
fuzzymatcher Documentation

Release 0.1

Robin Linacre

Aug 09, 2022

Contents:

1	fuzzymatcher	3
1.1	Installation	3
1.2	Usage	3
1.3	Simple example	3
2	Another title goes here	5

Note: fuzzymatcher is no longer actively maintained. Please see [splink](#) for a more accurate, scalable and performant solution

A Python package that allows the user to fuzzy match two pandas dataframes based on one or more common fields.

Fuzzymatches uses `sqlite3`'s Full Text Search to find potential matches.

It then uses [probabilistic record linkage](#) to score matches.

Finally it outputs a list of the matches it has found and associated score.

1.1 Installation

```
pip install fuzzymatcher
```

Note that you will need a build of `sqlite` which includes FTS4. This seems to be widely included by default, but otherwise [see here](#).

1.2 Usage

See [examples.ipynb](#) for examples of usage and the output.

You can run these examples interactively [here](#).

1.3 Simple example

Suppose you have a table called `df_left` which looks like this:

id	ons_name
0	Darlington
1	Monmouthshire
2	Havering
3	Knowsley
4	Charnwood
...	etc.

And you want to link it to a table `df_right` that looks like this:

id	os_name
0	Darlington (B)
1	Havering London Boro
2	Sir Fynwy - Monmouthshire
3	Knowsley District (B)
4	Charnwood District (B)
...	etc.

You can write:

```
import fuzzymatcher
fuzzymatcher.fuzzy_left_join(df_left, df_right, left_on = "ons_name", right_on = "os_
↪name")
```

And you'll get:

best_match_score	ons_name	os_name
0.178449	Darlington	Darlington (B)
0.133371	Monmouthshire	Sir Fynwy - Monmouthshire
0.102473	Havering	Havering London Boro
0.155775	Knowsley	Knowsley District (B)
0.155775	Charnwood	Charnwood District (B)
...	etc.	etc.

CHAPTER 2

Another title goes here

- `genindex`
- `modindex`
- `search`