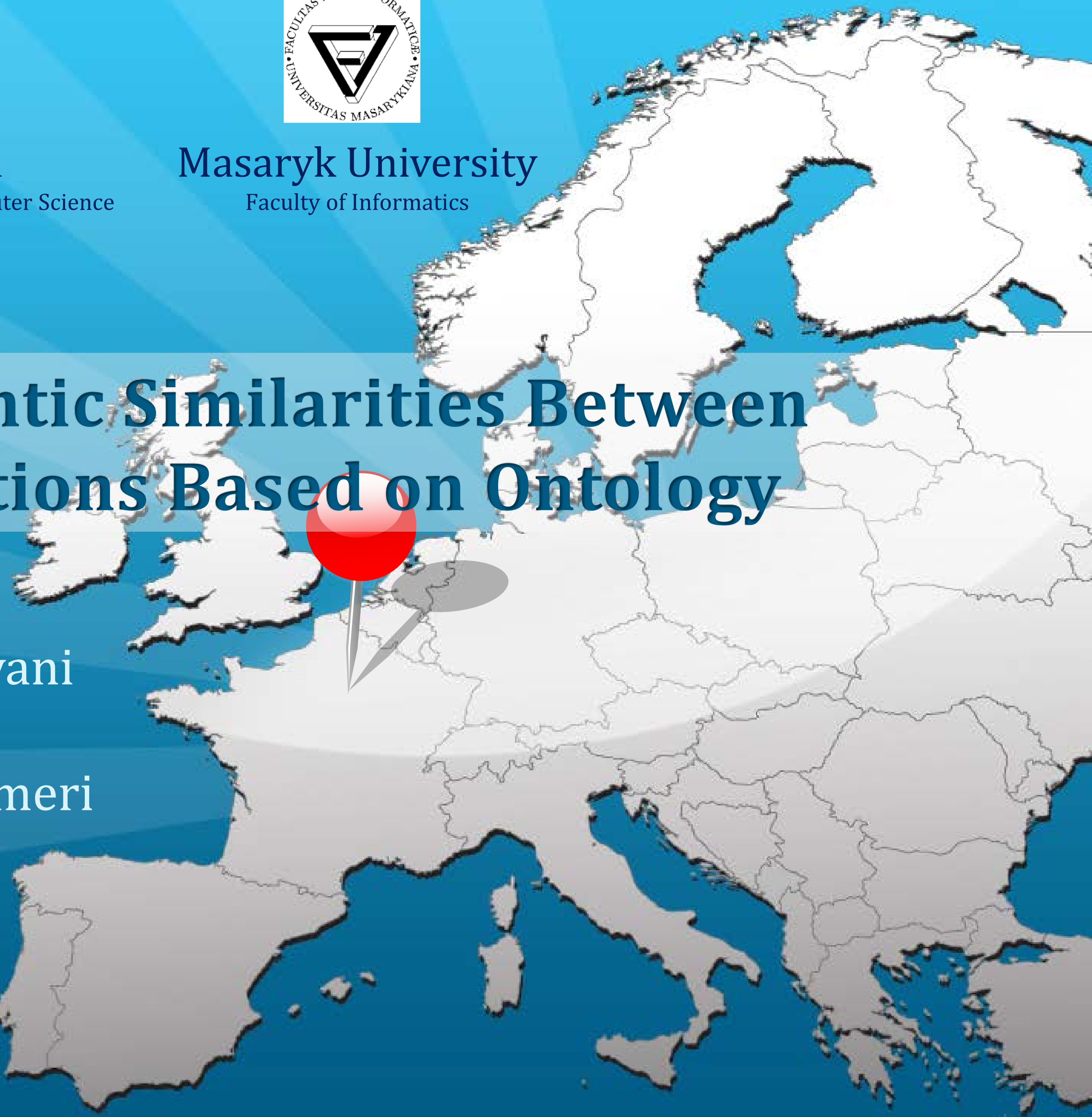
Università della Calabria
Department of Mathematics & Computer Science

Masaryk University
Faculty of Informatics

Semantic Similarities Between Locations Based on Ontology



Moiz Khan Sherwani
Dr. Petr Sojka
Dr. Francesco Calimeri

Question:

