#40: Cardinality and Similarity

Brad Solomon

Cardinality

Cardinality is a measure of:

Cardinality Estimation

If I randomly sampled values from 0 – 999 (no repeats) and told you that the minimum value was 300, what is your best estimate for the cardinality in the random set?

What if the minimum value was 20?

K-minimum Estimation

Will the k-th minimum give me a better, worse, or the same estimation accuracy as the minimum? Why?

K-minimum Estimation Equation

Given a range of values m and the k-th minimum value, what equation can be used to estimate the cardinality?

Set Review

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

$A \cup B =$

$A \cap B =$

$A \, / \, B =$

$A \, \Delta \, B =$

Jaccard Similarity

What is the equation for the Jaccard similarity? What is the similarity for the above A and B?

Estimating Similarity

Given the bottom 8 minimum hash values for A and B (below), estimate the similarity of the sets using an approximation of intersection and union.

A={3, 7, 8, 11, 15, 17, 22, 23}
B={2, 3, 6, 7, 9, 11, 17, 23}

Repeat the same calculation, but this time using the inclusion-exclusion principle (also known as 'double counting') to estimate the similarity without using the intersection.