Computer Science COMP-4250 - Winter 2021

Assignment 1

Due: End of Sunday, Feb. 7, 2021

****

Note: Submit your assignment as a **single zip file** on Blackboard. The zip file should be named as Assignment1_*StudentId*.zip, in which replace *StudentId* with your university student number.

****

Problem 1. (12 points)

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?

Problem 2. (16 points)

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

(a) The largest integer.
(b) The average of all the integers.
(c) The same set of integers, but with each integer appearing only once.
(d) The count of the number of distinct integers in the input.

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

Problem 3. (18 points)

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:

(a) If the support threshold is ***5***, which items are frequent?
(b) If the support threshold is ***5***, which pairs of items are frequent?
(c) which basket is the largest?
(d) what are the confidence and interest of the following association rules?

$$\{5,7\} \rightarrow 2$$

$$\{2,3,4\} \rightarrow 5$$

(e) Show all the association rules that have ***100%*** confidence for this market-basket data.

Problem 4. (18 points)

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:

(a) If the support threshold is **5**, which items are frequent?
(b) If the support threshold is **5**, which pairs of items are frequent?
(c) which basket is the largest?
(d) what are the confidence and interest of the following association rules?

$$\{24,60\} \rightarrow 8.$$
$$\{2,3,4\} \rightarrow 5.$$

(e) Show all the association rules that have **100**% confidence for this market-basket data.

Problem 5. (10 points)

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.

Problem 6. (16 points)

Here is a collection of twelve baskets. Each contains three of the six items **1** through **6**.

$$\begin{array}{cccc} \{1,2,3\} & \{2,3,4\} & \{3,4,5\} & \{4,5,6\} \\ \{1,3,5\} & \{2,4,6\} & \{1,3,4\} & \{2,4,5\} \\ \{3,5,6\} & \{1,2,4\} & \{2,3,5\} & \{3,4,6\} \end{array}$$

Suppose the support threshold is **4**. On the first pass of the PCY Algorithm we use a hash table with **11** buckets, and the set **{i,j}** is hashed to bucket **i×j** mod **11**.

(a) By any method, compute the support for each item and each pair of items.
(b) Which pairs hash to which buckets?
(c) Which buckets are frequent?
(c) Which pairs are counted on the second pass of the PCY Algorithm?

Problem 7. (10 points)

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**.

(a) Determine the counts of the buckets on the second pass.
(b) 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.