Assignment One

**Question One:**

Suppose hash-keys are drawn from the population of all positive integers that are multiples of some constant c, and hash function h(x) = x mod k, and k is a positive integer. For what values of c will h be a suitable hash function, i.e., a large random choice of hash-keys will be divided roughly equally into buckets?

The hash function give in x mod k, as well as all the values from the population being all positive integers, the value of c would have to be between c and k-1.

**Question Two:**

Design MapReduce algorithms to take a very large file of integers and produce as output:

1. The largest integer.

The map phase will find the max value within each chunk. Each chunk’s max value is sent to the reduce phase where a max value is then chosen there.

The reduce phase will take all the chunks max values and then reduce them to a single max value.

1. The average of all the integers.

The map phase will complete an average. It will take a key of 1 and a key value of weigh, which will dictate the amount of integers in that map job and average, the average of those integers.

The reduce phase will then average the values as the key between all jobs are the same.

1. The same set of integers, but with each integer appearing only once.

The map phase will return a key value pair for each input, key = value

The reduce phase will create a list of keys, which is the output.

1. The count of the number of distinct integers in the input.

The map phase will create key value pairs, the key being 1, the value being the input.

The reduce phase will then only take the key values pairs with a key = 1, and for each of these pairs a counter would be incremented.

Note: You don’t need to develop any program. You can follow the example we had for counting the words in some documents.

**Question Three:**

Suppose there are 100 items, numbered 1 to 100, and also 100 baskets, also numbered 1 to 100. Item *i* is in basket b if and only if *i* divides b with no remainder. Thus, item 1 is in all the baskets, item 2 is in all fifty of the even-numbered baskets, and so on. Basket 12 consists of items {1,2,3,4,6,12}, since these are all the integers that divide 12. Answer the following questions:

1. If the support threshold is 5, which items are frequent?

There are 100 items max in the list, and the support thresh hold is 5.

In this case we are looking for item that appear in at least 5 buckets.

To appear in 5 buckets, then 100/*i* >= 5, must hold true.

There for items: {1},{2},{3},{4},{5},{6},{7},{8},{9},{10},{11},{12},{13},{14},{15},{16},{17},{18},{19},{20} represent the frequent items.

1. If the support threshold is 5, which pairs of items are frequent?

These are pairs that appear in more than 5 buckets, these pairs are:

(1,2), (1, 3), (1, 4), (1, 5), (1, 6), (1, 7), (1, 8), (1, 9), (1, 10), (1, 11), (1, 12), (1, 13), (1, 14), (1, 15), (1, 16), (1, 17), (1, 18), (1, 19), (1, 20), (2, 3), (2, 4), (2, 5), (2, 6), (2, 7), (2, 8), (2, 9), (2, 10), (2, 12), (2, 14), (2, 16), (2, 18), (2, 20), (3, 4), (3, 5), (3, 6), (3, 9), (3, 12), (3, 15), (3, 18), (4, 5), (4, 6), (4, 8), (4, 10), (4, 12), (4, 16), (4, 20), (5, 10), (5, 15), (5, 20), (6, 9), (6, 12), (6, 18), (7, 14), (8, 16), (9, 18), (10, 20)

1. which basket is the largest?

The largest basket is basket 12.

1. what are the confidence and interest of the following association rules?

{5,7} → 2

{5,7} is found in 2 baskets, meaning it has a support of 2. S({5,7}) = 2

{5,7,2} is only found in 1 basket when explored. S({5,7,2}) = 1

Therefore, by following the association rule:

S({5,7, 2})/S({5, 7}) = ½

{2,3,4} → 5

{2,3,4} is found 8 baskets, meaning it has a support of 8. S({2,4,5}) = 8

{2,3,4,5} is found in 1 basket, meaning it has a support of 1. S({2,4,5,6}) = 1

Therefore, by following the association rule:

S({2,4,5,6})/ S({2,4,5}) = 1/8

1. Show all the association rules that have 100% confidence for this market-basket data.

The association rules where *i* -> 1 would all have a 100% confidence as 1 is in every basket, however, the denominator would need to be less than the numerator.

