Team 87 Group Project Presentation

Ashritha Shivakumaraswamy (as6601)
David Zhang (hz2811)
Kara Jiwachotkamjon (kj2545)
Zichen Wang(zw2913)

Executive Summary

EDA and Visualization

- Explored data pattern from left and right data set, leveraging venn-diagram to represent the initial match data across tables
- Identify unique data amount in each column to decide which ones to include when computing confidence score.
- Examine similarity between variables from left and right dataset

Data Preparation

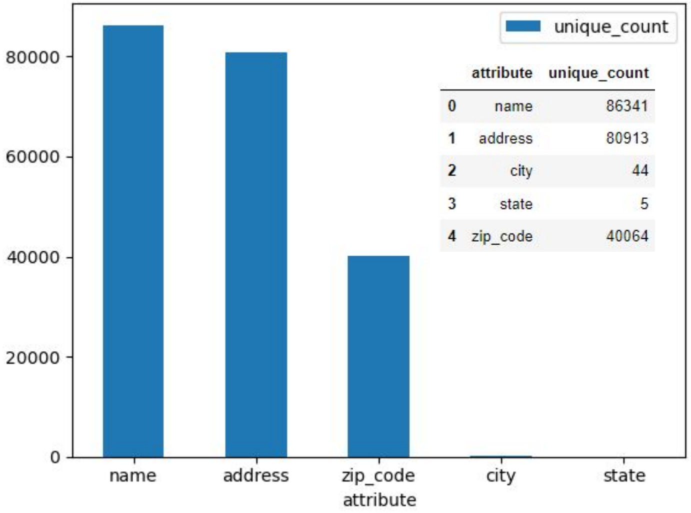
- Drop unnecessary columns to reduce data size
- Rename matching columns to allow merge for later algorithm
- Remove special characters, NA/missing values and set all strings to lowercase for both datasets.

Algorithm Development

- JW-Similarity(character comparison)
- Fuzzy-Wuzzy(String preprocessing)

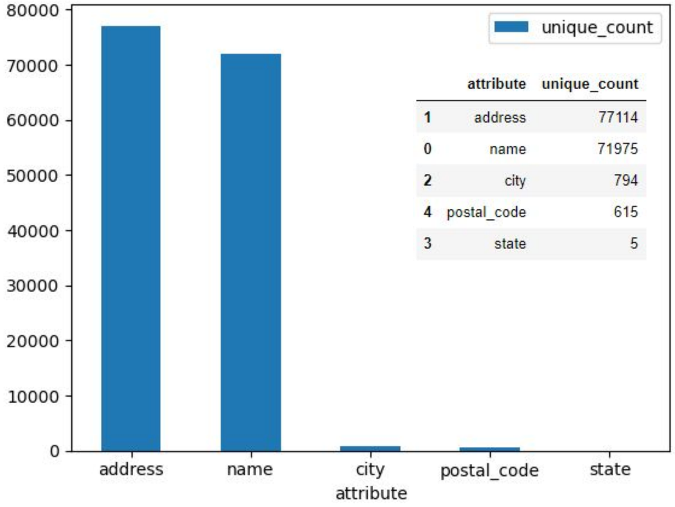
Exploratory Data Analysis (1/2)

Left Data Set



	attribute	unique_count
0	name	86341
1	address	80913
2	city	44
3	state	5
4	zip_code	40064

Right Data Set



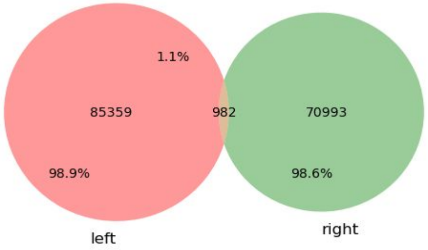
	attribute	unique_count
1	address	77114
0	name	71975
2	city	794
4	postal_code	615
3	state	5

Based on the unique count of each attribute between 2 data set, Left data set seems to have larger data compared to the Right data set. In addition, the Name and Address contributed to the largest between 2 groups. However, Zip_code in the Left data set displays a significant unique count of ~40k compared to the Right data set of only ~600.

