How can I calculate Jaccard Similarity using Python?

Authored by

stats writer

April 16, 2024

Jaccard Similarity is a mathematical measure used to determine the similarity between two sets of data. In order to calculate Jaccard Similarity using Python, you can follow these steps:

1. Import the necessary libraries, such as sklearn or numpy, to perform the calculations.

2. Create two sets of data, represented as arrays or lists, that you want to compare.

3. Use the intersection function to find the common elements between the two sets.

4. Use the union function to find the total elements in both sets.

5. Divide the number of common elements by the total number of elements to obtain the Jaccard Similarity score.

6. You can also use the Jaccard function from the sklearn library to directly calculate the Jaccard Similarity score.

By following these steps, you can easily calculate the Jaccard Similarity between two sets of data using Python. This measure is commonly used in data mining, machine learning, and text analysis to determine the similarity between different sets of data.

Calculate Jaccard Similarity in Python

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 Python.

Example: Jaccard Similarity in Python

Suppose we have the following two sets of data:

import numpy as np

a =
b =

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

#define Jaccard Similarity functiondef jaccard(list1, list2):
intersection = len(list(set(list1).intersection(list2)))
union = (len(list1) + len(list2)) - intersection
return float(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 =
d =

jaccard(c, d)

0.0
```

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

```
e =
f =

jaccard(e, f)

1.0
```

The function also works for sets that contain strings:

```
g =
h =
```

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 =
b =

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

0.6

How to Calculate Jaccard Similarity in R

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