Matching Algorithm

Yixuan Wang, Feifei Hong, Jianghui Hu,
Liutao Zhang, Fangzhou Yang, Yiran Zhao

Data Preparation

01 Data Cleaning

- Drop "category" column
- Clean business name
- Remove all comma and period
- Rename the column
  (give suffix 1 to the left dataset and suffix 2 to the right)

02 New Dataset

- Use cleaned dataset for matching

Left Dataset:

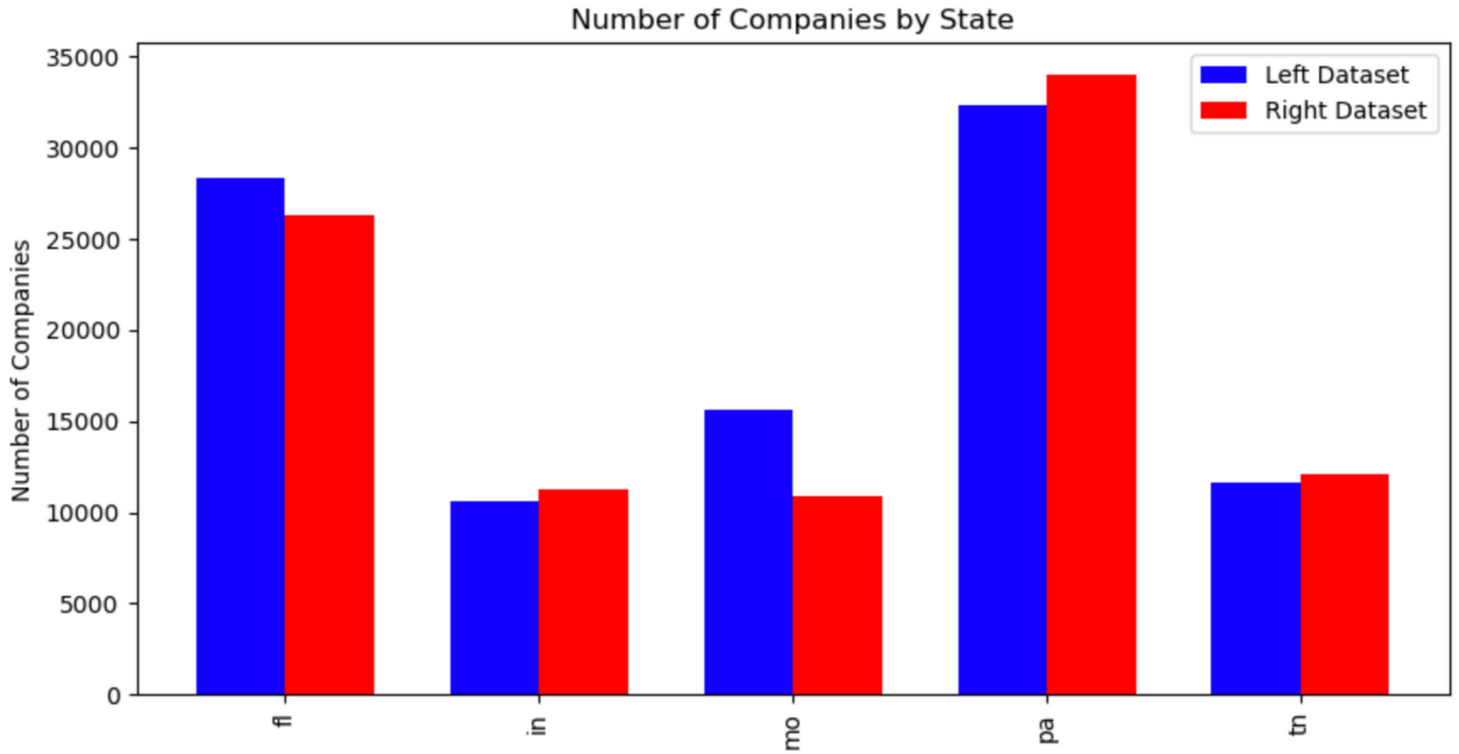
	id_1	name_1	address_1	city	state	zip_1
0	1	sourini painting inc	12800 44th st n	clearwater	fl	33762
1	2	wolff dolla bill llc	1905 e 19th ave	tampa	fl	33605
2	3	comprehensive surgery center llc	1988 gulf to bay blvd ste 1	clearwater	fl	33765
3	4	frank & adam apparel llc	13640 wright cir	tampa	fl	33626
4	5	moreno plus transport inc	8608 huron court unite 58	tampa	fl	33614

Right Dataset:

	id_2	name_2	address_2	city	state	zip_2
0	1	the ups store	87 grasso plaza shopping center	affton	mo	63123
1	2	st honore pastries	935 race st	philadelphia	pa	19107
2	3	perkiomen valley brewery	101 walnut st	green lane	pa	18054
3	4	sonic drive-in	615 s main st	ashland city	tn	37015
4	5	famous footwear	8522 eager road dierbergs brentwood point	brentwood	mo	63144

Data Examination

- About 10,000 companies on each dataset
- N^2 matching algorithm means 10 billions matches needs be completed
- Total of 5 states: FL, IN, PA, TN, MO
- Faster the matching process by group the data based on state



