






Jonathan Soma (/)

 (<http://twitter.com/dangerscarf>)  (<https://instagram.com/dangerscarf>)  (<https://github.com/jsoma>) 
(<https://tinyletter.com/jsoma>)  (<mailto:jonathan.soma@gmail.com>)

[Home \(/\)](#) > [Lede \(.././././.\)](#) > [Algorithms, Lede 2017 \(../././.\)](#) > [Fuzzing matching in pandas with fuzzywuzzy](#)

Fuzzing matching in pandas with fuzzywuzzy

This page is based on a Jupyter/IPython Notebook: download the original .ipynb (</lede/algorithms-2017/classes/fuzziness-matplotlib/fuzzing-matching-in-pandas-with-fuzzywuzzy.ipynb>)

Sometimes you don't want to use OpenRefine. Why not? I don't know, it's the *best* for cleaning up fuzzy matches. But yes, sure, sometimes maybe you don't.

```
%matplotlib inline
import pandas as pd
```

```
df = pd.read_csv("CD_Transactions_07-23-2017.CSV", index_col=False)
```

```
df.head()
```

	Result	Date	Transaction Type	Payment Type	Payment Detail	Amount	Last/Business Name	First Name	Address	City	...	----- --	Report Type
--	--------	------	------------------	--------------	----------------	--------	--------------------	------------	---------	------	-----	-------------	-------------

	Result	Date	Transaction Type	Payment Type	Payment Detail	Amount	Last/Business Name	First Name	Address	City	...	----- --	Report Type	
0	1	4/3/2017	Income	Check	12864	\$350.00	Alaska Republican Party State Account	NaN	NaN	NaN	...	NaN	24 Hour Report	20 An Mt Elk
1	2	4/3/2017	Income	Credit Card	NaN	\$500.00	Coffey	Dan	NaN	NaN	...	NaN	24 Hour Report	20 An Mt Elk
2	3	4/3/2017	Income	Check	3047	\$300.00	ACS Employees PAC	NaN	NaN	NaN	...	NaN	24 Hour Report	20 An Mt Elk
3	4	4/3/2017	Income	Credit Card	NaN	\$500.00	Holmes	Patrick	NaN	NaN	...	NaN	24 Hour Report	20 An Mt Elk
4	5	4/3/2017	Income	Credit Card	NaN	\$500.00	Gonzales	Mark L.	NaN	NaN	...	NaN	24 Hour Report	20 An Mt Elk

5 rows x 26 columns

What are all of our options for the “Alaska Sea Pilot PAC fund”?

```
df[df['Last/Business Name'] == 'Alaska Sea Pilot PAC fund'].shape
```

(3, 26)

```
df[df['Last/Business Name'] == 'ALASKA SEA PILOT PAC FUND'].shape
```

(6, 26)

```
df[df['Last/Business Name'] == 'Alaska Sea Pilot Pac Fund'].shape
```

(6, 26)

Maybe we can throw in a regex and catch some more?

```
df[df['Last/Business Name'].str.contains("Sea.*Pilot", na=False)]['Last/Business Name'].value_counts()
```

Alaska Sea Pilot PAC Fund	46
Alaska Sea Pilot PAC	17
Alaska Sea Pilots PAC Fund	13
Alaska Sea Pilot Pac Fund	6
Alaska Sea Pilots	5
Alaska Sea Pilot PAC fund	3
Alaska Sea Pilot Pac	3
Alaska Sea Pilot	3
Alaska Sea Pilots Pac Fund	2
Alaska Sea Pilot PAC Fund	2
Alaska Sea Pilot Fund	2
Alaska Sea Pilot PAC fund	2
Alaska Sea Pilot PAC	1
Alaska Sea Pilots, 1621 Tongass Ave., Ketchikan, AK. 99901	1
Ak Sea Pilot PAC	1
AlaskanSea Pilot PAC	1
Alaska SeaPilot PAC	1
Alaska Sea Pilots Ass'n	1
AK Sea Pilot Pac Fund	1
AK Sea Pilot PAC	1
Alaska Sea Pilots, LLC.	1
Alaska Sea Pilot	1
Alaska Sea Pilot Pac	1
AK Sea Pilot PAC Fund	1
Name: Last/Business Name, dtype: int64	

Using fuzzywuzzy for finding fuzzy matches

Fuzzy matches are incomplete or inexact matches. The Python package fuzzywuzzy (<https://github.com/seatgeek/fuzzywuzzy>) has a few functions that can help you, although they're a little bit confusing! I'm going to take the examples from GitHub and annotate them a little, then we'll use them.