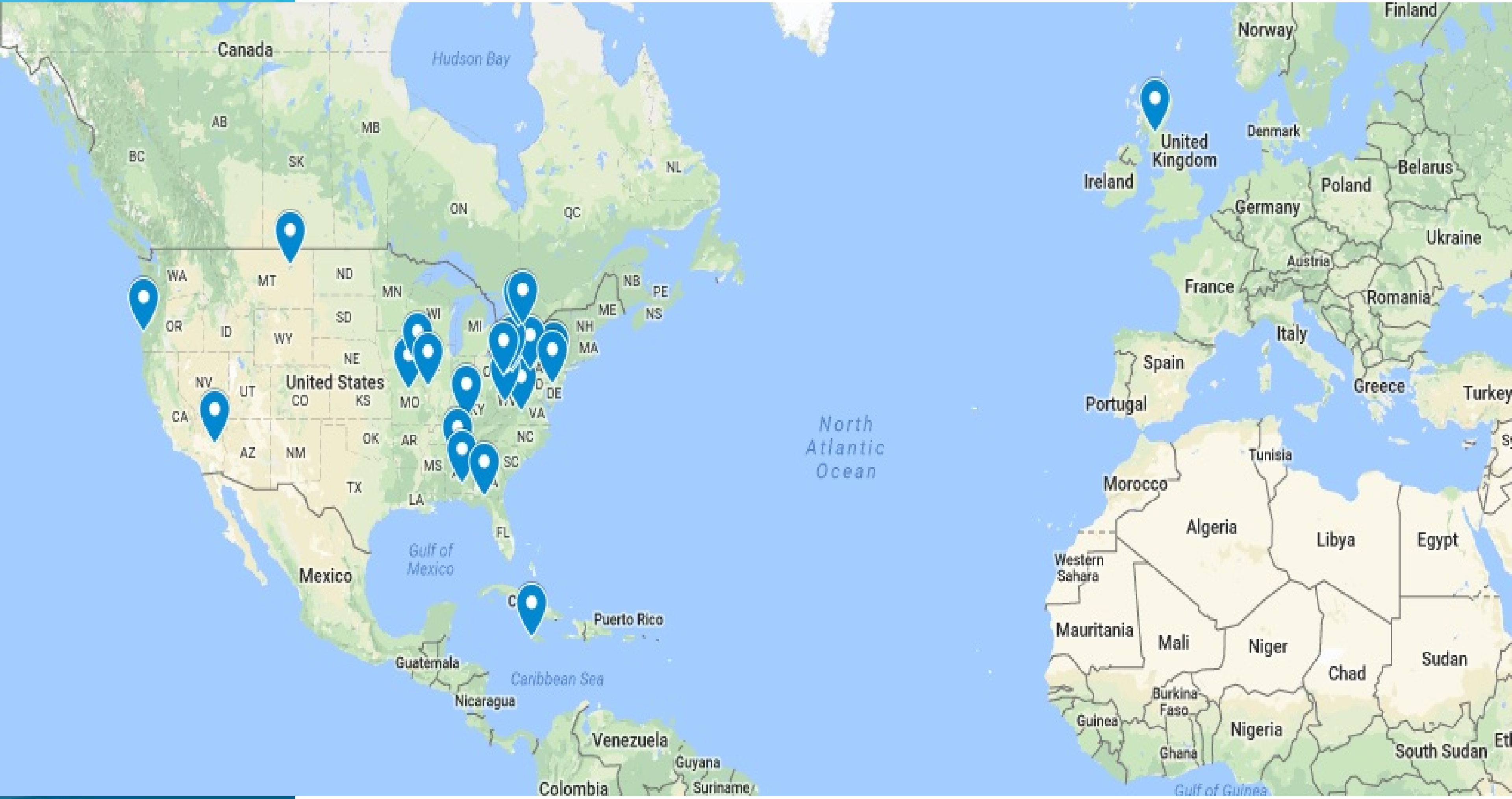
Do you like to travel often?

What do you do before travelling to a new place?

