Foursquare Point-of-Interest Matching

New Horizons: Yu Cao, Tim Gorman, Ling Zhou

Developing Machine Learning to Correctly Identify Points-of-Interest from Foursquare Location Data

[https://github.com/gormantt/foursquare-location-matching]

Foursquare - Location Matching

Match point of interest data across datasets

kaggle

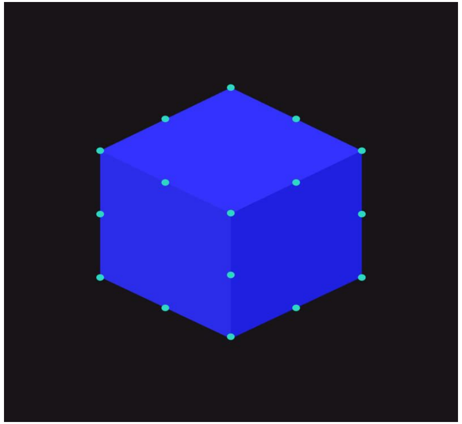
Location Data

Foursquare City Guide

FOURSQUARE Swarm

Points of Interest (POIs)

Insights Solutions in FSQ Location Intelligence



Who uses this data?

jetBlue airbnb Uber Coca-Cola Spotify

Goal: correctly identify duplicate points-of-interest

Datasets from Kaggle competition: train.csv pairs.csv

https://www.kaggle.com/competitions/foursquare-location-matching

Foursquare - Location Matching

Match point of interest data across datasets

kaggle

11 attribute fields

train.csv

id	name	latitude	longitude	address	city	state	zip	country	url	phone	categories	point_of_interest
E_263a142bd817bc	Susans	-33.934817	151.069265	NaN	NaN	NaN	NaN	AU	NaN	NaN	Women's Stores	P_d6898efecc074f
E_df41d13127f11d	清玉手調原味茶 King Tea	25.044300	121.515972	中正區南陽街26號	台北市	NaN	100	TW	http://qingyutea.blogspot.com	223709633	Tea Rooms, Bubble Tea Shops	P_691d41e345da73
E_473cd396422be0	Everybody Offer Center (7th mile)	1.469906	110.329860	NaN	NaN	NaN	NaN	MY	NaN	NaN	Clothing Stores	P_9c68d66db57814
E_73d94497af3313	Cancún Center	21.134743	-86.747002	Blvd. Kukulcan Km. 9.5	Cancún	QR	77500	MX	NaN	NaN	Convention Centers, Meeting Rooms, Event Spaces	P_e57277af20886d
E_f816e07f244bc9	Outback Steakhouse	39.991958	-74.176346	Hwy. 35 & Woodland Dr.	Middletown	NJ	07748	US	NaN	NaN	Steakhouses	P_bc5a827b040abf
...												

1 million+ places entries

Places entries were heavily altered to include noise, duplications, extraneous, or incorrect information.

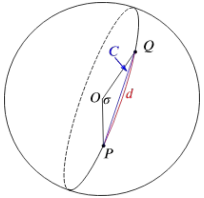
pairs.csv

A pregenerated subset of pairs of entries from train.csv:

'id_1' 'name_1' 'latitude_1' 'longitude_1' 'address_1' 'city_1' 'state_1' 'zip_1' 'country_1' 'url_1' 'phone_1' 'categories_1' 'id_2' 'name_2' 'latitude_2' 'longitude_2' 'address_2' 'city_2' 'state_2' 'zip_2' 'country_2' 'url_2' 'phone_2' 'categories_2' 'match'

'match'=True if two entries have the same POI value in train.csv

Plan Outline



Analyze train.csv → Analyze pairs.csv → Problem Reduction and Feature Engineering → Modeling → Results and Analysis