First, install fuzzywuzzy with

```
pip3 install fuzzywuzzy[speedup]
```

Then we'll get to importing it

```
# fuzz is used to compare TWO strings
from fuzzywuzzy import fuzz

# process is used to compare a string to MULTIPLE other strings
from fuzzywuzzy import process
```

MAKE SURE YOU INSTALLED USING `pip3 install fuzzywuzzy[speedup]` OR ELSE IT WILL COMPLAIN HERE AND WILL ALSO BE SLOWER

`fuzz.ratio` compares the entire string, in order

Every single thing in the string is important here!

```
fuzz.ratio("this is a test", "this is a fun")
```

74

`fuzz.partial_ratio` compares subsections of the string

Partial matches are fine! The exclamation mark at the end made `fuzz.ratio` not like the comparison last time, but this time it's OK.

```
fuzz.partial_ratio("this is a test", "test a is this")
```

57

fuzz.token_sort_ratio ignores word order

fuzz.token_sort_ratio orders all of the words first, so “KENNEDY JOHN” and “JOHN KENNEDY” would be the same.

```
fuzz.token_sort_ratio("fuzzy wuzzy was a bear", "wuzzy fuzzy was a bear")
```

100

```
fuzz.token_sort_ratio("this is a test", "is this a test")
```

100

```
fuzz.token_sort_ratio("fuzzy was a bear", "fuzzy fuzzy was a bear")
```

84

fuzz.token_set_ratio ignores duplicate words

I don't know why you'd ever have “JOHN KENNEDY KENNEDY” but if you use fuzz.token_set_ratio then it would definitely match “JOHN KENNEDY”.

```
fuzz.token_set_ratio("fuzzy was a bear", "fuzzy fuzzy was a bear")
```

```
100
```

Actually using fuzzywuzzy on our dataset, featuring `process.extract`

```
choices = ['fuzzy fuzzy was a bear', 'is this a test', 'THIS IS A TEST!!!']  
process.extract("this is a test", choices, scorer=fuzz.ratio)
```

```
[('THIS IS A TEST!!!', 100),  
 ('is this a test', 86),  
 ('fuzzy fuzzy was a bear', 33)]
```

```
choices = ['fuzzy fuzzy was a bear', 'is this a test', 'THIS IS A TEST!!!']  
process.extract("this is a test", choices, scorer=fuzz.token_sort_ratio)
```

```
[('is this a test', 100),  
 ('THIS IS A TEST!!!', 100),  
 ('fuzzy fuzzy was a bear', 28)]
```

Since we already imported, let's collect all of the business names into a list. We're going to search through the list to find names that are similar to **Alaska Sea Pilot PAC Fun**.

```
# If we grab a column and use .unique(), it gives us every business name with no repeats
choices = df['Last/Business Name'].unique()
choices[:15]
```

```
array(['Alaska Republican Party State Account', 'Coffey',
      'ACS Employees PAC', 'Holmes', 'Gonzales',
      'Anchorage Taxicab Permit Owners Association (ATPOA)', 'Abdullah',
      'Alimi', 'Barbosa', 'Bryant', 'Chamot', 'Farmer', 'Gautam',
      'Guevara', 'Lena'], dtype=object)
```

Now we'll use `process.extract` to find the top 15 matches

```
%%time
process.extract("Alaska Sea Pilot PAC Fund", choices, limit=30, scorer=fuzz.token_sort_ratio)
```