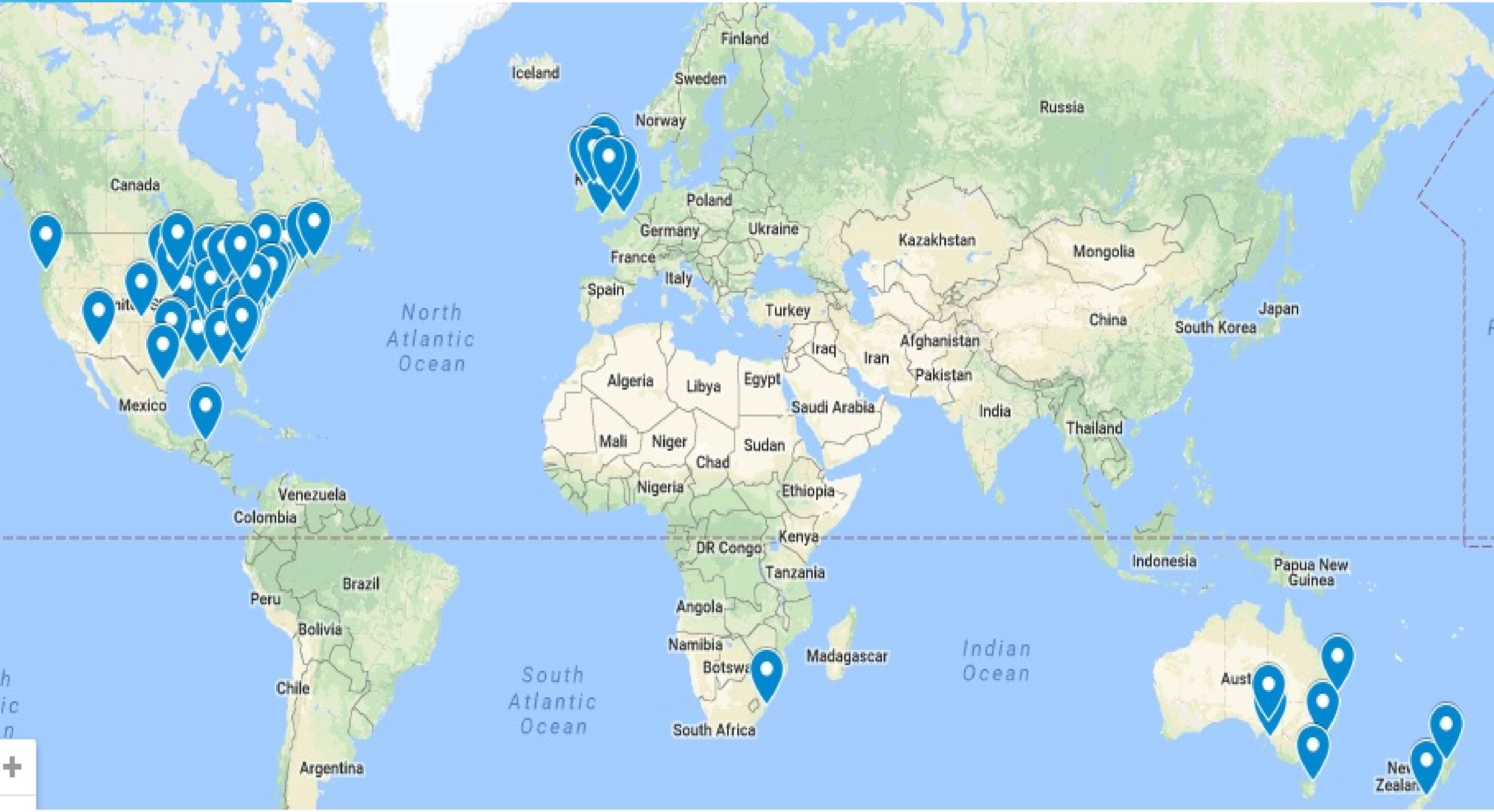
GOAL:

- Disambiguate toponym with highest precision.
- Create an interactive map based news retrieval system.

Problem: Which Glasgow?



Problem: Which Springfield?



What is Toponym & Gazetteer?

Toponym:

- It is a general name of any place or geographical entity.

Gazetteer:

- It is a geographical dictionary.
- It contains all information about the location (physical features).

Main focus of this Research

To carry out the toponym disambiguation, our main focus of this research was based on:

- **Semantic Similarities**
- **Ontology-based Approach**

To carry out the research based on these approaches, we have considered to work with the **Graph Database** instead of Relational Database.

What is Semantic Similarity & Ontology?

