GeoDD: End-to-end spatial data de-duplication system

Mykola Trokhymovych
Ukrainian Catholic University
trokhymovych@ucu.edu.ua

Oleksandr Kosovan
Ivan Franko National University of Lviv
Oleksandr.Kosovan.AEKE@lnu.edu.ua

October 13, 2022

Agenda

1. Introduction
2. Related work
3. Data observation
4. System architecture
5. Summary

Introduction. Motivation

- The information flow is huge and diverse
- Aggregated information can be valuable for business and government
- Automation of data-deduplication can provide close to real-time stats

Deduplication example:

USJ

Universal Studio JAPAN

Universal Studios Japan
(ユニバーサル・スタジオ・ジャパン)

Why Foursquare?

- Foursquare has a map of over 105 million places of interest in 190 countries. (Source: Techcrunch)
- Foursquare reaches ~1B check-ins per year. (Source: Financesonline)
- Collects information including texts, location, meta features, that allows to experiment with cross-domain models
- Open data within the competition

Introduction. Problem formulation

Location records duplicates:

Two records, that include location, text descriptions and other features are duplicates if they represent one specific physical entity

Location records de-duplication:

Given the set of location records find all existing duplicates and define groups of records that represent one specific physical entity

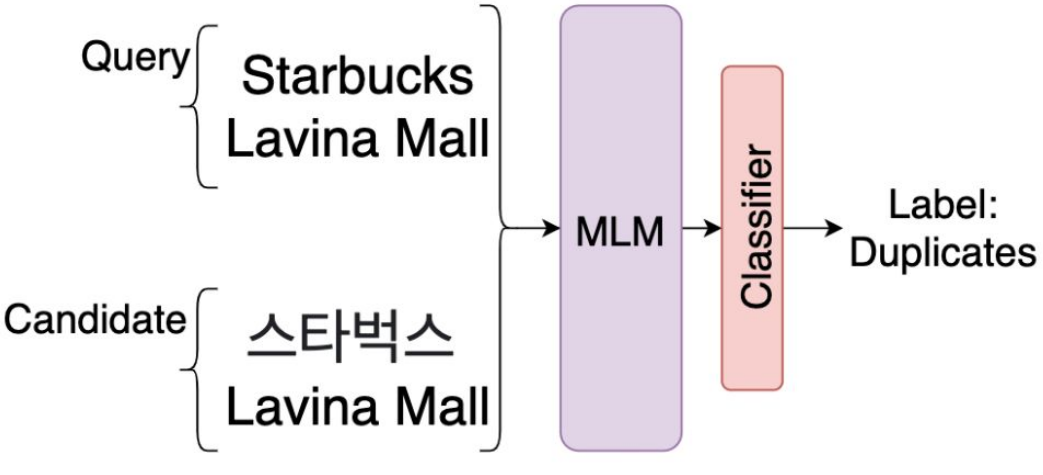
Research goals

- Analyze location records datasets. Define the specific data features and limitation, design a methodology for data preprocessing.
- Experiment with methods to search the duplicates within the DB
- Implement an end-to-end location data de-duplication system.

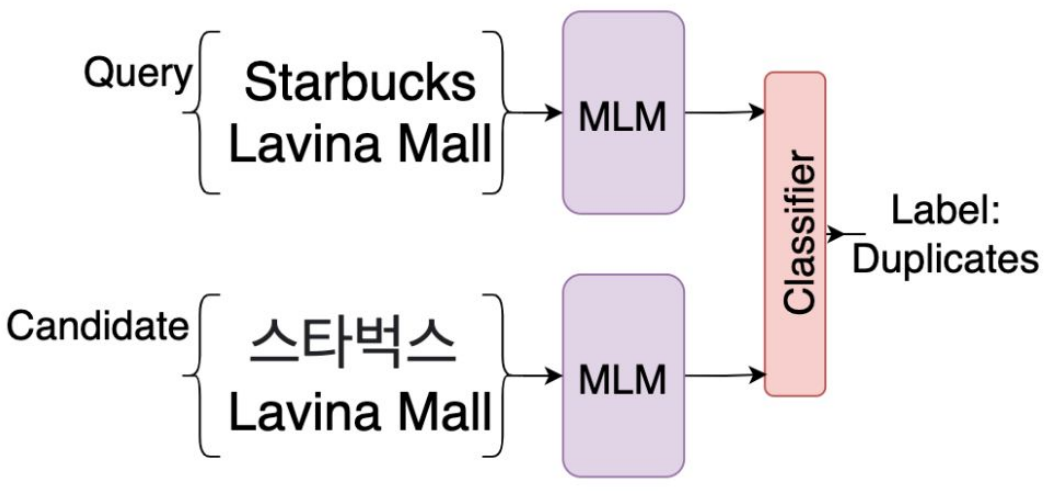
Related work

Natural language processing

Word-based approach



Sentence-based approach

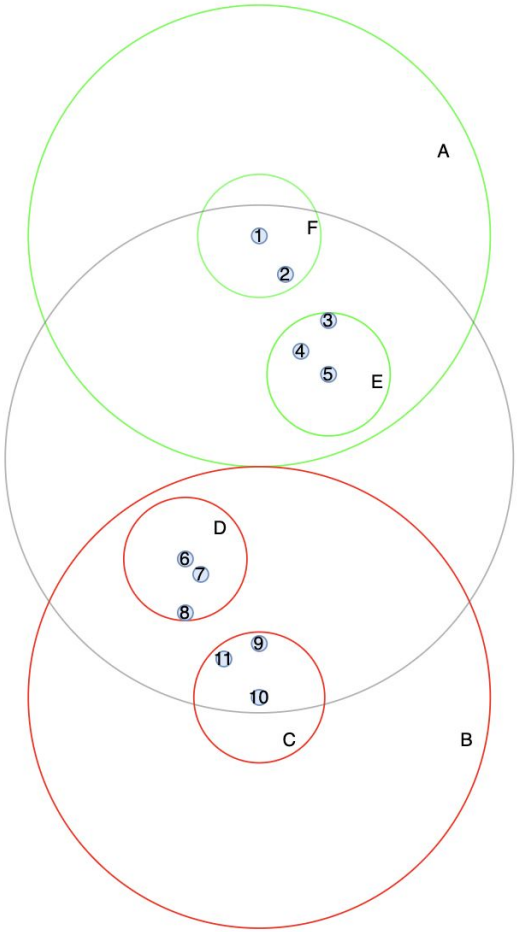


Why sentence-based approach:

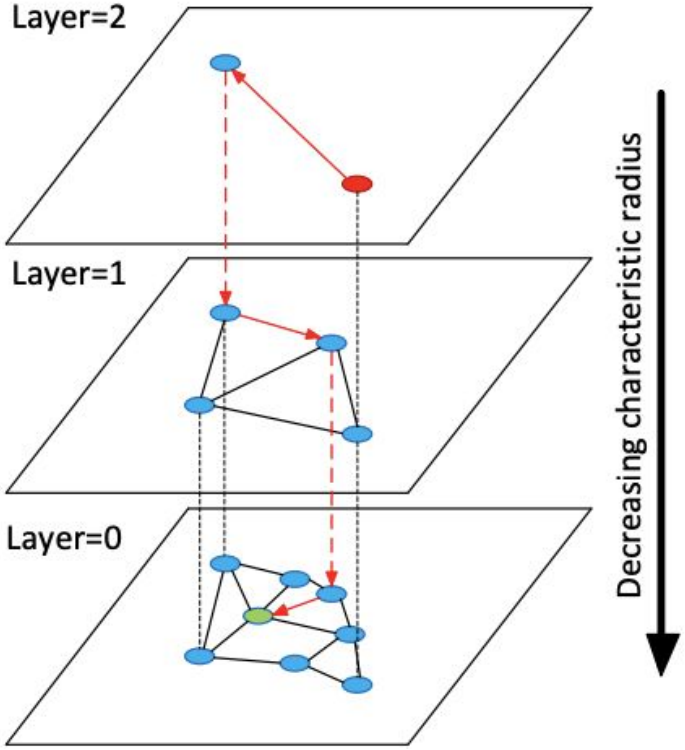
- Allows caching of sentence embeddings
- Allows processing query and candidates separately -> use ANN models
- Usually lighter and faster on inference
- Usually lower accuracy

Spatial data and search

BallTree index



NMSLIB index



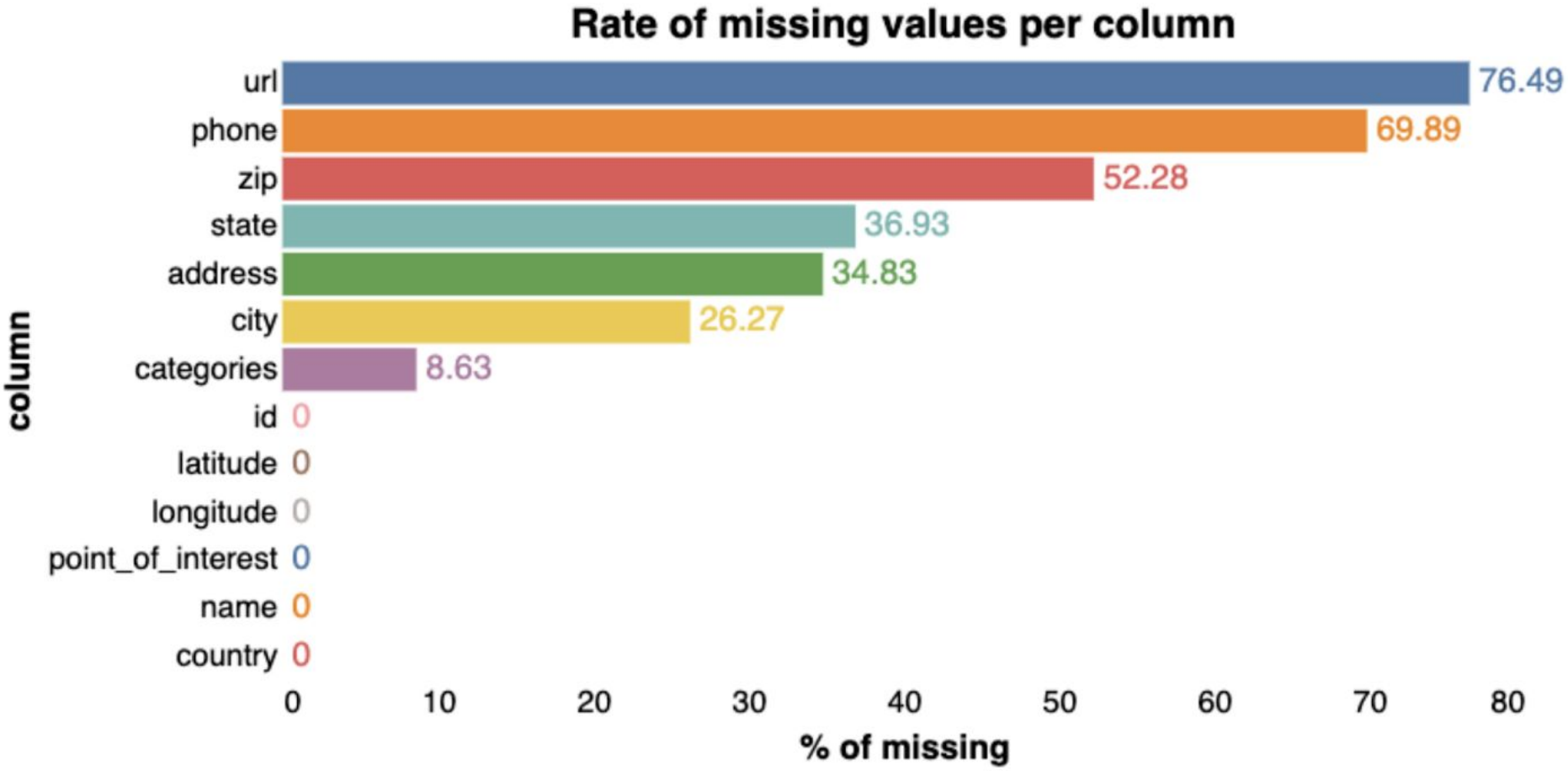
source: https://arxiv.org/abs/1603.09320

Data observation

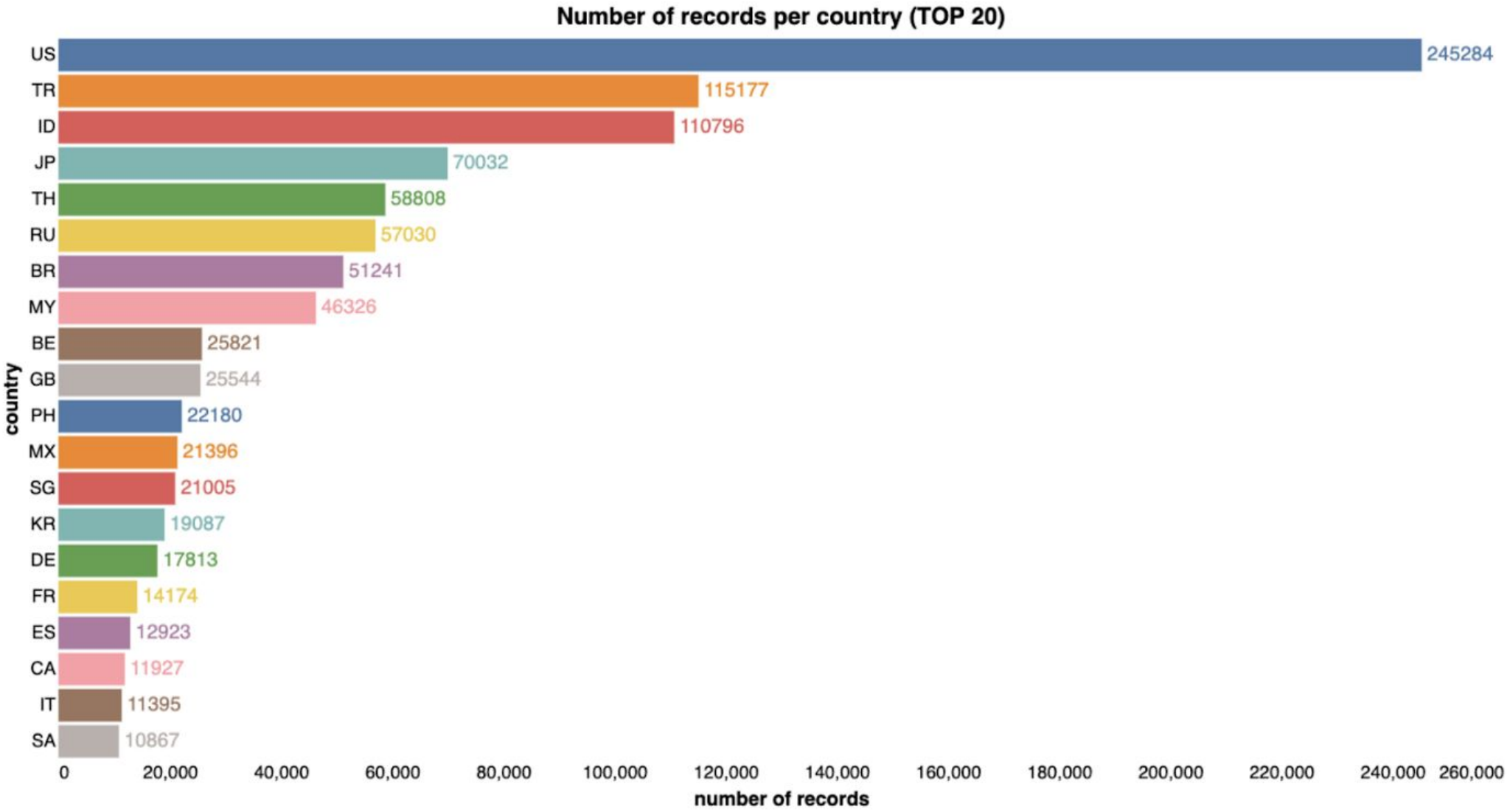
EDA. Data sample

	id	name	latitude	longitude	address	city	state	zip	country	url	phone	categories	point_of_interest
0	E_000001272c6c5d	Café Stad Oudenaarde	50.859975	3.634196	Abdijstraat	Nederename	Oost-Vlaanderen	9700	BE	NaN	NaN	Bars	P_677e840bb6fc7e
1	E_000002eae2a589	Carioca Manero	-22.907225	-43.178244	NaN	NaN	NaN	NaN	BR	NaN	NaN	Brazilian Restaurants	P_d82910d8382a83
2	E_000007f24ebc95	ร้านตัดผมการาเกด	13.780813	100.484900	NaN	NaN	NaN	NaN	TH	NaN	NaN	Salons / Barbershops	P_b1066599e78477
3	E_000008a8ba4f48	Turkcell	37.844510	27.844202	Adnan Menderes Bulvarı	NaN	NaN	NaN	TR	NaN	NaN	Mobile Phone Shops	P_b2ed86905a4cd3
4	E_00001d92066153	Restaurante Casa Cofiño	43.338196	-4.326821	NaN	Caviedes	Cantabria	NaN	ES	NaN	NaN	Spanish Restaurants	P_809a884d4407fb

EDA. Missing values



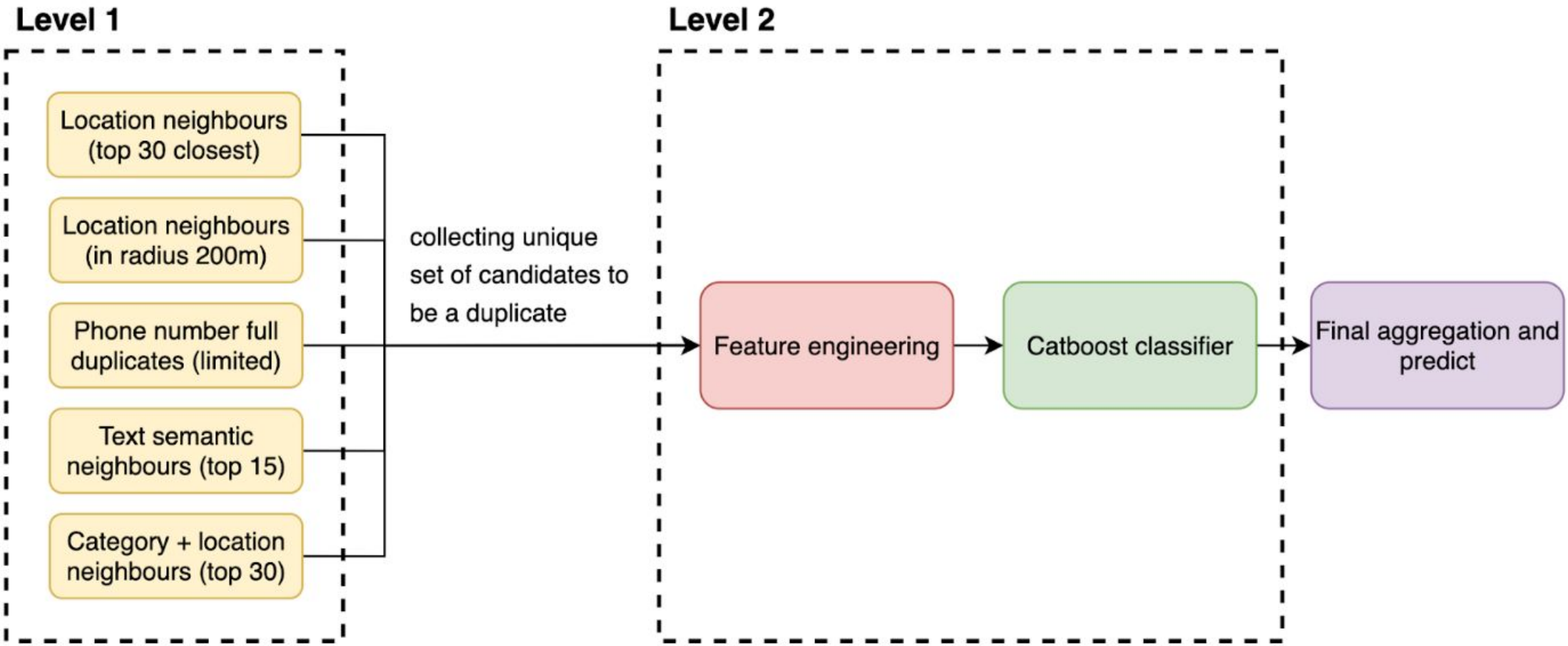
EDA. Counts per country



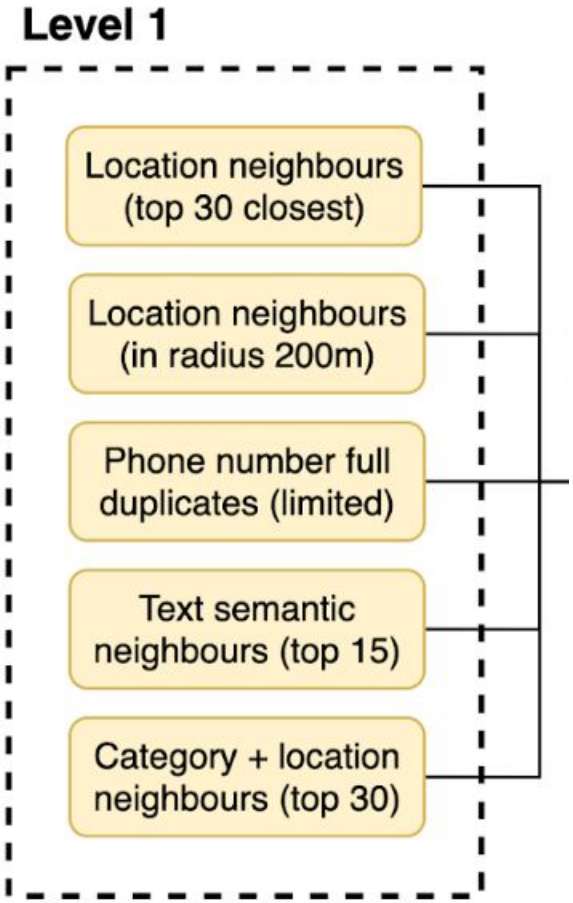
Distribution of points across the globe

System architecture

End-to-end system schema



Model level one. Results



Modelling details:

- Location search: BallTree index
- Text semantic: fine-tuned SentenceTransformers model
- Category semantics: W2V based model
- Semantic search: NMSLIB index

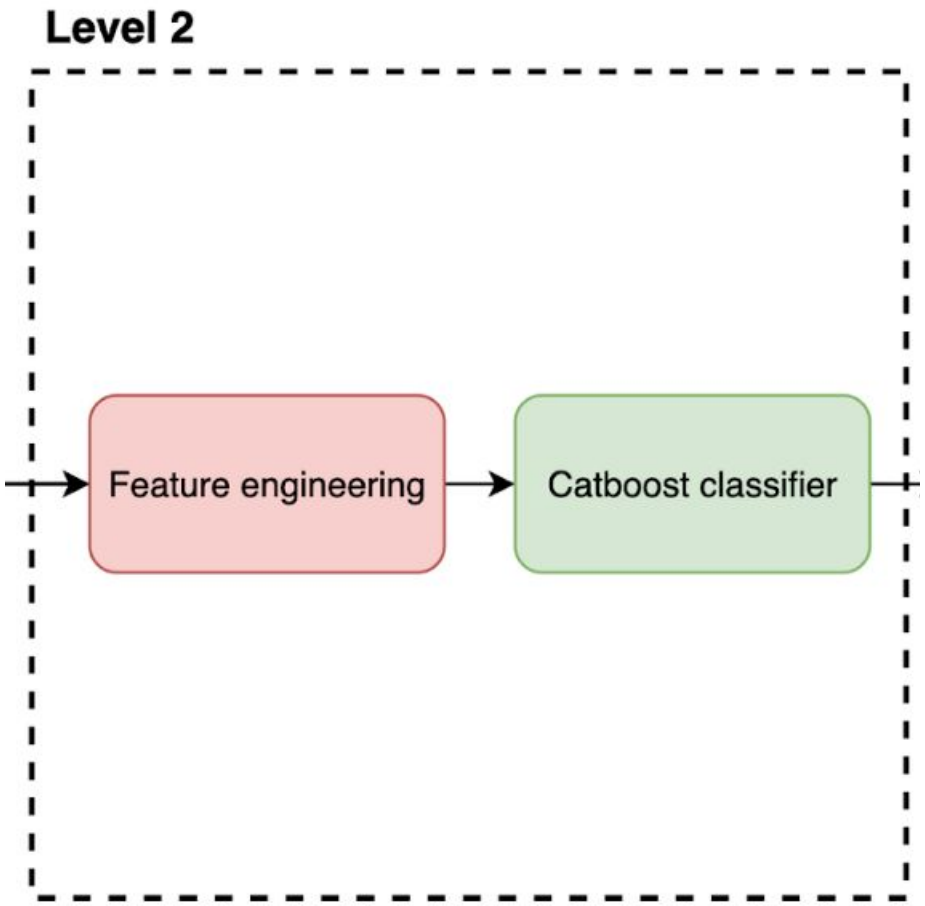
Evaluation metric

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

country code	recall (only location)	recall (only neighborhood)	recall (only category neighbors)	recall (only text similarity)	recall (all together)
(weighted)	0.9227	0.8947	0.9330	0.8869	**0.9786**

**metrics are different for different countries. Full results can be found in the paper*

Model level two. Results



Modelling details:

- Pair of records classification
- Feature engineering: ~45+ features
- Binary classifier per country: CatBoost
- Metric to evaluate end-to-end solution: IoU
- Apply specific post processing

Evaluation metric

The results were evaluated by the mean Intersection over Union (*IoU, aka the Jaccard index*) of the ground-truth entry matches and the predicted entry matches.

$$IoU = \frac{P(\{predicted\_duplicates\} \cup \{true\_duplicates\})}{P(\{true\_duplicates\})},$$

IoU before post processing **= 0.888, IoU** after post processing **= 0.905**

**metrics are different for different countries. Full results can be found in the paper*

Discussion. Further work

- Data specifics are important. Detailed EDA is needed prior to modeling
- Distributed computing should be investigated to tackle the problem of big data
- Multistage modeling leads to error accumulation -> alternatives should be tested

Thank you for attention

#standwithukraine

Q&A