WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
Deutsch
Русский
Српски / srpski
Українська
Edit links

String metric

From Wikipedia, the free encyclopedia
(Redirected from String metrics)

"String distance" redirects here. For the distance between strings and the fingerboard in musical instruments, see Action (music).

In mathematics and computer science, a **string metric** (also known as a **string similarity metric** or **string distance function**) is a metric that measures distance ("inverse similarity") between two text strings for approximate string matching or comparison and in fuzzy string searching. Necessary requirement for a string *metric* (e.g. in contrast to string matching) is fulfillment of the triangle inequality. For example the strings "Sam" and "Samuel" can be considered to be close. A string metric provides a number indicating an algorithm-specific indication of distance.

The most widely known string metric is a rudimentary one called the Levenshtein Distance (also known as Edit Distance). It operates between two input strings, returning a number equivalent to the number of substitutions and deletions needed in order to transform one input string into another. Simplistic string metrics such as Levenshtein distance have expanded to include phonetic, token, grammatical and character-based methods of statistical comparisons.

A widespread example of a string metric is DNA sequence analysis and RNA analysis, which are performed by optimized string metrics to identify matching sequences.

String metrics are used heavily in information integration and are currently used in areas including fraud detection, fingerprint analysis, plagiarism detection, ontology merging, DNA analysis, RNA analysis, image analysis, evidence-based machine learning, database data deduplication, data mining, Web interfaces, e.g. Ajax-style suggestions as you type, data integration, and semantic knowledge integration.

Contents [hide]

List of string metrics [edit]

- Sørensen–Dice coefficient
- Hamming distance
- Levenshtein distance and Damerau–Levenshtein distance
- Block distance or L1 distance or City block distance
- Simple matching coefficient (SMC)
- Jaccard similarity or Jaccard coefficient or Tanimoto coefficient
- Most frequent k characters
- Tversky index
- Overlap coefficient
- Variational distance
- Hellinger distance or Bhattacharyya distance
- Information radius (Jensen–Shannon divergence)
- Skew divergence
- Confusion probability
- Tau metric, an approximation of the Kullback–Leibler divergence
- Fellegi and Sunters metric (SFS)
- Maximal matches
- Lee distance
- Jaro–Winkler distance

Selected string measures examples [edit]

Name	Example
Hamming distance	"karolin" and "kathrin" is 3.
Levenshtein distance and Damerau–Levenshtein distance	1. **k**itten → **s**itten (substitution of "s" for "k") 2. sitt**e**n → sitt**i**n (substitution of "i" for "e") 3. sittin → sittin**g** (insertion of "g" at the end).
Most frequent k characters	MostFreqKeySimilarity('research', 'seeking', 2) = 2

See also [edit]

- edit distance
- approximate string matching
- String matching
- Carnegie Mellon University open source library
- StringMetric project a Scala library of string metrics and phonetic algorithms
- Natural project a JavaScript natural language processing library which includes implementations of popular string metrics

External links [edit]

- http://www.dcs.shef.ac.uk/~sam/stringmetrics.html [*dead link*] A fairly complete overview Archive copy at the Wayback Machine

Categories: String similarity measures | Metrics