Semantic Similarity

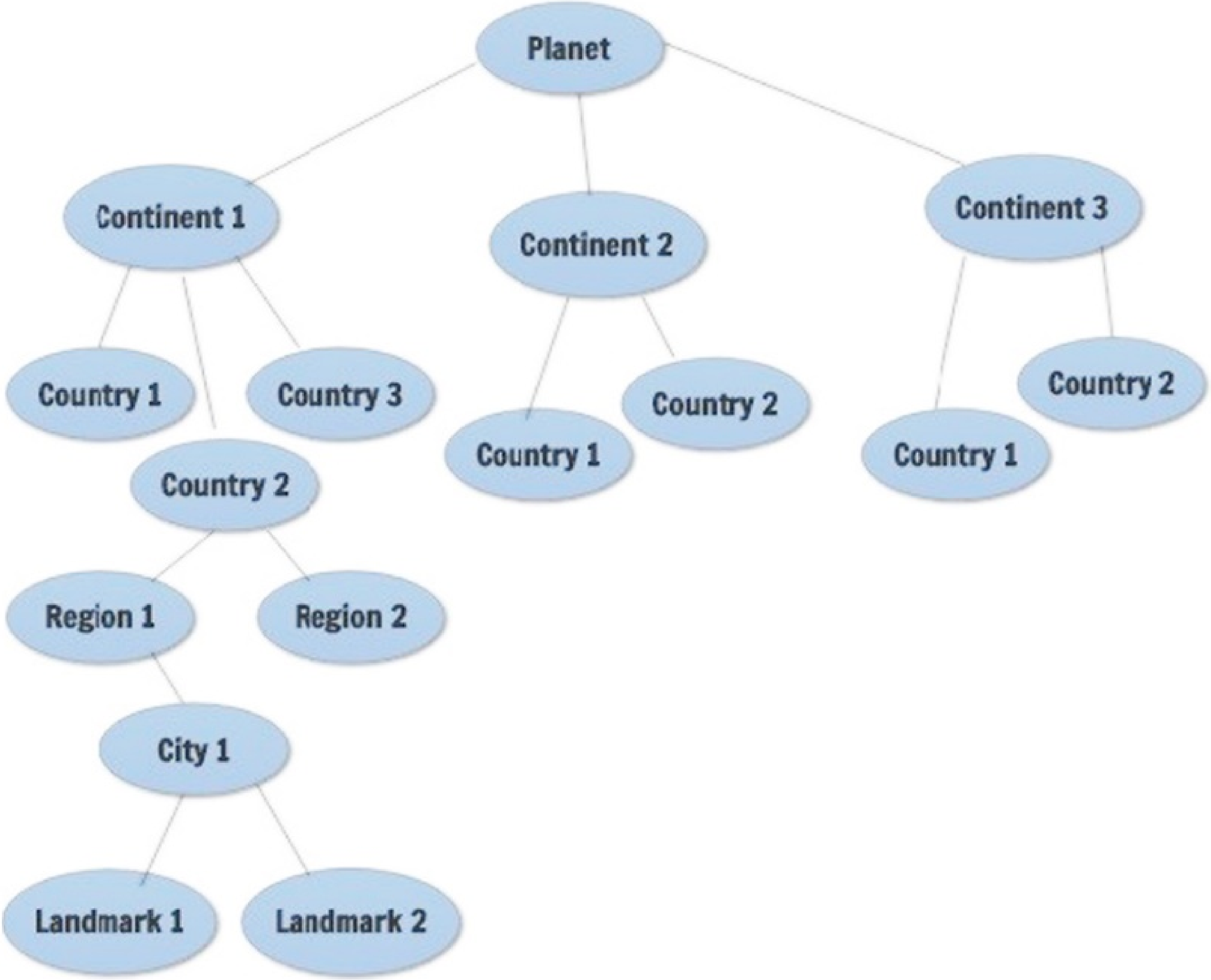
- Defines resemblance between two words
- Similar and dissimilar entries are related by lexical relationships
- Humans can judge easily unlike computers

Ontology

- An **ontology** is a formal naming and definition of the types, properties, and interrelationships of the entities.

- Ontologies are created to limit complexity and to organize information.

Geo-Ontology?



Datasets

- GEO-CLEF
  - 169,477 news articles that contains 1,238,686 toponym occurrences in the articles.

- Gazetteer (Geographical dictionary) sources:
  - GeoNames → over 10,000,000 geographical names corresponding to over 7,500,000 unique features. (latitude, longitude, elevation, population, administrative subdivision and postal codes.)
  - GNS → developed by the U.S. Geological Survey in cooperation with the U.S. Board on Geographic Names.