CPU times: user 634 ms, sys: 5.38 ms, total: 639 ms
Wall time: 642 ms

```
[('Alaska Sea Pilot PAC Fund', 100),  
 ('Alaska Sea Pilot PAC fund', 100),  
 ('ALASKA SEA PILOT PAC FUND', 100),  
 ('Alaska Sea Pilot PAC Fund ', 100),  
 ('Alaska SEA Pilot Pac Fund', 100),  
 ('Alaska SEA Pilot PAC Fund', 100),  
 ('Alaska Sea Pilot Pac Fund', 100),  
 ('Alaska Sea Pilot PAC fund', 100),  
 ('Alaska Sea Pilots PAC Fund', 98),  
 ('Alaska Sea Pilots Pac Fund', 98),  
 ('Alaska Sea Pilot Fund', 91),  
 ('AK Sea Pilot Pac Fund', 91),  
 ('ALASKA SEA PILOT FUND', 91),  
 ('AK Sea Pilot PAC Fund', 91),  
 ('Alaska Sea Pilot Pac', 89),  
 ('Alaska Sea Pilot PAC', 89),  
 ('Alaska Sea Pilot Pac ', 89),  
 ('Alaska Sea Pilot PAC ', 89),  
 ('ALASK SEA PILOT PAC', 86),  
 ('Alaska Sea Pilot', 78),  
 ('AK Sea Pilot PAC', 78),  
 (' Ak Sea Pilot PAC', 78),  
 ('Alaska Sea Pilot ', 78),  
 ('Alaska Sea Pilots, LLC.', 78),  
 ('Alaska Sea Pilots', 76),  
 ('AlaskanSea Pilot PAC', 76),  
 ('Alaska Marine Pilot PAC', 75),  
 ("Alaska Sea Pilots Ass'n", 75),  
 ('Alaska SeaPilot PAC', 73),  
 ('Alaska Senate Majority Fund', 73)]
```

Wow, those look pretty nice! Maybe instead we should just find the ones that are above a certain score? You can also specify a `scorer` if you want to get particular.

```
# Get 100 options so we're sure to have some non-matches
possibilities = process.extract("Alaska Sea Pilot PAC Fund", choices, limit=100, scorer=fuzz.token_sort_ratio)
```

```
# And let's see everything with a score above 73
[possible for possible in possibilities if possible[1] > 73]
```

```
[('Alaska Sea Pilot PAC Fund', 100),
 ('Alaska Sea Pilot PAC fund', 100),
 ('ALASKA SEA PILOT PAC FUND', 100),
 ('Alaska Sea Pilot PAC Fund ', 100),
 ('Alaska SEA Pilot Pac Fund', 100),
 ('Alaska SEA Pilot PAC Fund', 100),
 ('Alaska Sea Pilot Pac Fund', 100),
 ('Alaska Sea Pilot PAC fund', 100),
 ('Alaska Sea Pilots PAC Fund', 98),
 ('Alaska Sea Pilots Pac Fund', 98),
 ('Alaska Sea Pilot Fund', 91),
 ('AK Sea Pilot Pac Fund', 91),
 ('ALASKA SEA PILOT FUND', 91),
 ('AK Sea Pilot PAC Fund', 91),
 ('Alaska Sea Pilot Pac', 89),
 ('Alaska Sea Pilot PAC', 89),
 ('Alaska Sea Pilot Pac ', 89),
 ('Alaska Sea Pilot PAC ', 89),
 ('ALASK SEA PILOT PAC', 86),
 ('Alaska Sea Pilot', 78),
 ('AK Sea Pilot PAC', 78),
 (' Ak Sea Pilot PAC', 78),
 ('Alaska Sea Pilot ', 78),
 ('Alaska Sea Pilots, LLC.', 78),
 ('Alaska Sea Pilots', 76),
 ('AlaskanSea Pilot PAC', 76),
 ('Alaska Marine Pilot PAC', 75),
 ("Alaska Sea Pilots Ass'n", 75)]
```

Huh, pretty neat.

Filtering directly with fuzzywuzzy

We can also use this directly with our dataframe, if we'd like to use `fuzzywuzzy` to filter instead of giving us a list. It's going to be a lot slower, but that's life, I guess.

```
def get_ratio(row):
    name = row['Last/Business Name']
    return fuzz.token_sort_ratio(name, "Alaska Sea Pilot PAC Fund")

df[df.apply(get_ratio, axis=1) > 70]
```

	Result	Date	Transaction Type	Payment Type	Payment Detail	Amount	Last/Business Name	First Name	Address	City
50505	50506	12/3/2015	Income	Check	1278	\$500.00	Alaska Sea Pilot PAC Fund	NaN	1621 Tongass Ave, Ste 300	Ketchikan
51557	51558	11/23/2015	Income	Check	1279	\$500.00	Alaska Sea Pilot Pac	NaN	1621 Tongass Ave. Ste. 300	Ketchikan
61585	61586	11/21/2015	Income	Check	1090	\$500.00	Alaska Sea Pilots PAC Fund	NaN	1621 Tongass Avenue Ste. 300	Ketchikan
62351	62352	11/16/2015	Income	Check	1084	\$500.00	Alaska Sea Pilot PAC	NaN	1621 Tongass Ave	Ketchikan
74155	74156	11/16/2015	Income	Check	1087	\$500.00	Alaska Sea Pilot PAC Fund	NaN	1621 Tongass Ave	Ketchikan
75904	75905	9/13/2016	Income	Check	1289	\$1,000.00	Alaska Sea Pilot PAC Fund	Alaska Sea Pilot PAC Fund	1621 Tongass Avenue	Ketchikan

	Result	Date	Transaction Type	Payment Type	Payment Detail	Amount	Last/Business Name	First Name	Address	City
77090	77091	9/13/2016	Income	Check	1285	\$500.00	Alaska Sea Pilot	NaN	1621 Tongass Avenue, Suite 300	Ketchikan
77110	77111	10/24/2016	Income	Check	1297	\$1,000.00	Alaska Sea Pilot PAC Fund	NaN	1621 Tongass Ave.	Ketchikan
77259	77260	9/15/2016	Income	Check	1288	\$500.00	ALASK SEA PILOT PAC	NaN	1621 TONGASS AVE	KETCHIKAN
77473	77474	9/14/2016	Income	Check	1284	\$500.00	Alaska Sea Pilot Fund	n/a	1621 Tongass Avenue	Ketchikan
83512	83513	10/18/2016	Income	Check	1206	\$500.00	Alaska Sea Pilots PAC Fund	NaN	1621 Tongass Ave	Ketchikan
83824	83825	10/20/2016	Income	Check	1300	\$1,000.00	Alaska Sea Pilot Pac	NaN	1621 Tongass Ave. Ste. 300	Ketchikan
89576	89577	11/21/2015	Income	Check	1090	\$500.00	Alaska Sea Pilots PAC Fund	NaN	1621 Tongass Avenue Ste. 300	Ketchikan
91019	91020	10/18/2016	Income	Check	1202	\$1,000.00	Alaska Sea Pilots Pac Fund	NaN	1621 Tongass	Ketchikan

	Result	Date	Transaction Type	Payment Type	Payment Detail	Amount	Last/Business Name	First Name	Address	City
91031	91032	10/18/2016	Income	Check	1299	\$500.00	Alaska Sea Pilot PAC fund	NaN	1621 Tongass Ave	Ketchikan
91556	91557	10/26/2016	Income	Check	1208	\$1,000.00	Alaska Sea Pilot PAC Fund	NaN	1621 Tongass Avenue, Suite 300	Ketchikan
92062	92063	9/13/2016	Income	Check	1286	\$1,000.00	Alaska Sea Pilot PAC Fund	NaN	1621 Tongrass Ave	Ketchikan
92508	92509	12/21/2016	Income	Check	1212	\$1,000.00	Alaska Sea Pilots PAC Fund	NaN	1621 Tongass Ave Ste 300	Ketchikan
93756	93757	9/13/2016	Income	Check	1291	\$500.00	ALASKA SEA PILOT PAC FUND	NaN	1621 Tongass ave.	Ketchikan
93896	93897	9/13/2016	Income	Check	1287	\$500.00	Alaska Sea Pilot PAC Fund	NaN	1621 Tongass Ave	Ketchikan
95996	95997	11/25/2015	Income	Check	1277	\$500.00	Alaska Sea Pilot PAC Fund	NaN	1621 Tongass Ave.	Ketchikan
96262	96263	9/19/2016	Income	Check	1201	\$1,000.00	AK Sea Pilot PAC	NaN	1621 Tongass Ave	Ketchikan