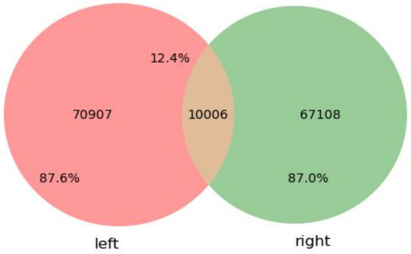
Exploratory Data Analysis (2/2)

Example of Venn-diagram visualization

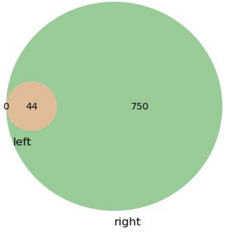
Name Attribute Similarity



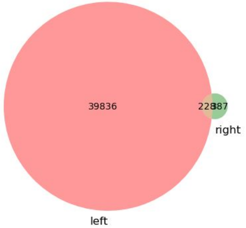
Address Attribute Similarity



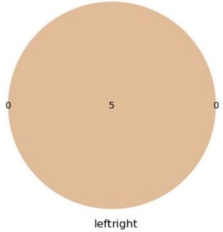
City Attribute Similarity



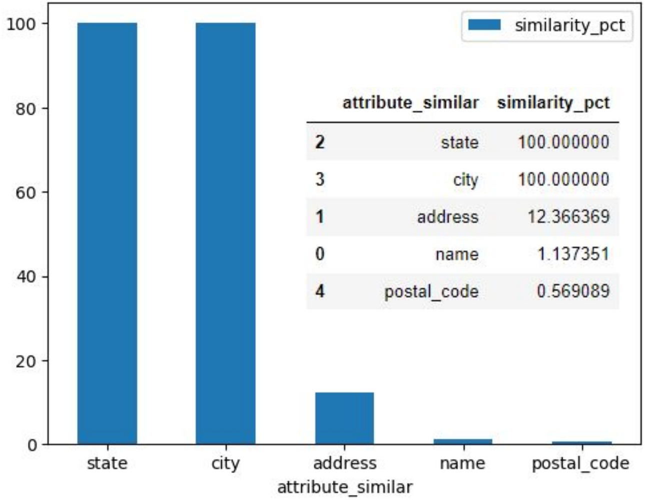
Zip code Attribute Similarity



State Attribute Similarity



Similarity Percentage Summary



	attribute_similar	similarity_pct
2	state	100.000000
3	city	100.000000
1	address	12.366369
0	name	1.137351
4	postal_code	0.569089

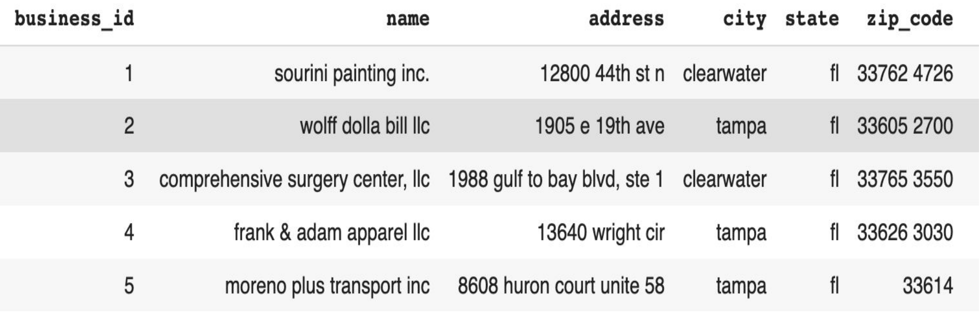
Without any format adjustments other than lower case, State and City attributes are the only attributes with 100% matches, while Address, Name and Postal code are the lowest matches.

