What is the definition of Jaro-Winkler similarity and can you provide an example?

Authored by

stats writer

June 28, 2024

Jaro-Winkler similarity is a metric used to measure the similarity between two strings. It takes into account the number of matching characters and the order in which they appear in the strings. The resulting similarity score ranges from 0 (no similarity) to 1 (perfect similarity).

For example, let's consider the strings "cat" and "hat". The Jaro-Winkler similarity between these two strings is 0.917, indicating a high level of similarity due to the matching characters "a" and "t" and their order in the strings. On the other hand, the strings "dog" and "horse" have a Jaro-Winkler similarity of 0.0, indicating no similarity due to the lack of matching characters and different order of characters. Overall, the Jaro-Winkler similarity provides a quantitative measure of how similar two strings are, which can be useful in various applications such as record linkage and spell checking.

An Introduction to Jaro-Winkler Similarity (Definition & Example)

In statistics, the Jaro-Winkler similarity is a way to measure the similarity between two strings.

The Jaro similarity (simj) between two strings is defined as:

simj = 1/3 * (m /|s1| + m/|s2| + (m-t)/m)

where:

m: Number of matching charactersTwo characters from s1 and s2 are considered matching if they are the same and not farther than - 1 characters apart.|s1|, |s2|: The length of the first and second strings, respectivelyt: Number of transpositionsCalculated as the number of

matching (but different sequence order) characters divided by 2.

The Jaro-Winkler similarity (simw) is defined as:

simw = simj + lp(1 - simj)

where:

simj: The Jaro similarity between two strings, s1 and s2l: Length of the common prefix at the start of the string (max of 4 characters)p: Scaling factor for how much the score is adjusted upwards for having common prefixes. Typically this is defined as p = 0.1 and should not exceed p = 0.25.

The Jaro-Winkler similarity between two strings is always between 0 and 1 where:

0 indicates no similarity between the strings1 indicates that the strings are an exact match

Note: The Jaro-Winkler *distance* would be defined as 1 - simw.

The following example shows how to calculate the Jaro-

Winkler similarity between two strings in practice.

Example: Calculating Jaro-Winkler Similarity Between Two Strings

Suppose we have the following two strings:

String 1 (s1): mouseString 2 (s2): mute

First, let's calculate the Jaro Similarity between these two strings:

where:

m: Number of matching charactersTwo characters from s1 and s2 are considered matching if they are the same and not farther than - 1 characters apart.

In this case, - 1 is calculated as 5/2 - 1 = 1.5. We would define three letters as matching: m, u, e. Thus, m = 3.

|s1|, |s2|: The length of the first and second strings, respectively

In this case, |s1| = 5 and |s1| = 4.

t: Number of transpositionsCalculated as the number of matching (but different sequence order) characters

divided by 2.

In this case, there are three matching characters but they're already in the same sequence order, so t = 0.

Thus, we would calculate the Jaro Similarity as:

simj = 1/3 * (3/5 + 3/4 + (3-0)/3) = 0.78333.

Next, let's calculate the Jaro-Winkler similarity (simw) as:

simw = simj + lp(1 - simj)

In this case, we would calculate:

simw = 0.78333 + (1)*(0.1)(1 - 0.78333) = 0.805.

The Jaro-Winkler similarity between the two strings is 0.805.

Since this value is close to 1, it tells us that the two strings are very similar.

We can confirm this is correct by calculating the Jaro-Winkler similarity between the two strings in R:

```
library(stringdist)

#calculate Jaro-Winkler similarity between 'mouse' and
'mute'
1 - stringdist("mouse", "mute", method = "jw", p=0.1)

0.805
```

This matches the value that we calculated by hand.

Additional Resources

The following tutorials explain how to calculate other similarity metrics: