EGU22-8323
https://doi.org/10.5194/egusphere-egu22-8323
EGU General Assembly 2022


# Multi-attribute geolocation inference from tweets

**Umair Qazi**, Ferda Ofli, and Muhammad Imran
Qatar Computing Research Institute, Social Computing, Doha, Qatar (uqazi@hbku.edu.qa)

Geotagged social media messages, especially from Twitter, can have a substantial impact on decision-making processes during natural hazards and disasters. For example, such geolocation information can be used to enhance natural hazard detection systems where real-time geolocated tweets can help identify the critical human-centric hotspots of an emergency where urgent help is required.

Our work can extract geolocation information from tweets by making use of five meta-data attributes provided by Twitter. Three of these are free-form text, namely tweet text, user profile description, and user location. The other two attributes are GPS coordinates and place tags.

Tweet text may or may not have relevant information to extract geolocation. In the cases where location information is available within tweet text, we follow toponym extraction from the text using Named Entity Recognition and Classification (NERC). The extracted toponyms are then used to obtain geolocation information using Nominatim (which is open-source geocoding software that powers OpenStreetMap) at various levels such as country, state, county, city.

Similar process is followed for user profile description where only location toponyms identified by NERC are stored and then geocoded using Nominatim at various levels.

User location field, which is also a free form text, can have mentions of multiple locations such as USA, UK. To extract location from this field a heuristic algorithm is adopted based on a ranking mechanism that allows it to be resolved to a single point of location which can be then mapped at various levels such as country, state, county, city.

GPS coordinates provide the exact longitude and latitude of the device's location. We perform reverse geocoding to obtain additional location details, e.g., street, city, or country the GPS coordinates belong to. For this purpose, we use Nominatim's reverse API endpoint to extract city, county, state, and country information.

Place tag provides a bounding box or an exact longitude and latitude or name information of location-tagged by the user. The place field data contains several location attributes. We extract location information from different location attributes within the place using different algorithms. Nominatim's search API endpoint to extract city, county, state, and country names from the Nominatim response if available.

Our geo-inference pipeline is designed to be used as a plug-in component. The system spans an elasticsearch cluster with six nodes for efficient and fast querying and insertion of records. It has already been tested on geolocating more than two billion covid-related tweets. The system is able to handle high insertion and query load. We have implemented smart caching mechanisms to avoid repetitive Nominatim calls since it is an expensive operation. The caches are available both for free-form text (Nominatim's search API) and exact latitude and longitude (Nominatim's reverse API). These caches help reduce the load on Nominatim and give quick access to the most commonly queried terms.

With this effort, we hope to provide the necessary means for researchers and practitioners who intend to explore social media data for geo-applications.