[illegible]

	Result	Date	Transaction Type	Payment Type	Payment Detail	Amount	Last/Business Name	First Name	Address	City
366646	366647	12/8/2014	Income	Check	1076	\$1,000.00	Alaska Sea Pilots Ass'n	NaN	1621 Tongass Avenue	Ketchikan
371350	371351	12/18/2013	Income	Check	1198	\$1,000.00	Alaska Sea Pilots PAC Fund	NaN	1621 Tongass Avenue	Ketchikan
371450	371451	10/3/2014	Income	Check	1324	\$1,000.00	Alaska Sea Pilot PAC	NaN	1621 Tongass Ave.	Ketchikan
375262	375263	10/6/2014	Income	Check	1314	\$500.00	Alaska Sea Pilot PAC Fund	NaN	1621 Tongass ave	Ketchikan
380165	380166	11/3/2014	Income	Check	1068	\$500.00	Alaska Sea Pilot PAC	PAC	NaN	NaN
380309	380310	11/3/2014	Income	Check	1070	\$1,000.00	Alaska Sea Pilot PAC Fund	NaN	NaN	NaN
380817	380818	12/14/2013	Income	Check	128	\$500.00	Alaska Sea Pilot PAC	NaN	1621 Tongass Fund, Suite 300	Ketchikan
382147	382148	10/9/2014	Income	Check	1304	\$500.00	ALASKA SEA PILOT PAC FUND	NaN	1621 TONGASS AVE	KETCHIKAN

	Result	Date	Transaction Type	Payment Type	Payment Detail	Amount	Last/Business Name	First Name	Address	City
382399	382400	12/5/2014	Income	Check	1079	\$500.00	Alaska Sea Pilot PAC Fund	NaN	1621 Tongass Ave	Ketchikan
383646	383647	11/5/2014	Income	Check	1072	\$1,000.00	AK Sea Pilot PAC Fund	NaN	1621 Tongass Ave. Suite 300	Ketchikan
384035	384036	8/7/2014	Income	Check	1301	\$1,000.00	Alaska Sea Pilot PAC	NaN	1621 Tongass Ave	Ketchikan
384493	384494	10/6/2014	Income	Check	1307	\$500.00	ALASKA SEA PILOT PAC FUND	NaN	1621 TONGASS AVENUE	KETCHIKAN
384629	384630	10/4/2014	Income	Check	1318	\$500.00	Alaska Sea Pilot PAC Fund	NaN	1621 Tongass Ave	Ketchikan
385218	385219	11/3/2014	Income	Check	1068	\$500.00	Alaska Sea Pilot PAC Fund	NaN	1621 Tongass Avenue	Ketchikan
385656	385657	11/7/2014	Income	Check	1075	\$1,000.00	Alaska Sea Pilot PAC	NaN	1621 Tongass Ave, Suit 300	Ketchikan
387266	387267	10/3/2014	Income	Check	1321	\$500.00	Alaska Sea Pilot PAC	NaN	1621 Tongass Avenue	Ketchikan

