In this case, which method among triangular matrix and tabular method should be preferred and why?
- The tabular method stores the counts of all possible pairs in a two-dimensional array. A tabular method is recommended because only 10% of the total pairs will have a non-zero count, this means that 90% of the entries in the array will be a zero value. In a triangular matrix method it only stores the non-zero count values in the lower or upper triangular part of the matrix. Given that 10% of the total pairs have a non-zero count the tabular method would be the option to go with since it would be more memory-efficient and easier to retrieve the count values than the compared method of the triangular matrix method.

Mark Shinozaki
11672355
Cpts 315 – Homework #1
Analytical Part

Q3. This question is about the PCY algorithm for counting frequent pairs of items. Suppose we have six items numbered 1,2,3,4,5,6. Consider the following twelve baskets.

1. {1,2,3}
2. {2,3,4}
3. {3,4,5}
4. {4,5,6}
5. {1,3,5}
6. {2,4,6}
7. {1,3,4}
8. {2,4,5}
9. {3,5,6}
10. {1,2,4}
11. {2,3,5}
12. {3,4,6}

Suppose the support threshold is 4. On the first pass the PCY algorithm, we use a hash table with 11 buckets, and the set $\{i,j\}$ is hashed to $i \times j$ mod 11.

a. By any method, compute-the support for each item and each pair of items.
   - (1,2): 2/(1,3): 4/(1,4): 2/(1,5): 2/(1,6): 0/(2,3): 4/(2,4): 4/(2,5): 2/(2,6): 2/(3,4): 4/(3,5): 4/(3,6): 2/(4,5): 4/(4,6): 2/(5,6): 2/
b. Which pairs hash to which buckets?
   Bucket 0: (6,6)/Bucket 1: (2,6),(6,2)/Bucket 2: (1,2), (2,1)/Bucket 3: (3,6), (6,3)/Bucket 4: (2,4), (4,2), (4,6), (6,4)/Bucket 5: (1,5), (5,1), (5,6), (6,5)/ Bucket 6: (3,4), (4,3)/Bucket 7: (3,5), (5,3)/Bucket 8: (2,3), (3,2)/Bucket 9: (1,4), (4,1)/Bucket 10: (1,3), (3,1), (2,5), (5,2)
c. Which buckets are frequent?
   - Bucket 1 – 5 and 9 are not frequent, Bucket 6 – 8 and 10 are frequent
d. Which pairs are counted on the second pass of the PCY algorithm?
   - (3,4), (3,5), (2,3), (1,3)

Mark Shinozaki
11672355
Cpts 315 – Homework #1
Analytical Part

Q4. Please read the following paper and write a brief summary of the main points in at most ONE page. You can skip the theoretical parts.

The article first starts out explaining the various ways people reuse media on the internet. It starts off by saying that digital content is interacted with in a few ways such as copying, quotation, revision, plagiarism and sharing. The article then goes on to explain the main focus of the article which is the topic of digital fingerprinting. Digital fingerprinting is the topic of being able to accurately identify copying, as well as small partial copies and large copies. Then, the authors talk more about how they came up with a solution which includes coming up with algorithms for finding/detecting copies. Then the authors talk about their own fingerprinting algorithm they developed knowing as winnowing. In the introduction the authors go over a few reasons as to why documents would be copied and how documents would be copied, as well as, the kinds of current plagiarism methods exist and how they are currently being detected. Then the authors go over how their algorithm is efficient for selecting fingerprints from a sequence of hashes that guarantee that at least part of any sufficiently long match is detected. Then the authors explain their algorithm step by step, firstly the authors explain what a k-gram is, which is an important first step in understanding their algorithm. K-gram is a contiguous substring of length k. Divide a document into k-grams, where k is a parameter chosen by the user. The author goes on to explain how it chooses an efficient algorithm for selecting fingerprints from a sequence of hashes that guarantees that at least part of any sufficiently long match is detected. Then after the authors go over how the algorithm works they then go over how the Karp-Rabin string matching algorithm works. The Karp-Rabin string matching algorithm is for fast substring matching and is the earliest version of fingerprinting based on k-grams. The Karp-Rabin string matching algorithm uses substring matching that uses a "rolling" hash function to find occurrences of a particular string of length "k" which is a longer string. The hash function treats the k-gram as a k-digit number in base b and calculates the hash by summing up the products of each character and its corresponding power of b. The rolling hash function allows the hash of the next k-gram to be quickly computed from the pervious one by subtracting out the high-order digit, multiplying by b, and adding the new low-order digit. Then the next part of the article, the authors talk about All-to-all matching. The idea of using fingerprinting for document comparison was developed by Manber, who applied the Karp-Rabin string matching algorithm to detect similar files in file systems. Then there's, DRM which is Digital Right Management, DRM systems aim to solve the problem of intellectual property use by controlling copying, but digital content is inevitably copied, leading to the need for efficient copy detection. Then, the authors talked about winnowing, winnowing algorithm is used to find substring matches within a set of documents. The definition of winnowing is, 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. Another way of explaining the definition, winnowing is a method to find the fingerprints of a document from a sequence of hashes. The selected hashes are then saved as the fingerprints of the document. Towards the end of the article the authors compare the performance of winnowing with other fingerprinting algorithms, the authors define a local algorithm as one where the choice of hash depends only on the contents of the window, and not on the position of the window in the file. They prove a lower bound for the density of a local algorithm given random inputs but they say the lower bound can be improved. Then lastly, the authors write in the conclusion that the winnowing algorithm guarantees the detection of matches, is efficient, and that they have conducted enough tests with sufficient data to prove as much.

Mark Shinozaki
11672355
Cpts 315 – Homework #1
Analytical Part
Programming and Experimental Part