csvmatch

Fuzzy matching text with Python

Dan Keemahill, he/him
USA TODAY data team
dkeemahill@gannett.com

Links!

csvmatch – a Python fuzzy-matching library by Max Harlow

fuzzy_pandas – an implementation of csvmatch for Pandas by Jonathan Soma

fuzzywuzzy – a Python fuzzy-matching library by Seatgeek developers

OpenRefine – an application for managing data

Fuzzy matching

"matches that may be less than 100% perfect" but "refer to the same thing"

San Jose = San José

Keemahill, Daniel = Dan Keemahill

Catherine Osterman = Cat Osterman

Fuzzy matching best practices

Verify matches

Keep subject matter and context in mind (are data dirty by design?)

Root out false positivites, watch for false negatives

There are many algorithms, but simple can go a long way

Larger datasets generally take more time to match

Common text matching gotchas

Upper case and lower case

Non-alphanumeric (punctuation, whitespace, etc)

Non-latin characters (é, å, ß, etc)

The order the words are in

The order the *letters* are in

Common name prefixes (Mr, Ms, etc)

A custom list of words (LLC, University, "City of")

csvkit can ignore common gotchas

Specify arguments on the command line

Common text matching gotchas

	`csvkit arguments`	
Upper case and lower case	`-i`	`--ignore-case`
Non-alphanumeric (punctuation, whitespace, etc)	`-a`	`--ignore-nonalpha`
Non-latin characters (é, å, ß, etc)	`-n`	`--ignore-nonlatin`
The order the words are in	`-s`	`--ignore-order-words`
The order the *letters* are in	`-e`	`--ignore-order-letters`
Common name prefixes (Mr, Ms, etc)	`-t`	`--ignore-titles`
A custom list of words (LLC, University, "City of")	`-l`	`--ignore-custom`

csvkit uses fuzzy-matching algorithms

Levenshtein – percentage of characters in common (a type of "edit distance")

Metaphone – English pronunciation

Bilenko – interactively train a machile learning model with dedupe

Demo: csvkit

This Google Colab notebook demonstrates csvkit algorithms with datasets of the names of world leaders collected by Max Harlow.

Example: PPP loans and WARN notices

Center for Public Integrity used its own algorithm to identify companies that received federal money and laid off workers.

Match companies where:

- states are the same

- the first 18 non-whitespace, alphanumeric characters the same

- company filed a WARN notice within six months of receiving a PPP loan

Example: PPP loans and WARN notices

CPI manually verified one thousand matches.

Result was still likely an undercount due to false negatives (companies operating in multiple states, differences in the first 18 characters).

Always be learning!

Join USA TODAY Network Data in Microsoft Teams

Visit training.usatodaynetwork.com