**Question Four:**

Suppose there are 100 items, numbered 1 to 100, and also 100 baskets, also numbered 1 to 100. Item i is in basket b if and only if b divides i with no remainder. For example, basket 12 consists of items {12,24,36,48,60,72,84,96}, since these are all the integers less than 100 that are dividable by 12.

Answer the following questions:

1. If the support threshold is 5, which items are frequent?

A basket *b* will consist of items that are a multiple of *b*. That being said, a frequent item if it has 5 factors that are less than 100, which is the number of items in the set.

The list of frequent items is:

{12}, {16}, {18}, {20}, {24}, {28}, {30}, {32}, {36}, {40}, {42}, {44}, {45}, {48}, {50}, {52}, {54}, {56}, {60}, {63}, {64}, {66}, {68}, {70}, {72}, {75}, {76}, {78}, {80}, {81}, {84}, {88}, {90}, {92}, {96}, {98}, {99}, {100}

1. If the support threshold is 5, which pairs of items are frequent?

(12, 24), (12, 36), (12, 48), (12, 60), (12, 72), (12, 84), (12, 96), (16, 32), (16, 48), (16, 64), (16, 80), (16, 96), (18, 36), (18, 54), (18, 72), (18, 90), (20, 40), (20, 60), (20, 80), (20, 100), (24, 36), (24, 48), (24, 60), (24, 72), (24, 84), (24, 96), (28, 56), (28, 84), (30, 60), (30, 90), (32, 48), (32, 64), (32, 80), (32, 96), (36, 48), (36, 54), (36, 60), (36, 72), (36, 84), (36, 90), (36, 96), (40, 60), (40, 80), (40, 100), (42, 84), (44, 88), (45, 90), (48, 60), (48, 64), (48, 72), (48, 80), (48, 84), (48, 96), (50, 100), (54, 72), (54, 90), (56, 84), (60, 72), (60, 80), (60, 84), (60, 90), (60, 96), (60, 100), (64, 80), (64, 96), (72, 84), (72, 90), (72, 96), (80, 96), (80, 100), (84, 96)

1. which basket is the largest?

The largest basket is basket 12

1. what are the confidence and interest of the following association rules?

{24,60} → 8.

{24,60} is found in 6 baskets, meaning it has a support of 6. S({24,60}) = 6

{8, 24, 60} is only found in 3 basket when explored. S({8,24,60}) = 3

Therefore, by following the association rule:

S({8, 24, 60})/S({24, 60}) = 3/6 = 1/2

{2,3,4} → 5.

{2,3,4} is found in 8 baskets, meaning it has a support of 8. S({2, 3, 4}) = 8

{2, 3, 4, 5} is only found in 1 basket when explored. S({2, 3, 4, 5}) = 1

Therefore, by following the association rule:

S({2, 3, 4, 5}) S({2, 3, 4})= 1/8

1. Show all the association rules that have 100% confidence for this market-basket data.

The association rules where *i* -> 1 would all have a 100% confidence as 1 is in every basket, however, the denominator would need to be less than the numerator.

**Question Five:**

Suppose the items are numbered 1 to 10, and each basket is constructed by including item i with probability 1/i, each decision being made independently of all other decisions. That is, all the baskets contain item 1, half contain item 2, a third contain item 3, and so on. Assume the number of baskets is sufficiently large that the baskets collectively behave as one would expect statistically. Let the support threshold be 1% of the baskets.

(a)Find the frequent itemsets.

(b)Prove that in this data there are no interesting association rules, i.e., the interest of every

association rule is 0.

**Question Six:**

Text

Description automatically generated

1. **By any method, compute the support for each item and each pair of items.**

Singletons:

support({1}) = 4

support({2}) = 6

support({3}) = 8

support({4}) = 8

support({5}) = 6

support({6}) = 4

Pairs:

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}) = 4

support({3,6}) = 2

support({4,5}) = 3

support({4,6}) = 3

support({5,6}) = 2

1. **Which pairs hash to which buckets?**

Hash function: h(i, j) = (i\*j) mod 11

h(1,2) = (1\*2) mod 11 = 2

h(1,3) = (1\*3) mod 11 = 3

h(1,4) = (1\*4) mod 11 = 4

h(1,5) = (1\*5) mod 11 = 5

h(1,6) = (1\*5) mod 11 = 6

h(2,3) = (2\*3) mod 11 = 6

h(2,4) = (2\*4) mod 11 = 8

h(2,5) = (2\*5) mod 11 = 10

h(2,6) = (2\*6) mod 11 = 1

h(3,4) = (3\*4) mod 11 = 1

h(3,5) = (3\*5) mod 11 = 4

h(3,6) = (3\*6) mod 11 = 7

h(4,5) = (4\*5) mod 11 = 9

h(4,6) = (4\*6) mod 11 = 2

h(5,6) = (5\*6) mod 11 = 8

The first table is the buckets with the values.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|  | 1 | 2 | 3 | 2 | 1 | 0 | 2 | 4 | 3 | 2 |
|  | 4 | 3 |  | 4 |  | 3 |  | 2 |  |  |

The second table is the pairs in each bucket.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | (1,2) | (1,3) | (1,4) | (1,5) | (1,6) |  |  |  |  |
|  | (2,6) |  |  |  |  | (2,3) |  | (2,4) |  | (2,5) |
|  | (3,4) |  |  | (3,5) |  |  | (3,6) |  |  |  |
|  |  | (4,6) |  |  |  |  |  |  | (4,5) |  |
|  |  |  |  |  |  |  |  | (5,6) |  |  |

The third table is the frequency table for each bucket

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Totals |  | 5 | 5 | 3 | 6 | 1 | 3 | 2 | 6 | 3 | 2 |

1. **Which buckets are frequent?**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Buckets | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Totals |  | 5 | 5 | 3 | 6 | 1 | 3 | 2 | 6 | 3 | 2 |

As we can see here, any bucket with a frequency greater than or equal to the threshold, which is 4, is considered frequent. As shown above, the frequent buckets are bucket 1, 2, 4, and 8.

1. **Which pairs are counted on the second pass of the PCY Algorithm?**

This is an index

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |

The pairs in which are counted in the second pass of the PCY Algorithm will be pairs found in buckets containing bit value of 1.

The list of pairs are:

(1,2)

(1,4)

(2,4)

(2,6)

(3,4)

(3,5)

(4,6)

(5,6)

**Question Seven:**

Suppose we run the Multistage Algorithm on the data of Problem 6, with the same support threshold of 4. The first pass is the same as in that problem, and for the second pass, we hash pairs to nine buckets, using the hash function that hashes {i,j} to bucket i+j mod 9.

1. Determine the counts of the buckets on the second pass.

This is the information from the first pass, as we can see the last table expressed the frequency of each bucket and the second table expressed the pairs in each bucket, two very import pieces of information for the second pass.

Table

Description automatically generated

Text

Description automatically generatedFor the second pass we will take the pairs found in the buckets with frequency >=4 and run them into the second pass hash function which is:

Hash function: h(*x*) = (i+j) mod 9

h(*x*) = (2+6) mod 9 = 8

h(*x*) = (3+4) mod 9 = 7

h(*x*) = (1+2) mod 9 = 3

h(*x*) = (4+6) mod 9 = 1

h(*x*) = (1+4) mod 9 = 5

h(*x*) = (3+5) mod 9 = 8

h(*x*) = (2+4) mod 9 = 6

h(*x*) = (5+6) mod 9 = 2

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Buckets | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|  |  | 3 | 2 | 2 |  | 2 | 4 | 4 | 1 |
|  |  |  |  |  |  |  |  |  | 4 |
| Totals |  | 3 | 2 | 2 |  | 2 | 4 | 4 | 4 |
| Pairs: |  | {4,6} | {5,6} | {1,2} |  | {1,4} | {2,4} | {3,4} | {2,6} |
|  |  |  |  |  |  |  |  |  | {3,5} |

The second pass produces a new set of candidate pairs: {{2,4}, {3,4}, {2,6}. {3,5}}

1. Does the second pass reduce the set of candidate pairs? Note that all items are frequent, so the only reason a pair would not be hashed on the second pass is if it hashed to an infrequent bucket on the first pass.

Yes, the second pass reduces the set of candidate pairs. This is because we are using only the candidate pairs from the first pass and will be removing false positives.