How can I calculate the Jaccard Similarity in R?

Authored by

stats writer

April 21, 2024

The Jaccard Similarity is a metric used to measure the similarity between two sets of data. In order to calculate it in R, you can use the "jaccard" function from the "proxy" package. This function takes two vectors as inputs and returns the Jaccard Similarity coefficient, which ranges from 0 (completely dissimilar) to 1 (identical). It is a useful tool for data analysis and can be easily implemented in R for various applications.

Calculate Jaccard Similarity in R

The measures the similarity between two sets of data. It can range from 0 to 1. The higher the number, the more similar the two sets of data.

The Jaccard similarity index is calculated as:

Jaccard Similarity = (number of observations in both sets) / (number in either set)

Or, written in notation form:

J(A, B) = |A∩B| / |A∪B|

This tutorial explains how to calculate Jaccard Similarity for two sets of data in R.

Example: Jaccard Similarity in R

Suppose we have the following two sets of data:

a <- c(0, 1, 2, 5, 6, 8, 9)b <- c(0, 2, 3, 4, 5, 7, 9)

We can define the following function to calculate the Jaccard Similarity between the two sets:

```
#define Jaccard Similarity function
jaccard <- function(a, b) {
intersection = length(intersect(a, b))
union = length(a) + length(b) - intersection
return (intersection/union)
}

#find Jaccard Similarity between the two sets
jaccard(a, b)
```

0.4

The Jaccard Similarity between the two lists is 0.4.

Note that the function will return 0 if the two sets don't share any values:

```
c <- c(0, 1, 2, 3, 4, 5)
d <- c(6, 7, 8, 9, 10)

jaccard(c, d)
```

0

And the function will return 1 if the two sets are identical:

```
e <- c(0, 1, 2, 3, 4, 5)
f <- c(0, 1, 2, 3, 4, 5)

jaccard(e, f)

1
```

The function also works for sets that contain strings:

```
g <- c('cat', 'dog', 'hippo', 'monkey')
h <- c('monkey', 'rhino', 'ostrich', 'salmon')

jaccard(g, h)

0.142857
```

You can also use this function to find the Jaccard distance between two sets, which is the *dissimilarity* between two sets and is calculated as 1 - Jaccard Similarity.

```
a <- c(0, 1, 2, 5, 6, 8, 9)
```

```
b <- c(0, 2, 3, 4, 5, 7, 9)

#find Jaccard distance between sets a and b
1 - jaccard(a, b)

0.6
```

Refer to this Wikipedia page to learn more details about the Jaccard Similarity Index.