	Result	Date	Transaction Type	Payment Type	Payment Detail	Amount	Last/Business Name	First Name	Address	City
387476	387477	10/3/2014	Income	Check	1320	\$500.00	Alaska Sea Pilots PAC Fund	NaN	1621 Tongass Avenue Ste. 300	Ketchikan
387715	387716	11/11/2014	Income	Check	1074	\$1,000.00	Alaska Sea Pilot	NaN	1621 Tongass Ave	Ketchikan
387783	387784	10/6/2014	Income	Check	1306	\$500.00	Alaska Sea Pilots PAC Fund	NaN	1621 Tongass Ave ste 300	Ketchikan
388013	388014	10/8/2014	Income	Check	1313	\$1,000.00	Alaska Sea Pilots PAC Fund	NaN	1621 Tongass Ave	Ketchikan
389046	389047	12/9/2014	Income	Check	1077	\$500.00	Alaska Sea Pilot Pac Fund	NaN	1621 Tongass Ave	Ketchikan
389085	389086	10/11/2014	Income	Check	1312	\$500.00	Alaska Sea Pilot PAC	NaN	1621 Tongass Ave	Ketchikan
389117	389118	11/3/2014	Income	Check	1071	\$1,000.00	Alaska Sea Pilot PAC Fund	NaN	1621 Tongass Ave	Ketchikan
389256	389257	10/6/2014	Income	Check	1302	\$750.00	AlaskanSea Pilot PAC	NaN	1621 Tongass Ave	Ketchikan

	Result	Date	Transaction Type	Payment Type	Payment Detail	Amount	Last/Business Name	First Name	Address	City
389400	389401	12/18/2014	Income	Check	1078	\$250.00	Alaska Sea Pilot PAC	NaN	1621 Tongass Ave	Ketchikan
389470	389471	12/17/2015	Income	Check	1086	\$500.00	Alaska Sea Pilots	NaN	1621 Tongass Ave.	Ketchikan
389473	389474	12/21/2015	Income	Check	6531	\$250.00	Alaska Sea Pilots, LLC.	Richard Murphy	PO Box 920226	Dutch Harbor
393750	393751	11/5/2014	Income	Check	1073	\$1,000.00	Alaska Sea Pilot PAC	NaN	1621 Tongass Avenue	Ketchikan
394209	394210	12/19/2013	Income	Check	1195	\$1,000.00	Alaska Sea Pilot PAC fund	NaN	1621 Tongass Ave.	Ketchikan
394336	394337	12/18/2013	Income	Check	1196	\$500.00	Alaska Sea Pilot PAC Fund	NaN	1621 Tongass Ave	Ketchikan

128 rows × 26 columns

You could also do it using a lambda if you wanted

```
df[df.apply(lambda row: fuzz.token_sort_ratio(row['Last/Business Name'], "Alaska Sea Pilot PAC Fund"), axis=1) > 70]
```

We could technically clean it like below, but... it seems risky. Because it

is risky!

```
df.loc[df.apply(get_ratio, axis=1) > 75, "Last/Business Name"] = "Alaska Sea  
Pilot PAC Fund"
```

Want to hear when I release new things?

My infrequent and sporadic newsletter can help with that.

Hi, I'm Soma

I run a fake school in Brooklyn (<http://brooklynbrainery.com>) and a data journalism program (<http://ledeprogram.com>) at Columbia University's Journalism School.

I also co-host talks about food science and culture in a semi-monthly lecture series called Masters of Social Gastronomy (<http://omgmsg.com>). We have a podcast (<https://soundcloud.com/msgpodcast>) that doesn't get updated nearly often enough, too.

Sign up for my newsletter (<http://tinyletter.com/jsoma>) and I will *definitely* disappoint you.

Track me down

Miscellaneous projects

Handsome Atlas (<http://www.handsomeatlas.com/>)

✉ jonathan.soma@gmail.com
(mailto:jonathan.soma@gmail.com)

✉ soma@brooklynbrainery.com
(mailto:soma@brooklynbrainery.com)

🐦 dangerscarf (<https://twitter.com/dangerscarf>)

📷 dangerscarf (<https://instagram.com/dangerscarf>)

🐙 jsoma (<https://github.com/jsoma>)

✍ jsoma (<https://tinyletter.com/jsoma>)

Dabbler (<https://dabbles.in/>)

Vintage Visualizations
(<http://vintagevisualizations.com/>)

Interactive Singles Map
(<http://jonathansoma.com/singles/>)

visualizing.nyc (<http://visualizing.nyc/>)

Open-Source Language Map
(<http://jonathansoma.com/open-source-language-map/>)