COUNT VECTORIZATION

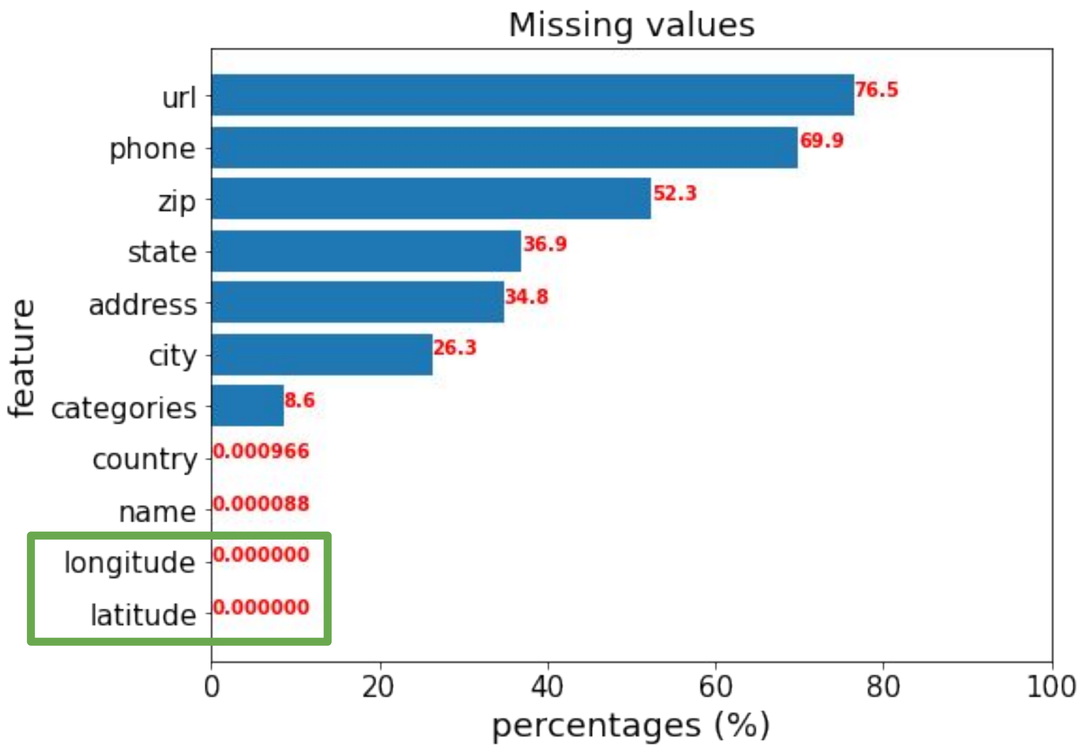
sequence matching

The Levenshtein Algorithm

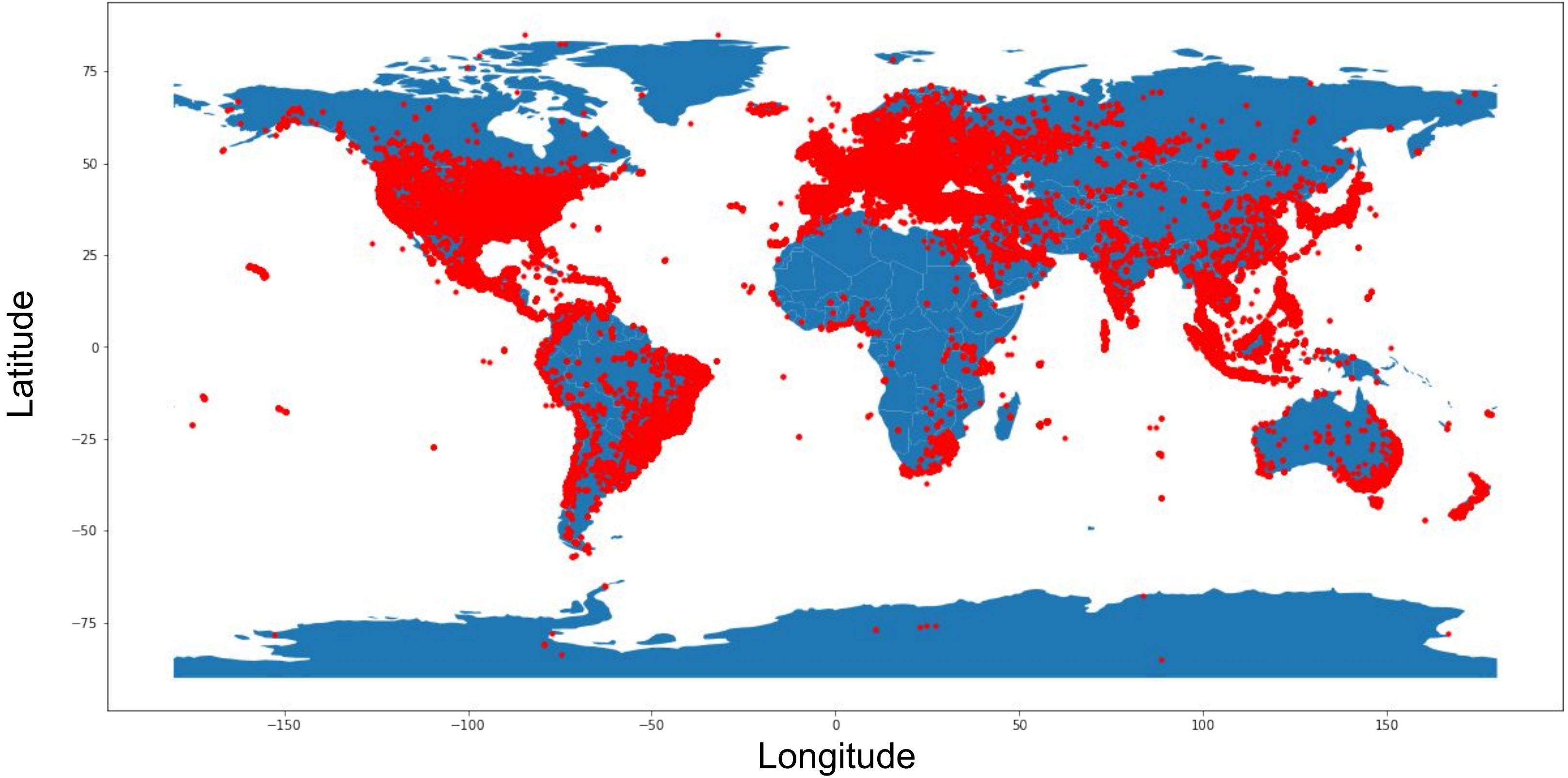
Baseline

Exploratory Data Analysis: train.csv

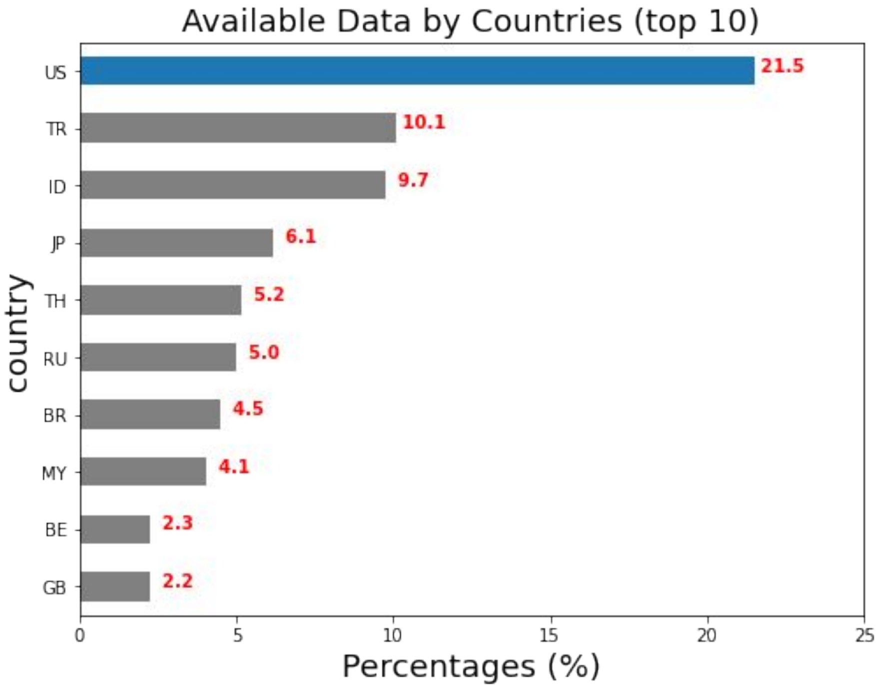
Exploratory Data Analysis: train.csv



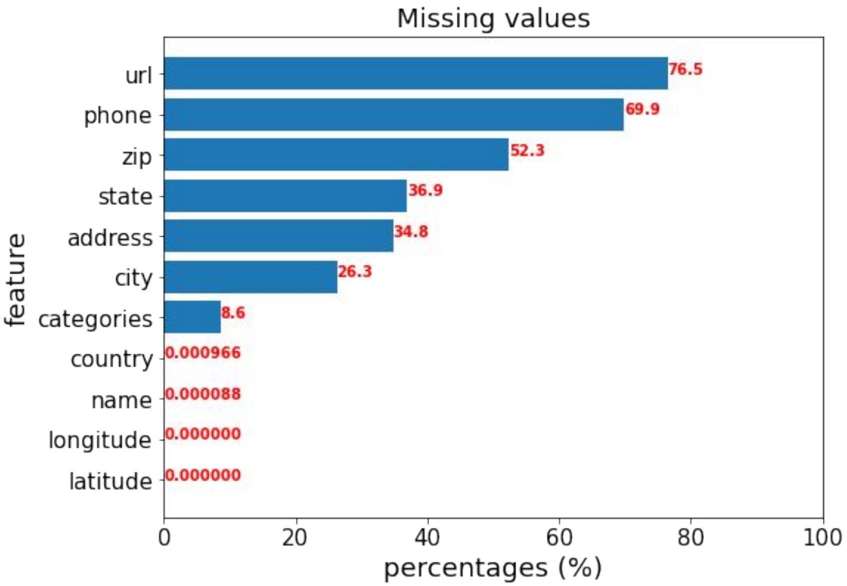
Exploratory Data Analysis: train.csv

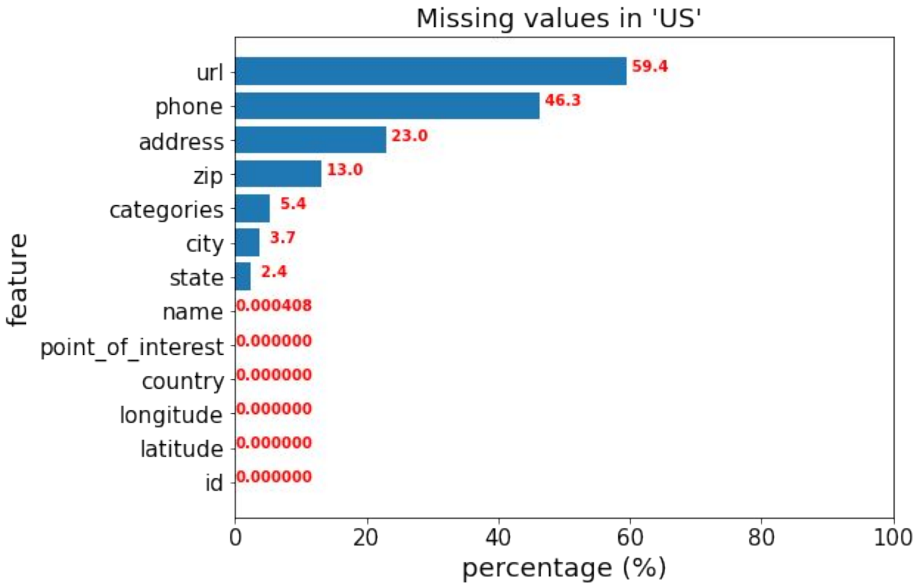


Exploratory Data Analysis: train.csv

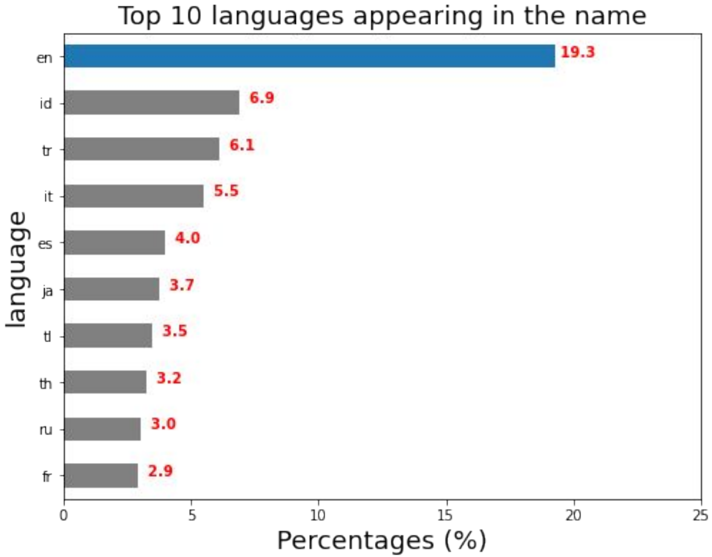


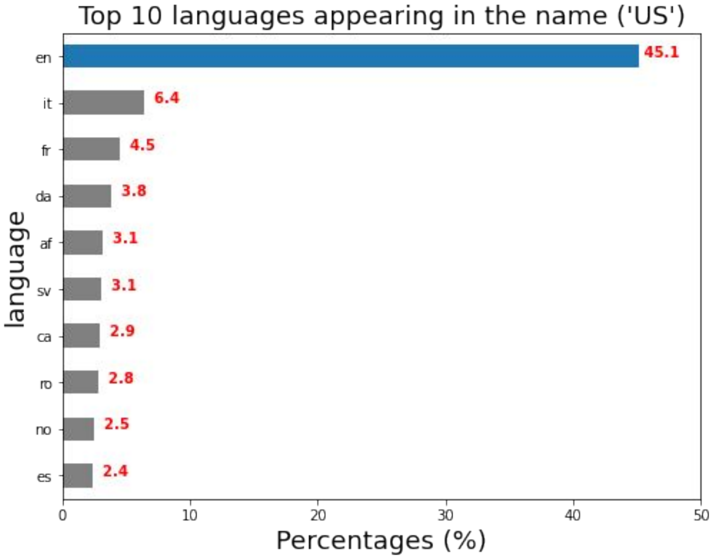
Exploratory Data Analysis: train.csv US Data





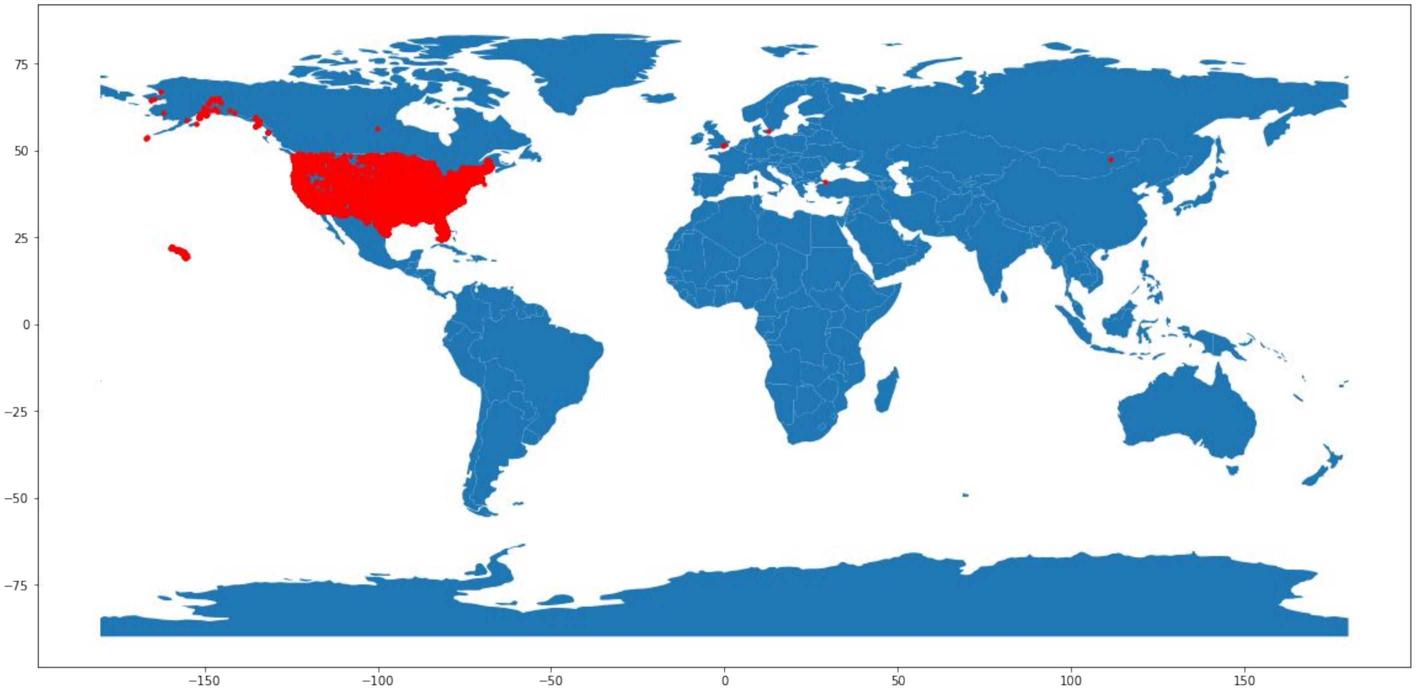
Exploratory Data Analysis: train.csv US Data





Exploratory Data Analysis: train.csv US Data

Data with country = “US”



Exploratory Data Analysis: train.csv US Data

Noise in the Data: Same POI, Same Brand, but Different Places

name	latitude	longitude	address	city	state	zip	country	url	phone	categories	point_of_interest
Gregg Steiner - Nationwide Insurance & Financi...	33.607341	-117.877248	3 Corporate Plaza Dr Ste 250	Newport Beach	CA	92660	US	http://t.co/4eDIQ4ptQn	+19492200701	Financial or Legal Services	P_399ab9d64f2a2e
James S Wills Agency - Nationwide Insurance	34.002130	-81.772107	100 North Main St	Saluda	SC	29138	US	http://agency.nationwide.com/agent/james-wills...	8644450011	NaN	P_399ab9d64f2a2e
Nationwide Insurance	39.962014	-82.885847	583 S Yearling Rd	Columbus	OH	43213	US	http://t.co/4eDIQ4ptQn	+16145752643	Financial or Legal Services	P_399ab9d64f2a2e
Nationwide Insurance	39.280624	-76.611005	209 Key Hwy	Baltimore	MD	21230	US	http://agency.nationwide.com/agent/lawrence-l-...	+14108376400	Financial or Legal Services	P_399ab9d64f2a2e
Nationwide Insurance	41.224360	-73.071814	333 Boston Post Rd	Milford	CT	06460	US	http://t.co/4eDIQ4ptQn	+12038789003	Financial or Legal Services	P_399ab9d64f2a2e

Exploratory Data Analysis: train.csv US Data

Noise in the Data: Same POIs, Different Names

name	latitude	longitude	address	city	state	zip	country	url	phone	categories	point_of_interest
Amanda C. Castro, MD	39.779379	-75.555298	1600 Rockland Rd	Wilmington	DE	19803	US	https://findaprovider.nemours.org/details/1666...	+18004164441	Mental Health Offices, Doctor's Offices	P_ce9291000a8f0b
Arieda Gjikopulli, MD	39.779379	-75.555298	1600 Rockland Rd	Wilmington	DE	19803	US	https://findaprovider.nemours.org/details/1572...	+18004164441	Doctor's Offices	P_ce9291000a8f0b
Cara J. Lasley, MD	39.779379	-75.555298	1600 Rockland Rd	Wilmington	DE	19803	US	https://findaprovider.nemours.org/details/1877...	+18004164441	Doctor's Offices	P_ce9291000a8f0b
Charles D. Vinocur, MD	39.779379	-75.555298	1600 Rockland Rd	Wilmington	DE	19803	US	https://findaprovider.nemours.org/details/1148...	+18004164441	Doctor's Offices	P_ce9291000a8f0b
Christian Pizarro, MD	39.779379	-75.555298	1600 Rockland Rd	Wilmington	DE	19803	US	https://findaprovider.nemours.org/details/1312...	+18004164441	Doctor's Offices	P_ce9291000a8f0b

Exploratory Data Analysis: train.csv US Data

Noise in the Data: Same POI, but Inaccurate Information (e.g. zip)

name	latitude	longitude	address	city	state	zip	country	url	phone	categories	point_of_interest
12 hour work day = lost voice	28.473537	-81.464653	NaN	Orlando	FL	32819	US	NaN	NaN	Coworking Spaces	P_a3fddc2f0a77e7
@ Work	28.473057	-81.272141	NaN	Orlando	FL	32829	US	NaN	NaN	Fairs	P_a3fddc2f0a77e7
@ work	28.588776	-81.416936	Parkway Commerce Blvd.	Orlando	FL	32808	US	NaN	NaN	NaN	P_a3fddc2f0a77e7
Heading to work!	28.484512	-81.408905	NaN	Orlando	FL	32839	US	NaN	NaN	Bowling Alleys	P_a3fddc2f0a77e7
Hell Aka Work	28.798779	-81.296295	132 commerce way	Sanford	FL	NaN	US	NaN	NaN	Tech Startups	P_a3fddc2f0a77e7

EDA Summary of train.csv

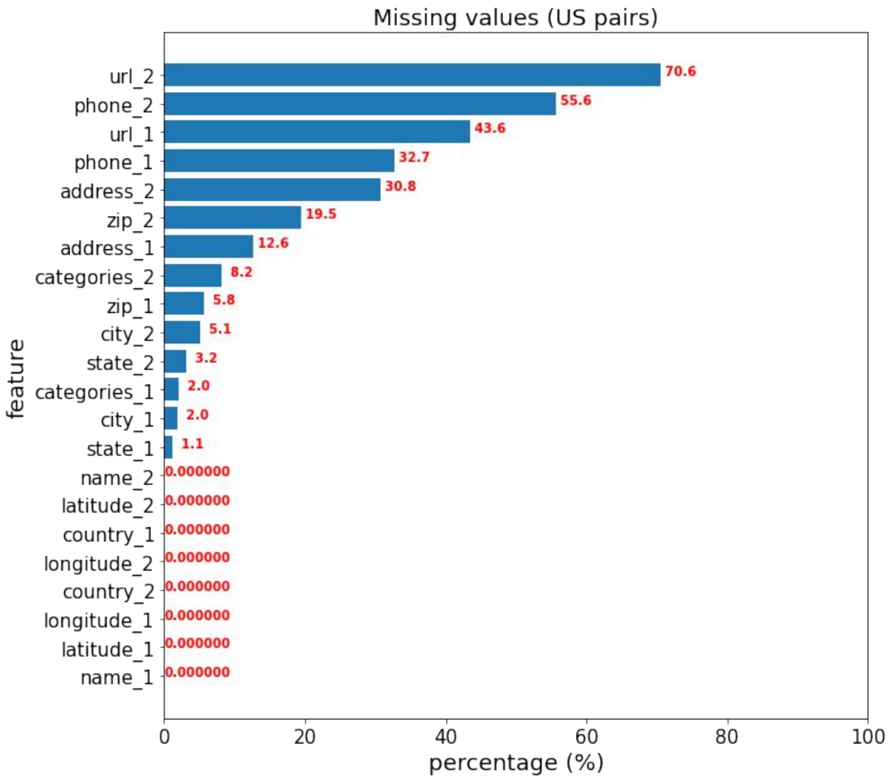
- `train.csv` is full of complexity, missing values, and errors.
- Reducing to US data simplifies classification problem:
  - Reduces Missing Values
  - Reduces language variation
- Many-to-many classification will be a difficult first step
  - Solution: Build first models on pairs.csv
    - Reduces many-to-many classification to binary classification problem

EDA and Data Preparation of pairs.csv

EDA and Data Preparation of pairs.csv (US)

- 117708 US pairs (578907 in total)
- Missing values vary by feature
- US True/False Imbalance:

  True / False: 72% / 28%



EDA and Data Preparation of pairs.csv (US)

Steps to clean the (US) data:

- Fill the missing (string) values with the empty string
- Change the format in the "state" pairs

EDA and Data Preparation of pairs.csv: State Data

After Cleaning:

```
unique values in state_1:
['CA' 'GA' 'NM' 'FL' 'VA' 'TN' 'NJ' 'UT' 'IN' 'NC' 'WI' '' 'NV' 'KS' 'MA'
 'MS' 'AZ' 'MI' 'NY' 'TX' 'IL' 'AL' 'PA' 'OK' 'AR' 'KY' 'MO' 'WV' 'CO'
 'NE' 'OH' 'OR' 'MT' 'CT' 'NH' 'MD' 'HI' 'WA' 'WY' 'RI' 'VT' 'IA' 'MN'
 'LA' 'SC' 'ND' 'DE' 'DC' 'SD' 'AK' 'ID' 'ME' 'CE']
unique values in state_2:
['CA' 'GA' 'NM' 'FL' 'VA' 'TN' '' 'NJ' 'UT' 'IN' 'NC' 'WI' 'NV' 'KS' 'MA'
 'MS' 'AZ' 'MI' 'NY' 'TX' 'IL' 'AL' 'PA' 'OK' 'AR' 'KY' 'WV' 'CO' 'NE'
 'OH' 'OR' 'MT' 'SC' 'CT' 'NH' 'MO' 'MD' 'HI' 'WA' 'WY' 'VT' 'IA' 'MN'
 'LA' 'ND' 'DE' 'DC' 'SD' 'AK' 'ID' 'ME' 'RI' 'UK' 'CE' 'NU' '国外']
```

Feature Engineering

- Quantify relationship between (US) pairs in pairs.csv for modeling
  - String Feature Differences (everything but latitude and longitude):
    - Sequence Matching
    - Levenshtein Distance
    - Count Vectorization and Cosine Similarity
  - Location Feature Difference:
    - Angular difference between lat/long. pairs

These features will be used as model input.

