TOOL

Similarity and Distance Metrics

Similarity Metrics	
Metric	**Description**
Jaccard Similarity	Produces a score in the interval [0,1], indicating the degree of similarity of two sets. Zero represents no common elements, and 1 represents a perfect similarity. Jaccard similarity can be calculated for any two sets (or numbers, characters, words, objects, etc.). Jaccard similarity is $\|I\|/\|U\|$, or the fraction of the sets' cardinalities, where I is the intersection set of distinct shared elements and U is the union set of all distinct elements. That is, you compute the fraction of overlap of two sets. Before you apply Jaccard similarity to two sequences of strings, these must be converted to sets of tokens (whether characters or words, or phrases, or sentences, etc).
Binary Similarity	Produces false (or 0 integer) when compared objects differ or true (or 1) when the objects are equivalent (i.e., equal in some sense). An example below demonstrates the equivalency built into Python. If the two string objects have different characters (including capitalization) or order, they are considered unequal. Notice that equivalence needs to be defined in advance. Typically, character strings are considered to be equivalent whenever they have exactly the same sequence of characters, but it may be convenient to allow capitalization, misspellings, punctuation and other modifications, which do not affect the semantic meaning. For example, Python treats 'Cornell', 'cornell', 'Cornel' as different strings of characters, but our domain of expertise suggests that these should all be treated as equal in some documents. Python, however, understands arithmetic representations of the same quantity and, for example, treats 0, 1-1, 4/2-2, 2+2-4 as equal.

Similarity Metrics	
Metric	**Description**
Correlation	A measure of linear relation with values in the interval $[-1,1]$. Correlation can only be computed on two same size numeric vectors. • A value of 1 indicates a perfectly linear relation, such as for vectors $x=[1,2,3]$ and $y=[1,3,5]$, where you can express one vector as a linear combination of another: $y=2x-1$ • A value of 0 indicates no linear relation between the two numeric vectors. Example: $x=[1,2,3]$ and $y=[2,2,2]$, where $y=0 \cdot x+2$, i.e. while elements of x increase with index, elements of y remain constant. • A value of -1 indicates a perfectly negative linear relation, such as for vectors $x=[1,2,3]$ and $y=[3,2,1]$, where you can express the relation between corresponding elements via a linear formula. I.e. $y=-x+1$

Distance Metrics	
Metric	**Description**
Hamming Distance	This metric is opposite to similarity in that it counts the number of disagreements in elements of corresponding positions of two sequences of equal length. Just like Jaccard similarity, Hamming distance works on elements of any data type (not just numbers). Just like correlation, it requires sequences (not just sets) of the same length. If sequence lengths mismatch, you can return an infinity or **np.inf** non-numeric value.
Levenshtein Distance	*Also referred to as edit distance.* A metric which counts the number of edits needed to convert one sequence to another (add, delete, replace). The sequence can be a word (sequence of characters), sentence, binary code (sequence of zeros and ones), DNA (sequence of nucleotides A,C,G,T), etc.