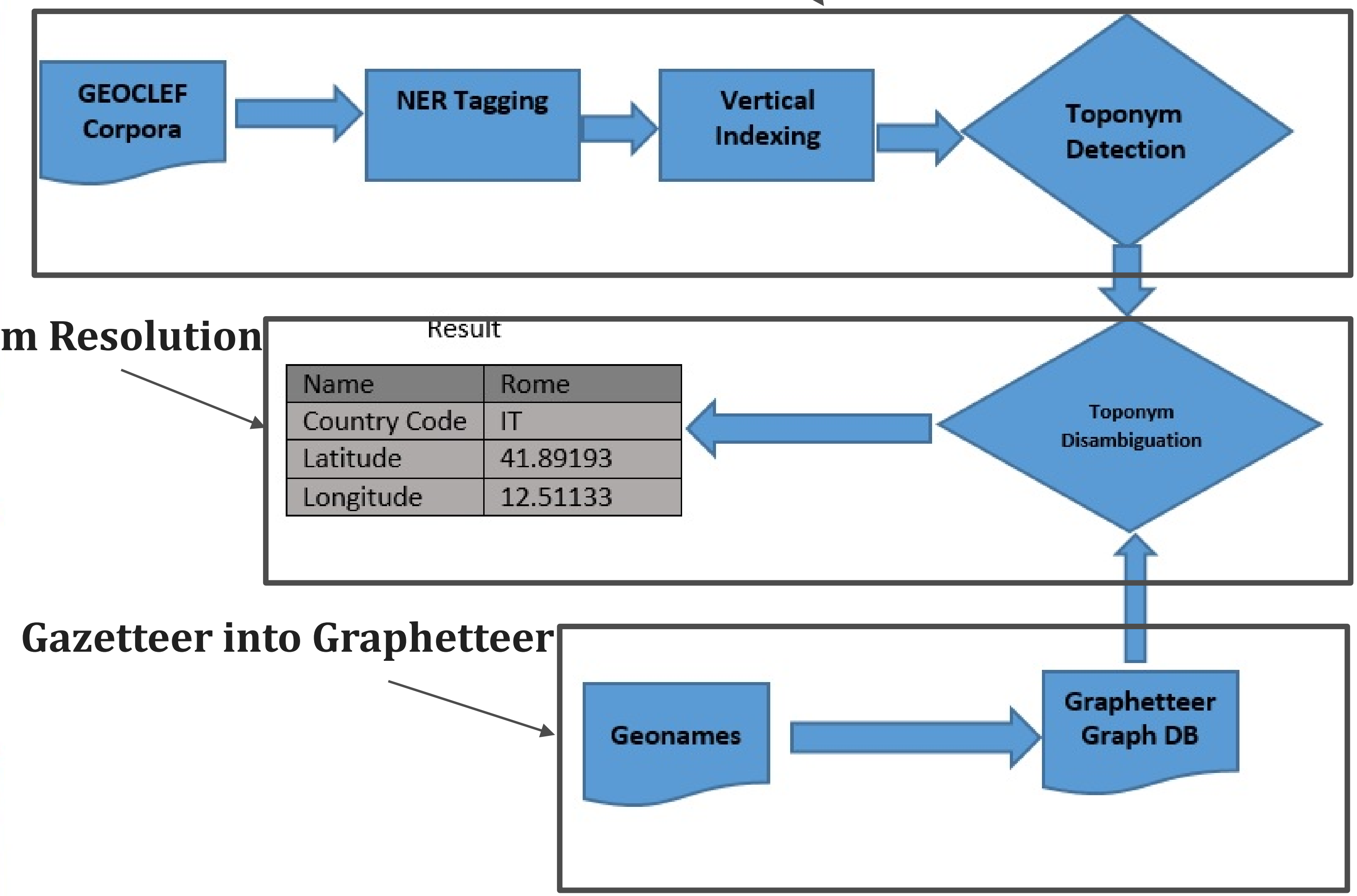
- Stanford NER datasets → It contains the training and test sets to fetch the location names.