Modeling

Modeling Process

Compare String Diff. Metrics through Logistic Regression with Hyperparameter tuning and K-fold CV

→

Best Performing Metric: “Levenshtein” Difference

→

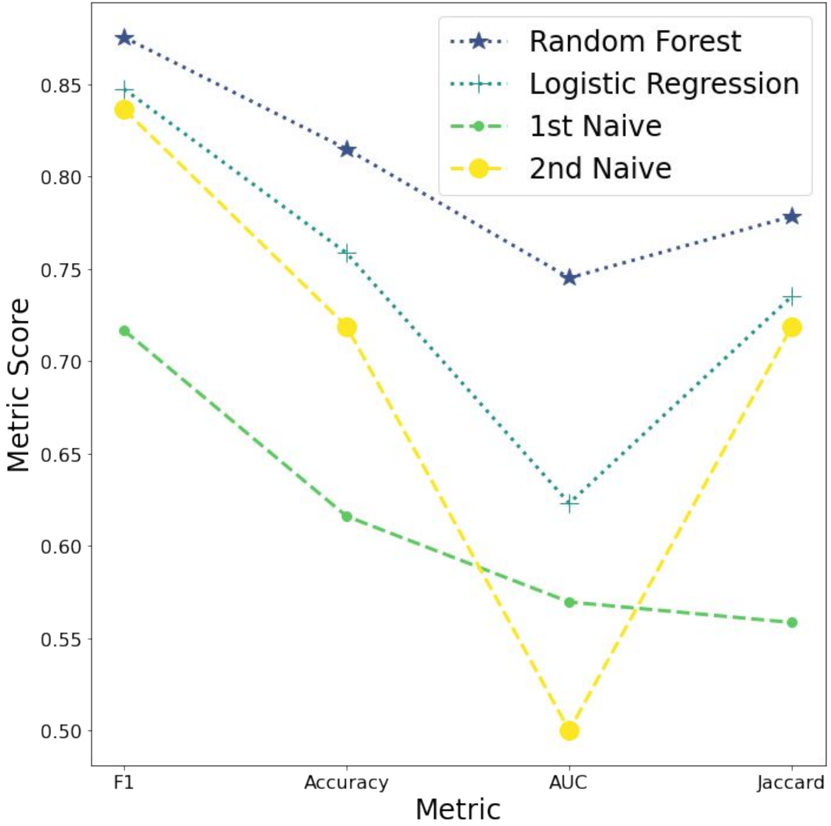
Random Forest with Levenshtein String Differences, Hyperparameter tuning and K-fold CV

↓

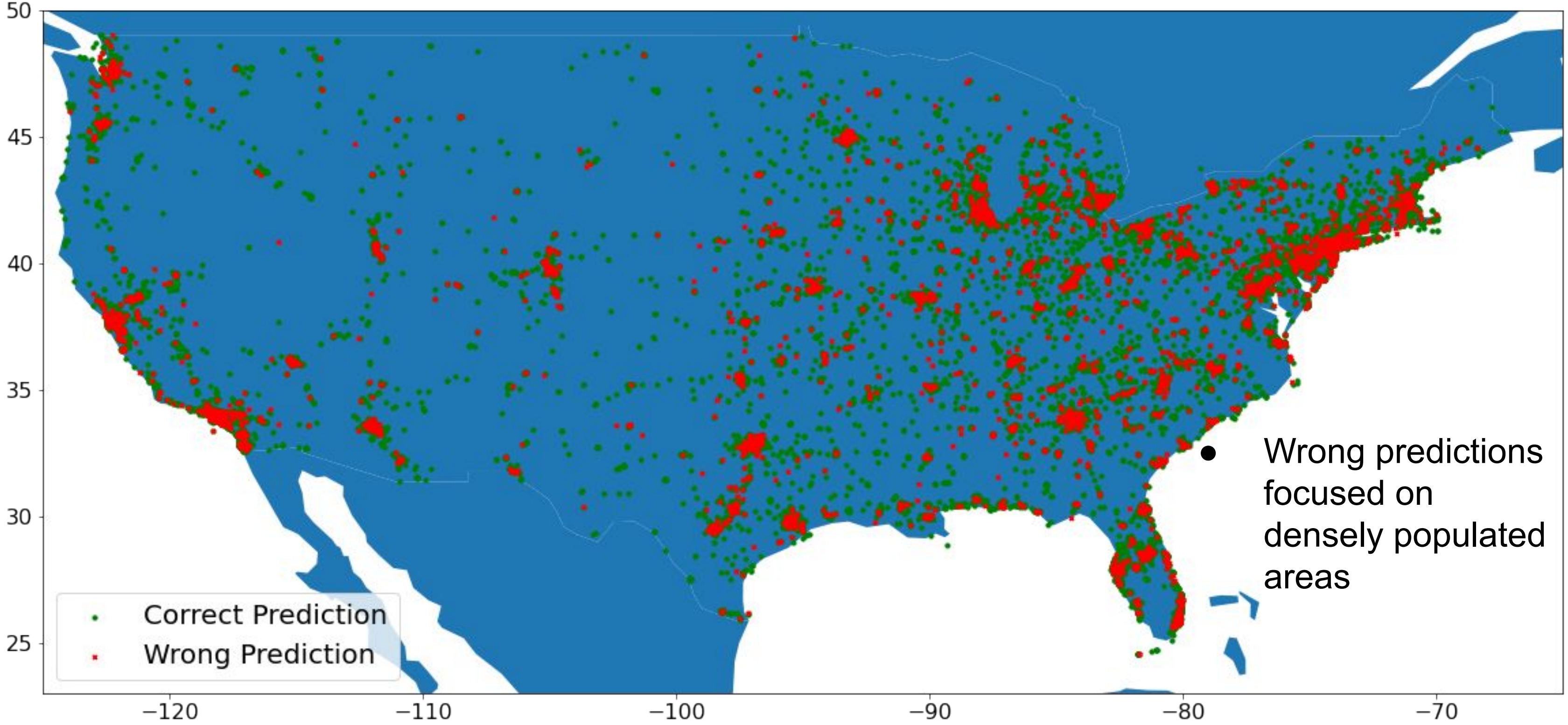
Compare Log Reg and Random Forest to Naive Models on train-test-split

- 1st Naive Model: Lev. Diff < 0.1→ match = True
    - Else: Random Assignment
    - Differenced Features:
        - Address
        - Name
        - Category
- 2nd Naive Model: Guess all match = True

Analysis of Results: Random Forest Wins!



