

A = {1,2,3,4} B = {2,3,5,7} C = {2,4,6}

Text

Description automatically generated

|A| = 4 |A intersection B| = 2

|B| = 4 |A intersection C| = 2

|C| = 3 |B intersection C| = 1

1. J(A,B) = 2/(4+4-2) = 0.33
2. J(A,C) = 2/(4+3-2) = 0.4
3. J(B,C) = 1/(4+3-1) = 0.16667



According to the Surprising Property

Text, letter

Description automatically generated

P[ h(A) = h(B) ] = J(A,B)

The value of J(A,B) => 0<= J(A,B) <=1

If J(A,B) = 0

It means that A,B have an empty intersection

Thus, the similarity of these signatures is zero and it must be.

A picture containing chart

Description automatically generated

A picture containing text, clock

Description automatically generated



Shape, circle

Description automatically generated

When the Jacquard Similarity is zero, it is impossible for a min-hash function to return the same value for these two columns.

Thus, the similarity of these signatures is zero and it must be.

Text

Description automatically generated

There are no intersections, so

Sim(C1,C2) = 0/7 = 0

* The probability (over all permutations of the rows) that h(C1) = h(C2) is the same as Sim(C1,C2)

C1 C2

0 0

0 0

0 0

0 0

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated

A spokesperson for the Sudzo Corporation revealed today that studies have shown it is good for people to buy Sudzo products.

A spokesperson for

for the Sudzo

the Sudzo Corporation

that studies have

have shown it

it is good

is good for

for people to

to buy sudzo