Flowchart

Toponym Extraction

Toponym Resolution

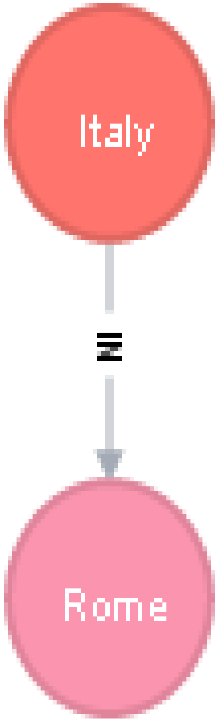
Gazetteer into Graphetteer

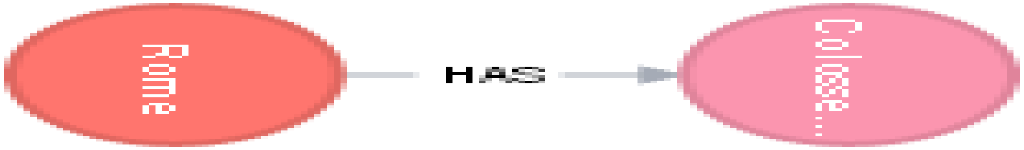


Name	Rome
Country Code	IT
Latitude	41.89193
Longitude	12.51133

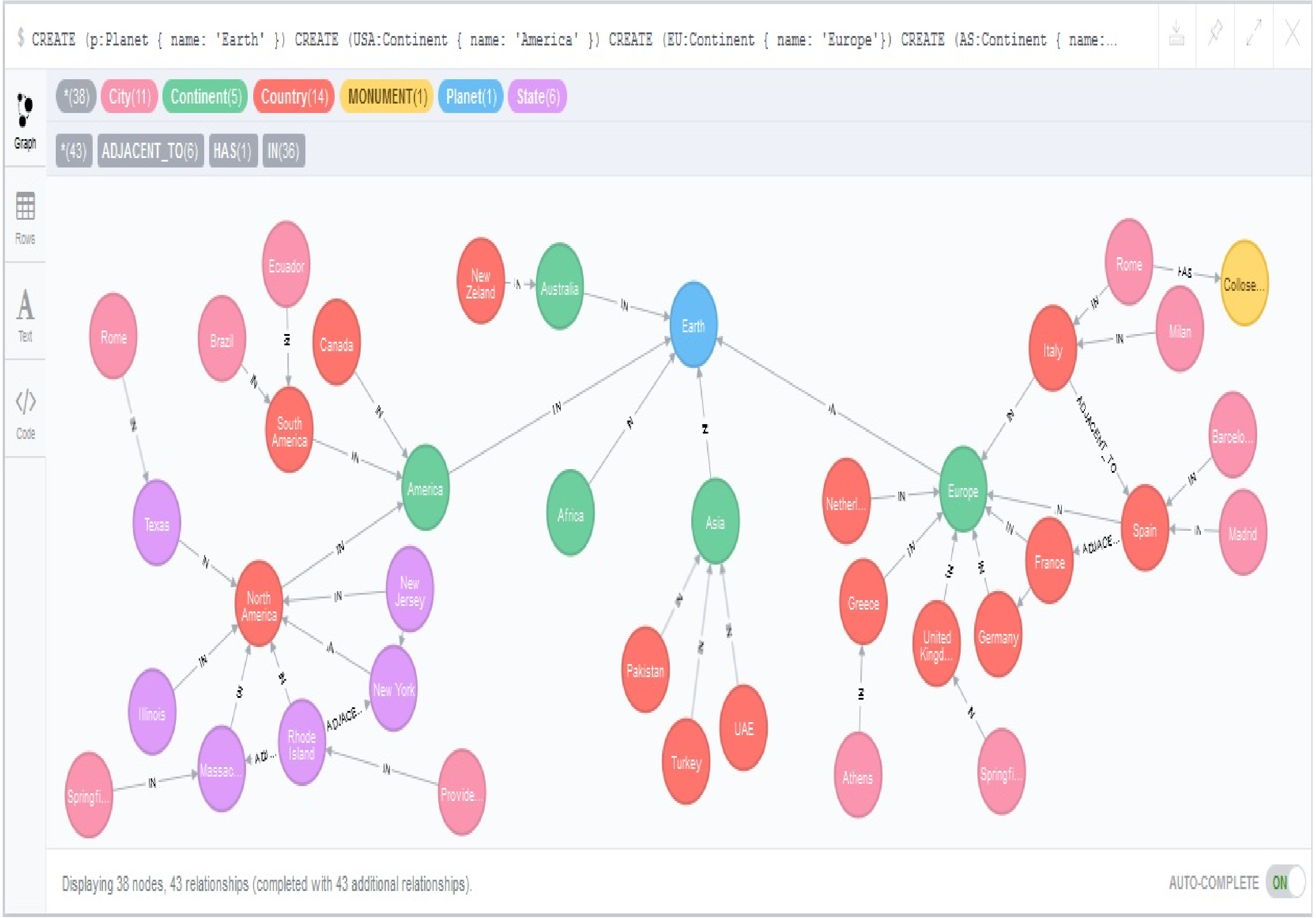
Methodology: Process-1

- **Gazetteer into Graphetteer:**
  - Gazetteer taken from the GeoNames and GNS website is converted into the graph database using Cypher Query language or Py2neo.
  - Application used for graph database is NEO4J.