Our Final Algorithm - Part I

```
best_name_match = process.extractOne(row['name_normalized'],
right_data['name_normalized'].tolist(), scorer=fuzz.partial_ratio)

best_address_match = process.extractOne(row['address_normalized'],
right_data['address_normalized'].tolist(), scorer=fuzz.partial_ratio)

best_postal_code_match = process.extractOne(row['zip_code_normalized'],
right_data['postal_code_normalized'].tolist(), scorer=fuzz.partial_ratio)
```

Elaboration

01 Leveraging fuzz.partial_ratio:

1. Finding the longest common contiguous substring and scaling the similarity score
2. More robust and less sensitive to small variations

02 Process.extractOne:

1. A query string to search for
2. A list of strings to compare the query string against
3. A scoring function to compute the similarity between strings

Our Final Algorithm - Part II

```
def find_matches_extract_one(row, right_data, name_threshold=80,
postal_code_threshold=80, address_threshold =80):

if best_name_match[1] >= name_threshold and best_address_match[1] >=
address_threshold and best_postal_code_match[1] >= postal_code_threshold:
    matched_rows = right_data[(right_data['name_normalized'] ==
    best_name_match[0]) &
    (right_data['address_normalized'] == best_address_match[0]) &
    (right_data['postal_code_normalized'] == best_postal_code_match[0])]

    if len(matched_rows) > 0:
    index_right = matched_rows.index[0]
    row_right = right_data.loc[index_right]
return (row['business_id'], row_right['entity_id'], (best_name_match[1] +
best_address_match[1] + best_postal_code_match[1]) / 3)

    return None
```

Elaboration

01 This code block ensures that only high-quality matches are considered and finds the corresponding rows in 'right_data' for further processing.

02 Scaled the score with average of the similarity scores for the best name, address, and postal code matches.

And, ensures that if there's a valid match found in the 'right_data' dataset

Result

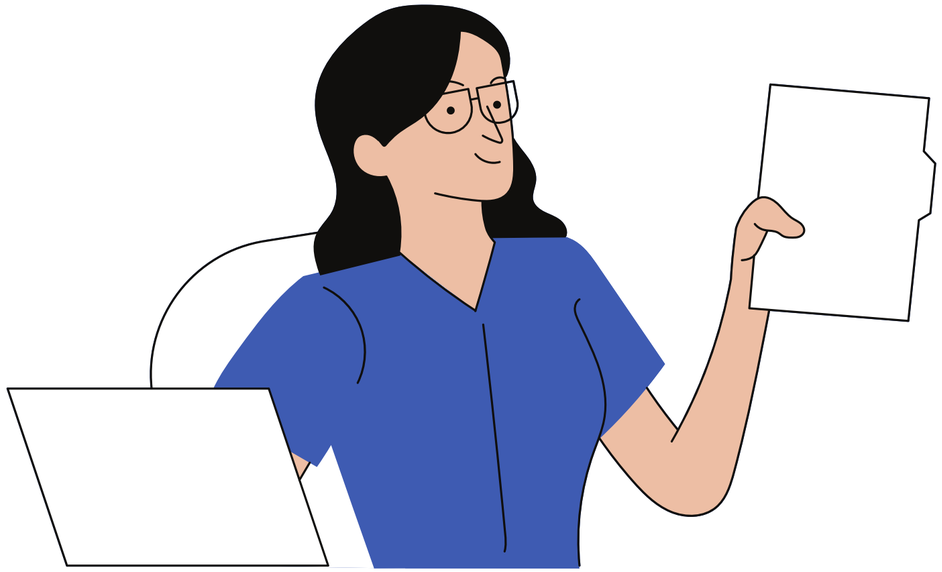
01 5065 matchd result

	id_1	id_2	confidence_score
4050	80504	57690.0	1.0
2423	53239	77112.0	1.0
3911	78240	54044.0	1.0
3073	64240	50358.0	1.0
3890	79022	39620.0	1.0
2490	55510	80398.0	1.0
3889	77852	39620.0	1.0
3884	77807	56066.0	1.0
3875	77670	9558.0	1.0
1061	23292	74450.0	1.0

02 Merged table

id_1	id_2	confidence_score	name_1	name_2	address_1	address_2
7	15925.0	1.0	jazz house supper club llc	jazz house supper club	9331 e adamo dr	9331 e adamo dr
16985	93741.0	1.0	integrity tire & automotive centers	integrity tire & automotive centers	4029 little rd	4029 little rd
88173	7971.0	1.0	cafe coco llc	cafe coco	210 louise ave	210 louise ave
16910	11874.0	1.0	hawthorne bottle shoppe	hawthorne bottle shoppe	2927 central ave	2927 central ave
20324	11874.0	1.0	hawthorne bottle shoppe	hawthorne bottle shoppe	2927 central ave	2927 central ave
16950	53466.0	1.0	heavenly nails inc	heavenly nails	1155 s dale mabry hwy ste 18	1155 s dale mabry hwy ste 18

Further Improvment



01

- More precise threshold for each attribute
- Could be more accurate

02

- Parallelize the matching process to speed up the execution.