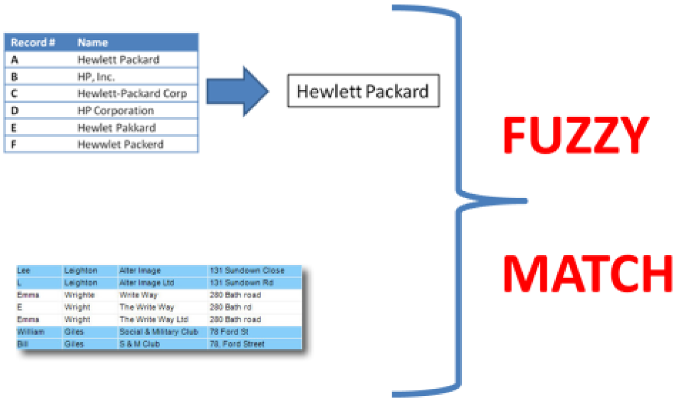
Fuzzy Matching Algorithms To Help Data Scientists Match Similar Data

- Posted by Megter on January 20, 2016 at 2:25pm
- Send Message View Blog



Record #	Name
A	Hewlett Packard
B	HP, Inc.
C	Hewlett-Packard Corp
D	HP Corporation
E	Hewlet Pakkard
F	Hewwlet Packerd

Lee	Leighton	Alter Image	131 Sundown Close
L	Leighton	Alter Image Ltd	131 Sundown Rd
Emma	Wrighte	Write Way	280 Bath road
E	Wright	The Write Way	280 Bath rd
Emma	Wright	The Write Way Ltd	280 Bath road
William	Giles	Social & Military Club	78 Ford St
Bill	Giles	S & M Club	78, Ford Street

Fuzzy Match Graphic by Megter.com

A common scenario for data scientists is the marketing, operations or business groups give you two sets of similar data with different variables & asks the analytics team to normalize both data sets to have a common record for modelling.

Here is an example of two similar data sets:

Data Set 1		**Data Set 2**	
Organization Name	Sales	Organization Name	# of Customers
John Doe Inc	$300	Sally Harper Cntr	10
Saint Rogers	$400	John Doe Incorporated	50
Sally Harper Center	$500	St. Rogers	100

How would you as a data scientist match these two different but similar data sets to have a master record for modelling?

Short of doing it manually, the most common method is fuzzy matching.

So, what is Fuzzy matching? Here is a short description from Wikipedia:

Fuzzy matching is a technique used in computer-assisted translation as a special case of record linkage. It works with matches that may be less than 100% perfect when finding correspondences between segments of a text and entries in a database of previous translations. It usually operates at sentence-level segments, but some translation technology allows matching at a phrasal level. It is used when the translator is working with translation memory.

Given below is list of algorithms to implement fuzzy matching algorithms which themselves are available in many open source libraries:

Levenshtein distance Algorithm

Levenshtein distance is a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other.

Damerau–Levenshtein distance

Damerau–Levenshtein distance is a distance (string metric) between two strings, i.e., finite sequence of symbols, given by counting the minimum number of operations needed to transform one string into the other, where an operation is defined as an insertion, deletion, or substitution of a single character, or a transposition of two adjacent characters.

Bitap algorithm with modifications by Wu and Manber

Bitmap algorithm is an approximate string matching algorithm. The algorithm tells whether a given text contains a substring which is "approximately equal" to a given pattern, where approximate equality is defined in terms of Levenshtein distance — if the substring and pattern are within a given distance k of each other, then the algorithm considers them equal.

n-gram

n-gram is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a (n − 1)–order Markov model.

BK-tree

A BK-tree is a metric tree suggested by Walter Austin Burkhard and Robert M. Keller specifically adapted to discrete metric spaces.To understand, let us consider integer discrete metric d(x,y). Then, BK-tree is defined in the following way. An arbitrary element a is selected as root node. The root node may have zero or more subtrees. The k-th subtree is recursively built of all elements b such that d(a,b) = k. BK-trees can be used for approximate string matching in a dictionary

Soundex

Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English. The goal is for homophones to be encoded to the same representation so that they can be matched despite minor differences in spelling.

Views: 31243

Like
7 members like this

Share Tweet G+ Facebook

Like 0

- < Previous Post
- Next Post >

Comment

Visual Mode HTML Editor