Contd.



```
$ CREATE (p:Planet { name: 'Earth' }) CREATE (USA:Continent { name: 'America' }) CREATE (EU:Continent { name: 'Europe'}) CREATE (AS:Continent { name:...
```

*(38) City(11) Continent(5) Country(14) MONUMENT(1) Planet(1) State(6)

*(43) ADJACENT_TO(6) HAS(1) IN(36)

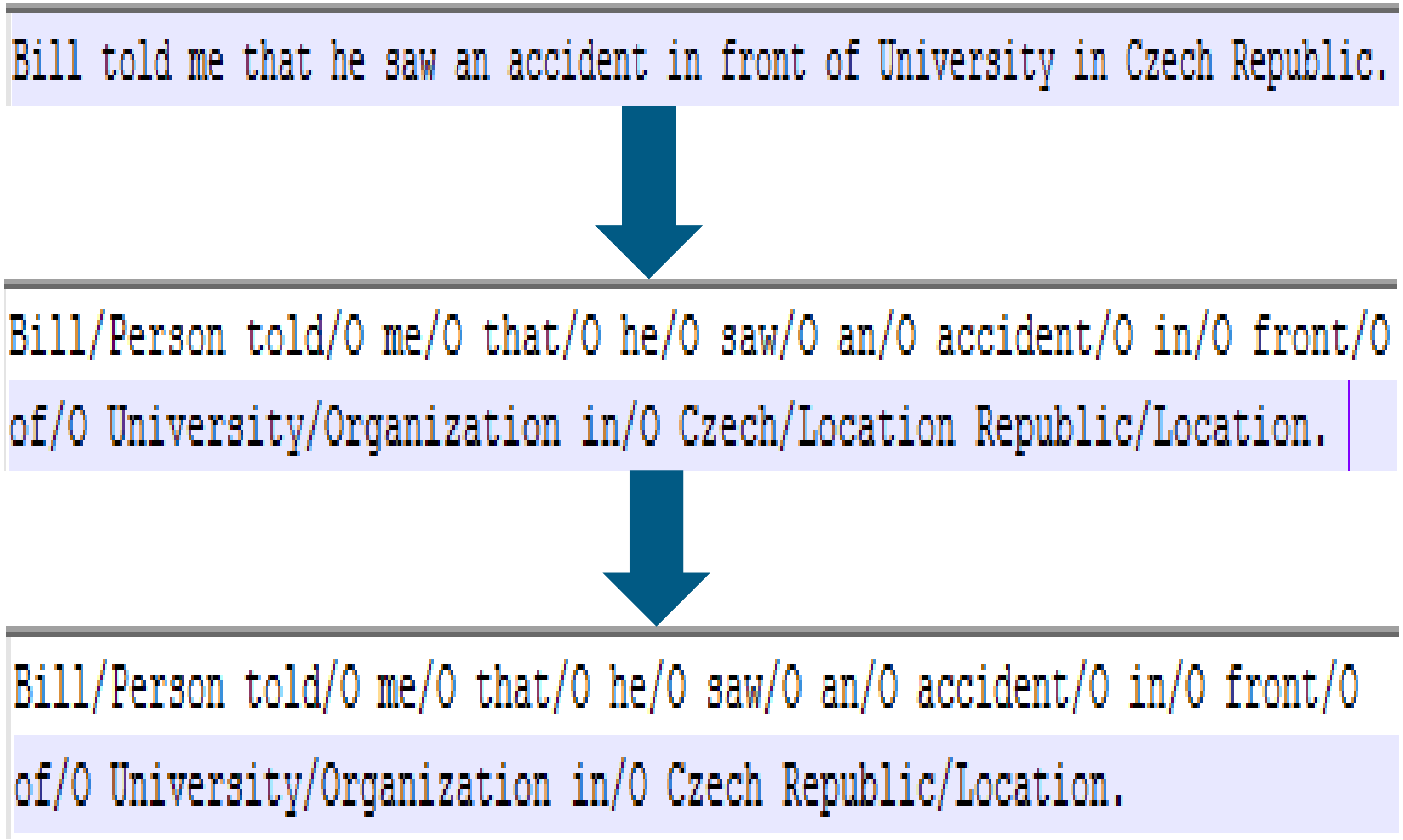
Displaying 38 nodes, 43 relationships (completed with 43 additional relationships).

AUTO-COMPLETE ON

Methodology: Process-2

- **Toponym Extraction from the Articles:**
  - Geo-CLEF corpus is tagged using NER Tagger (Stanford NER tool).
  - Vertical indexing for each word is performed.
  - All the location names are fetched out with the number of occurrences.