Data preparation/cleaning

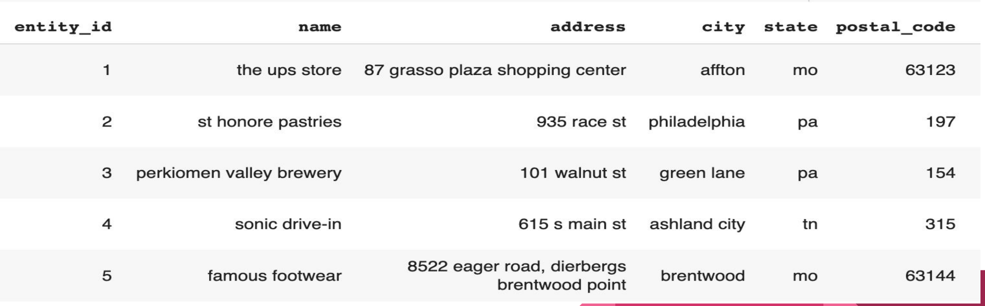
-Remove unnecessary columns including size, categories

-Rename ID columns from both datasets to allow merging

-Formatting data for algorithm by removing special characters and set all strings to lowercase.



business_id	name	address	city	state	zip_code
1	sourini painting inc.	12800 44th st n	clearwater	fl	33762 4726
2	wolff dolla bill llc	1905 e 19th ave	tampa	fl	33605 2700
3	comprehensive surgery center, llc	1988 gulf to bay blvd, ste 1	clearwater	fl	33765 3550
4	frank & adam apparel llc	13640 wright cir	tampa	fl	33626 3030
5	moreno plus transport inc	8608 huron court unite 58	tampa	fl	33614



entity_id	name	address	city	state	postal_code
1	the ups store	87 grasso plaza shopping center	affton	mo	63123
2	st honore pastries	935 race st	philadelphia	pa	197
3	perkiomen valley brewery	101 walnut st	green lane	pa	154
4	sonic drive-in	615 s main st	ashland city	tn	315
5	famous footwear	8522 eager road, dierbergs brentwood point	brentwood	mo	63144

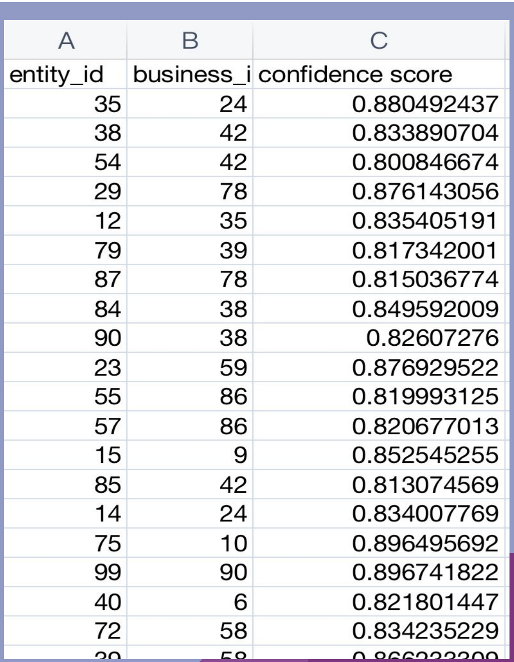
Algorithm Explanation and Confidence score

Jaro-Winkler Similarity

- Similarity Metrics: Jaro Distance
- Techniques: Character comparison - Transposition detection - Common prefix bonus
- Pros: Good performance on large datasets & Suitable for fuzzy matching tasks
- Cons: Limited to string comparisons & Sensitivity to variations in string length

FuzzyWuzzy

- Similarity Metrics: Levenshtein Distance
- Techniques: Preprocessing of strings - Scoring of similarity metrics - Selection of matching algorithm
- Pros: Flexible and customizable & Can handle multiple similarity metrics
- Cons: Slower performance compared to Jaro-Winkler & sensitive to choice of similarity metrics



A	B	C
entity_id	business_i	confidence score
35	24	0.880492437
38	42	0.833890704
54	42	0.800846674
29	78	0.876143056
12	35	0.835405191
79	39	0.817342001
87	78	0.815036774
84	38	0.849592009
90	38	0.82607276
23	59	0.876929522
55	86	0.819993125
57	86	0.820677013
15	9	0.852545255
85	42	0.813074569
14	24	0.834007769
75	10	0.896495692
99	90	0.896741822
40	6	0.821801447
72	58	0.834235229
39	58	0.866222209

Thank You