Random Forest Results Spatial Distribution



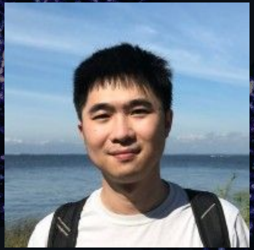
- Wrong predictions focused on densely populated areas

Conclusions and Future Work

- A random forest does the best job correctly predicting POI matches
- Expand the random forest model to other countries in pairs.csv
  - May require adapting for different languages
- Apply updated model to training set (train.csv)

Thank you for your attention!

Ling Zhou

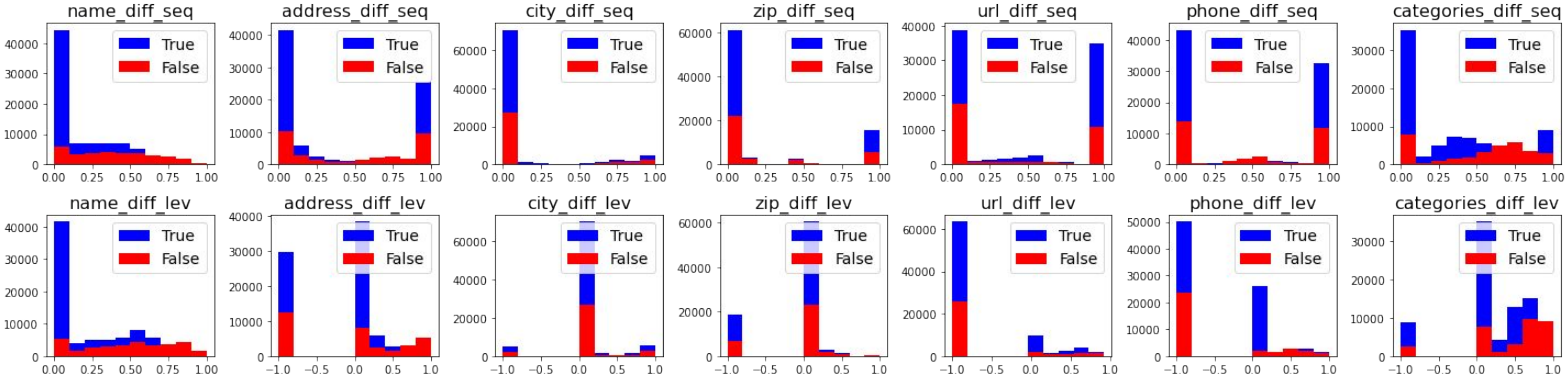


Yu Cao

Tim Gorman

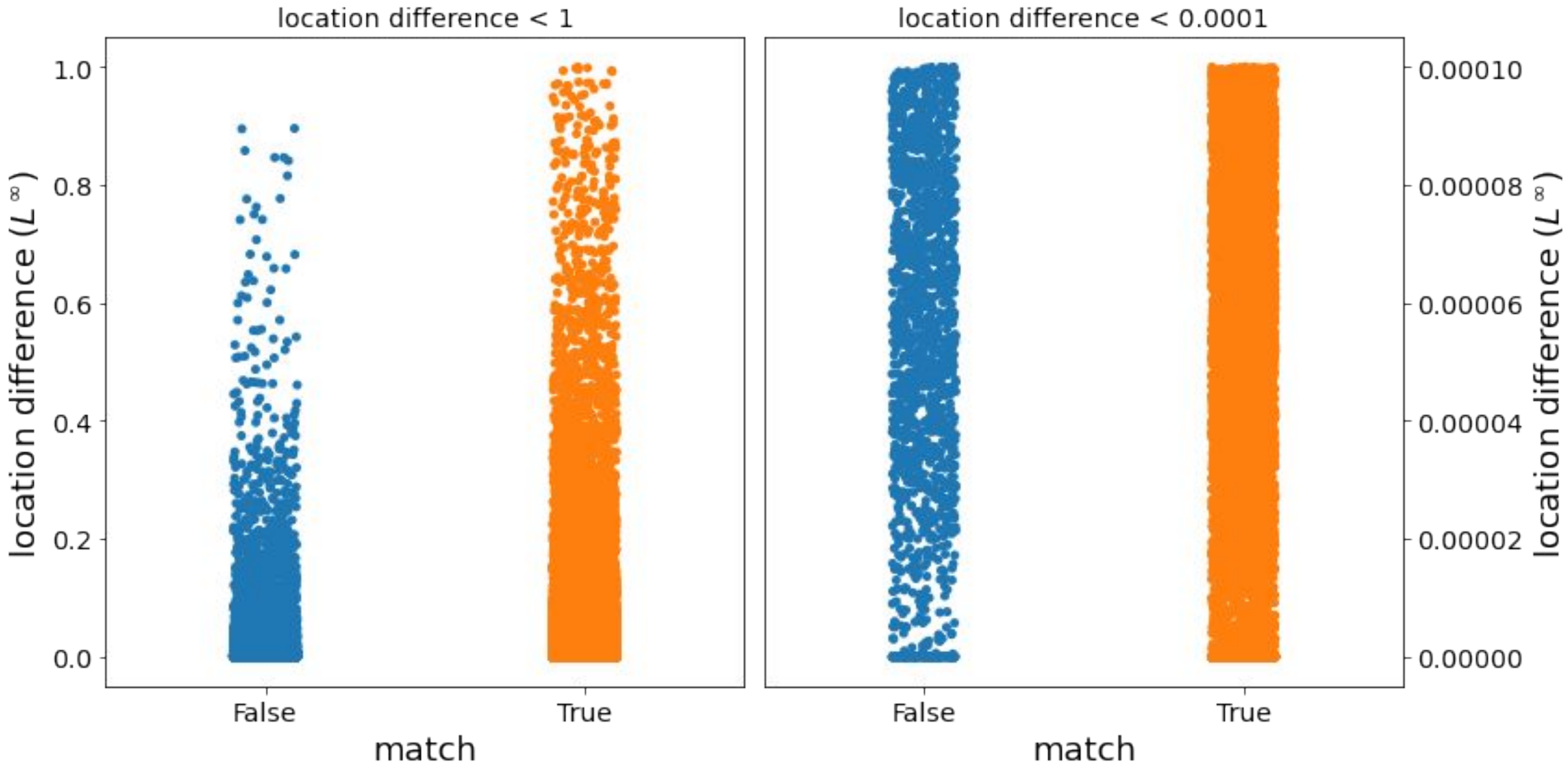
Extra Slides

SequenceMatcher vs. Levenshtein



Logistic Regression: Verifying Best String Metric

String Metric	**Best Accuracy Score**	**Best Hyper Parameters**
Sequence Matching	0.752	'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'
Levenshtein Distance	0.761	C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'
Cosine Similarity	0.739	'C': 0.05, 'penalty': 'l2', 'solver': 'newton-cg'


location difference < 1
location difference < 0.0001
location difference (L^∞)
1.0
0.8
0.6
0.4
0.2
0.0