Contd.

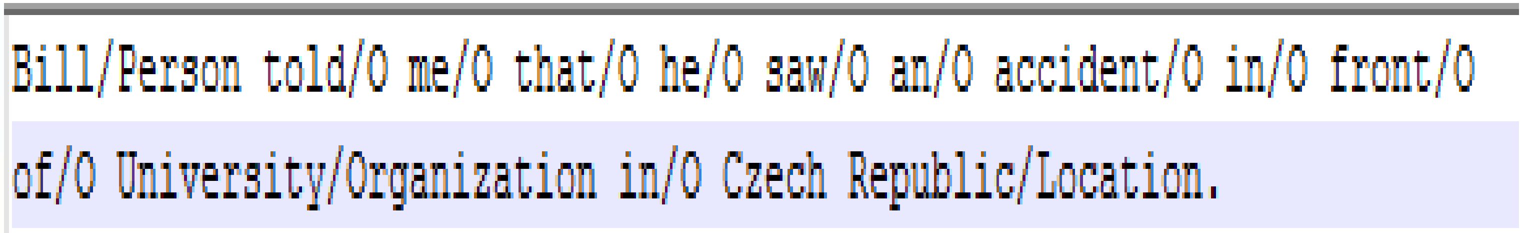


Bill told me that he saw an accident in front of University in Czech Republic.

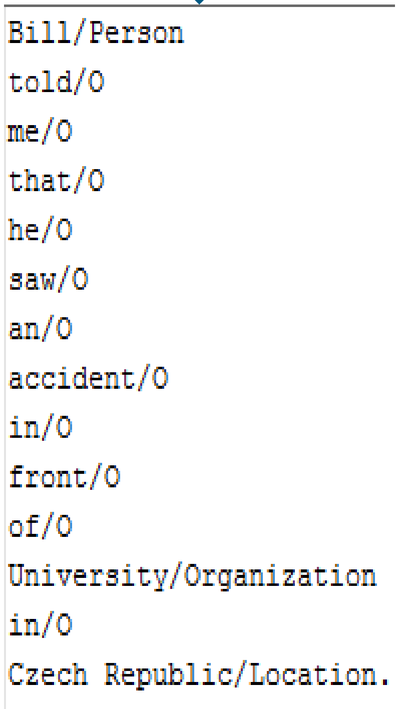
Bill/Person told/O me/O that/O he/O saw/O an/O accident/O in/O front/O of/O University/Organization in/O Czech/Location Republic/Location.

Bill/Person told/O me/O that/O he/O saw/O an/O accident/O in/O front/O of/O University/Organization in/O Czech Republic/Location.

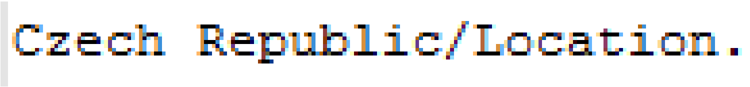
Contd.



```
Bill/Person told/O me/O that/O he/O saw/O an/O accident/O in/O front/O
of/O University/Organization in/O Czech Republic/Location.
```



```
Bill/Person
told/O
me/O
that/O
he/O
saw/O
an/O
accident/O
in/O
front/O
of/O
University/Organization
in/O
Czech Republic/Location.
```



```
Czech Republic/Location.
```

Methodology: Process-3

Toponym Resolution :

Previous Researches:

- Leidner (2007) in Toponym Resolution in Text: Disambiguation based on the population and distance.
- Hauptmann (1999) in TR for speech data: Disambiguation based on Countries, Continents reference.
- Weissenbacher (2015) in Knowledge driven geo-spatial location: disambiguation based on population, distance and meta-data approach.

Three evaluation methods are used while performing this step:

Node-based approach: All toponyms are evaluated based on the population property of the location in the database.

Geographic distance approach: All toponyms within an article are paired to find the smallest distance between them.

Edge-based approach: We introduced this approach based on graph database and it uses the shortest distance between the locations including population property in it.

Results

- As per comparison, edge-based approach resulted with highest precision.

Approach	Precision	Recall	F1-Measure
Node-based	0.70	0.89	0.78
Geographic distance-based	0.39	0.89	0.54
Edge-based	0.74	0.89	0.8

Conclusion and Future Work

We have investigated the datasets, and got satisfactory results based on the edge-based methodology.

Future works:

- Vector representation
- Weighting
- Meta data
- Alternate toponym names

Thank You