# FASDH25-portfolio2

A repository for students' portfolios for mini-project 2

Syed Azhar Uddin

# 2B Use stanza to extract all place names from (part of) the corpus

This project aims to automatically identify, count, and map geographical place names found in news articles. By focusing on content written in January 2024, it compares two distinct methods—NER (using Stanza) and a regex+gazetteer approach—highlighting their advantages and limitations.

## 1. Named Entity Recognition with Stanza

Adapted a Colab notebook (Gaza\_NER2\_groupname.ipynb) to use a larger corpus.

Filtered the dataset to include only January 2024 articles.

Used Stanza’s NER module to extract locations only.

Cleaned and normalized names (e.g., “Gaza’s” merged into “Gaza”).

Wrote results into ner\_counts.tsv, showing placename and count.

## 2. Building a Gazetteer

Script build\_gazetteer.py reads ner\_counts.tsv.

Used geocoding (via geopy or API) to find coordinates.

Output saved as NER\_gazetteer.tsv with placename, latitude, and longitude.

Manual geocoding was done for places not found via automatic methods:

Beit Hanoun

Jabalia refugee camp

Nuseirat

Maghazi

Bureij camp

## 3. Mapping the Places

Visual maps created from the NER and regex-based datasets.

Stored in the maps/ folder as:

ner\_map\_jan2024.png

regex\_map\_jan2024.png

## structure of repository

This project is clearly organized, and the folder structure is explained in this section. The main notebook, Gaza\_NER2\_groupname.ipynb, extracts place names from January 2024 news articles using Stanza. These names and their counts are saved in ner\_counts.tsv. Then, build\_gazetteer.py geocodes the names and saves the coordinates in gazetteer/NER\_gazetteer.tsv. Two maps—one using NER data and one using regex—are stored in the maps folder.

## Regex + Gazetteer

Advantages:

Simple and fast.

High accuracy if the gazetteer list is complete.

Easy to implement with minimal setup.

Disadvantages:

Misses place names that are not in the gazetteer.

cannot handle different spellings or variations.

Requires manual updates to the list.

## NER (Named Entity Recognition using Stanza)

Advantages:

Can detect new and unknown place names.

Understands context and can work across different sentence structures.

Does not rely on a fixed list.

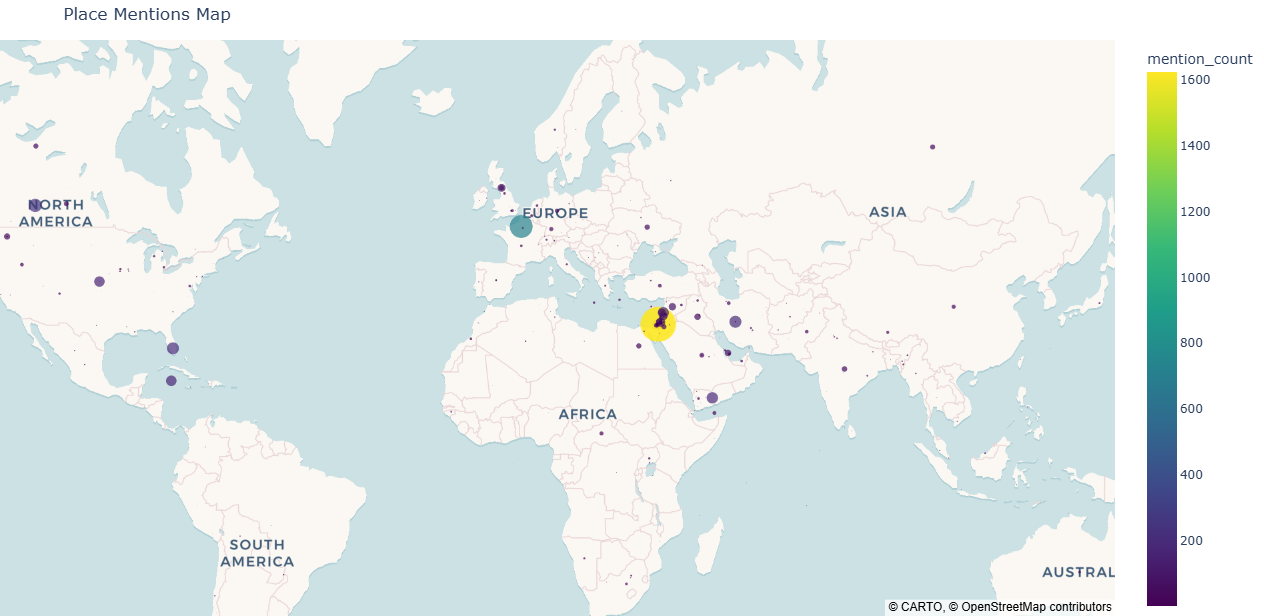
Disadvantages:

Slower and more resource intensive.

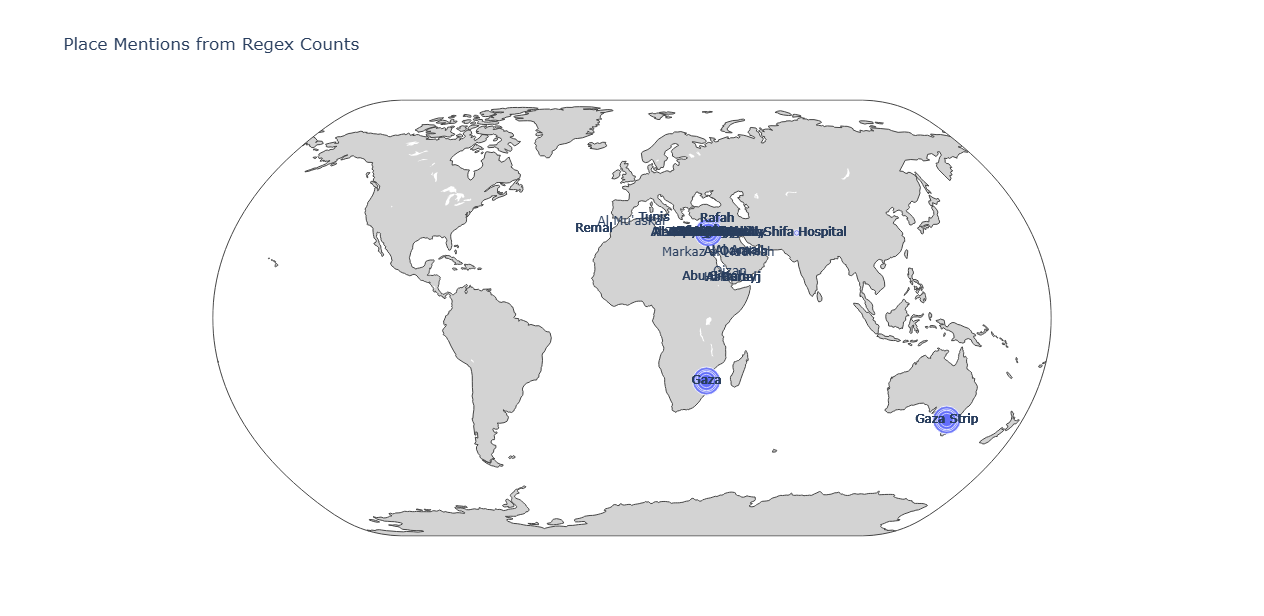
Sometimes extracts incorrect or irrelevant entities.

Requires post-processing to clean the output.

NER Map



Regex Map



Style used for regex map and why

The map generated in this project uses a **natural earth projection** with a **scatter geo style**, which plots geographical coordinates as individual points on a global map. This style was chosen because it effectively visualizes the spatial distribution of place name mentions across different regions. The **natural earth projection** provides a visually balanced and proportionally accurate view of the continents, making it ideal for global datasets. The **scatter plot** format allows each place to be marked with a circle whose size corresponds to its frequency of mention, enabling quick visual comparison of prominence. This style is both intuitive and informative, making it easy for viewers to interpret the density and spread of place references without overwhelming them with unnecessary geographic detail. It’s particularly suitable for digital humanities work, where clarity, simplicity, and contextual insight into textual geographies are essential.