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
False
True
match
match

Foursquare - Location Matching

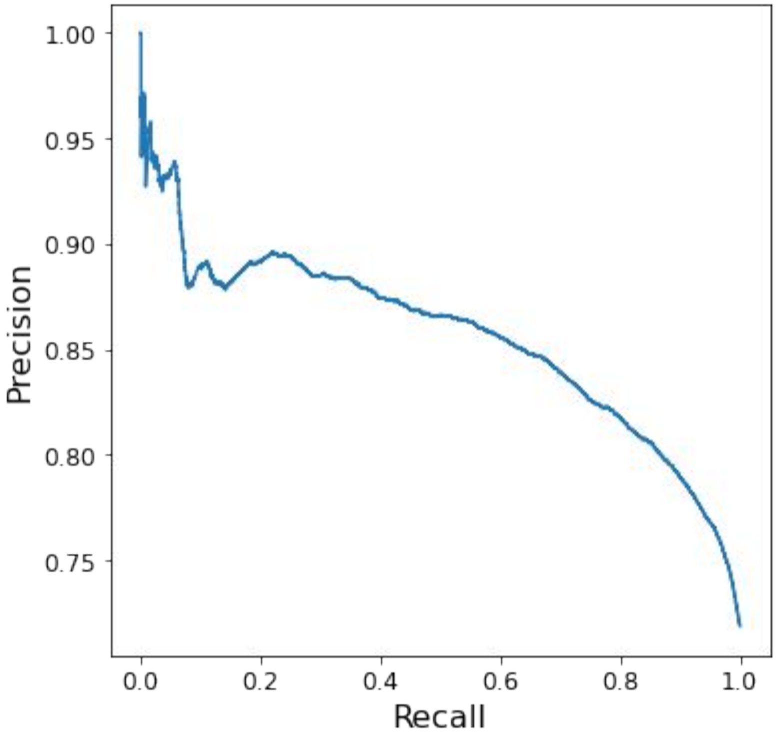
Match point of interest data across datasets

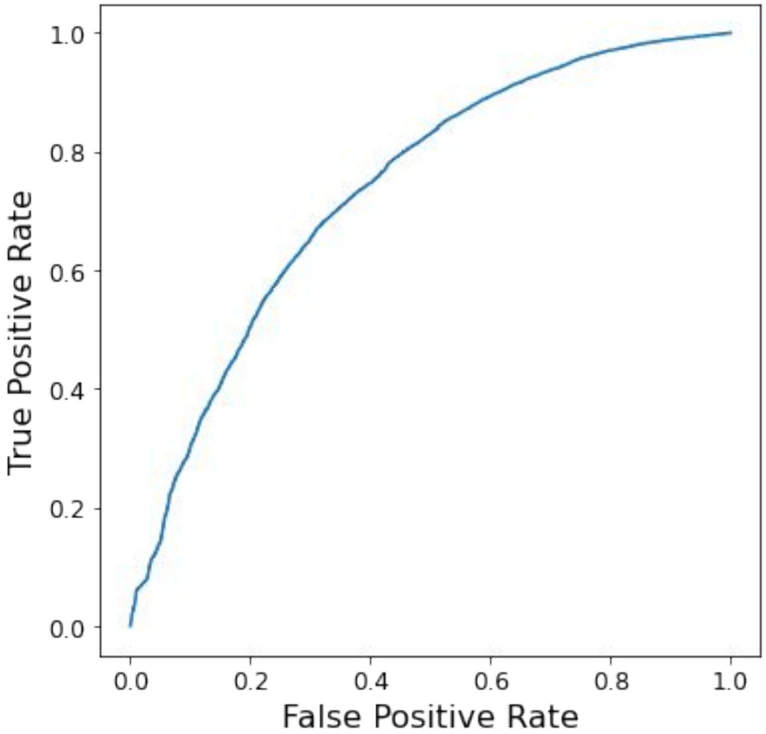
kaggle

Example Data from “train.csv”

id	name	latitude	longitude	address	city	state	zip	country	url
E_00002a131a2bf6	ministry of youth	29.36435235908348	47.97136230015956					KW	
E_0000764d65557e	McDonald's	-7.265894412994385	112.74938201904295	Plaza Surabaya, Pemuda Building				ID	
E_00007dcd2bb53f	TOGO'S Sandwiches	38.25779696430681	-122.06459937900875	1380 Holiday Ln., Ste. B	Fairfield	CA	94534	US	https://locations.togos.com/ll/US/CA/Fairfield/1380-Holiday-Ln_*-Ste_-B
E_0000890af22ff5	Flohmarkt Am Rathaus Steglitz	52.4570449854665	13.32247549214842					DE	
E_00009ab517afac	Starbucks	26.305219795470677	50.12944377493889	Ibis Avenue	Dhahran	Ash Sharqiyah	34465	SA	
E_0000c362229d93	Coffee Cat	7.082217570120776	125.61024431048877	F. Torres St.	Davao City	Davao Region	8000	PH	

Log. Reg. Results for Lev. Diff.





Log. Reg. Results for Lev. Diff.