Eliciting fuzzy location data from social media posts with Natural Language Processing

Gullick, D,S*[1], Whyatt, J,D,†[1] and Richardson, J,W‡[1]

[1]Lancaster Environment Centre, Lancaster University, UK

March 27, 2018

Summary

Social Media Platforms such as Twitter are collecting large volumes amounts of user generated content every day, much of which is location aware. Whilst there are clear use cases for data harvested from these platforms in research, harvesting and analysis of these datasets represents a substantial challenge. Existing work exploits geotagged or geocoded metadata collected alongside user generated content (and the challenges that accompany its use). However, the majority of user generated content lacks explicit locational information, but may still contain location information albeit in a less explicit, or "fuzzy", form such as textual descriptions in the main body of the social media posting. This work explores the analysis of such datasets, and presents a novel generalisable methodology for extracting implicit of location data from such data.

KEYWORDS: Geocoding, Geotagging, Fuzzy Location Data, Social Networking, Twitter, Natural Language processing;

1 Introduction

AT GISRUK 2017, the authors presented a novel tree failure modelling technique named Tree Risk Evaluation Environment for Failure and Limb Loss (TREEFALL) which aimed to quantify the risk failing trees posed to nearby critical infrastructure in severe weather events. Unfortunately, the data needed to validate such predictions (a database of where and when trees failed historically) does not exist. In an effort to create such a database, alternative socially generated sources were considered.

Social media platforms such as Facebook, Twitter, LinkedIn or Flickr are all Web 2.0 internet based applications facilitating the collection and communication of user generated content. For example,

*d.gullick@lancaster.ac.uk

†d.whyatt@lancaster.ac.uk

‡j.w.richardson@lancaster.ac.uk

Twitter allows registered users to send short messages, named "tweets" consisting of up to 280 characters (historically 140) to multiple other users. These messages can contain links to images, mention other users, and contain unique tags that may denote relevent themes for the tweet (called "hashtags"). These social media platforms process large volumes of user generated content: Twitter alone generates over 500 million "tweets" per day. [1].

Many of these social media platforms are becoming increasingly location aware, and often give the user the ability to tag their uploaded content with extra information regarding relevant locations either by "geotagging" (usually in the form of a GPS coordinate) or "geocoding" (converting an address or place into a location). Users can also supply information regarding their location in their public personal profile.

Whilst there are clear ways that datasets such as those generated by Twitter users can be an asset to a multitude of research questions, especially regarding complex on-line social interactions, disaster event detection (Krumm & Horvitz, 2015), language analysis (Morstatter, Lubold, Pon-Barry, Pfeffer, & Liu, 2014), or location estimation (Gonzalez, Figueroa, & Chen, 2012; Ikawa, Enoki, & Tatsubori, 2012; Kinsella, Murdock, & O'Hare, 2011; Matsuo, Shimoda, & Yanai, 2017) - using them is not without caveats. Unlike Volunteered Geographic Information (VGI) platforms such as OpenStreetMap[2] where geographic data is contributed voluntarily by individuals (Goodchild, 2007), the main purpose of platforms such as Twitter is the collation of user generated content, such as pictures and text, and not geographic data. Some of this content is accompanied by uncurated geographic data at varying degrees of specificity (e.g place names or road names embedded within the tweet).

Much of the previous work in the area explores the caveats working with the geotagged information, such as false hotspotting (Huck, Whyatt, & Coulton, 2012). However, the vast majority of tweets (99%) are neither geotagged or geocoded and are therefore excluded from most analyses. However, these tweets often contain information pertaining to location; such as textual descriptions of a place, or road name. This information is harder to elicit, and in most cases is less precise than either geotagged or geocoded content, but is still valuable information that can be used for analysis.

As such, location within these tweets can be considered on a scale of explicit (coordinate geotag) to implicit (mention of a place within a text) and precise (point coordinate) to imprecise (e.g the name of a town or county).

The analysis of these datasets represents a substantial challenge given its position intersecting multiple disciplines such as geography, computational social sciences, linguistics, and computer science. Another challenge, given the implicit nature of location data collected using this method is to extract and georeference locational data of different precisions from the content of the social media posting.

This project, Social Tree UpRooting DirectorY (STURDY), primarily aims to collect and analyse real time tweets in order to build a database of tree failure events which the aforementioned

[1] See: https://www.dsayce.com/social-media/tweets-day/

[2] See: https://www.openstreetmap.org/

TREEFALL system can be validated. In working to do so, an automated methodology that harvests data not only from relevant geotagged or geocoded tweets, but also relevant non geotagged or geocoded tweets through a process of Natural Language Processing (NLP), (the computerised analysis of text considering the symantic rules that govern the language) and geocoding (the process of translating a description of a place to a location on the earth's surface) is proposed.

2 Methodology

The official Twitter Application Programming Interface (API) for python called "Tweepy"[3] facilitated the collection of 18 million tweets between October 2016 and October 2017, each of which contain specific key words related to tree failure. Of this 18 million, few are relevant to TREEFALL and the area of interest. To focus the content of this dataset on relevant tweets within the UK, a filter process was carried out for each tweet, as shown in Figure 1.

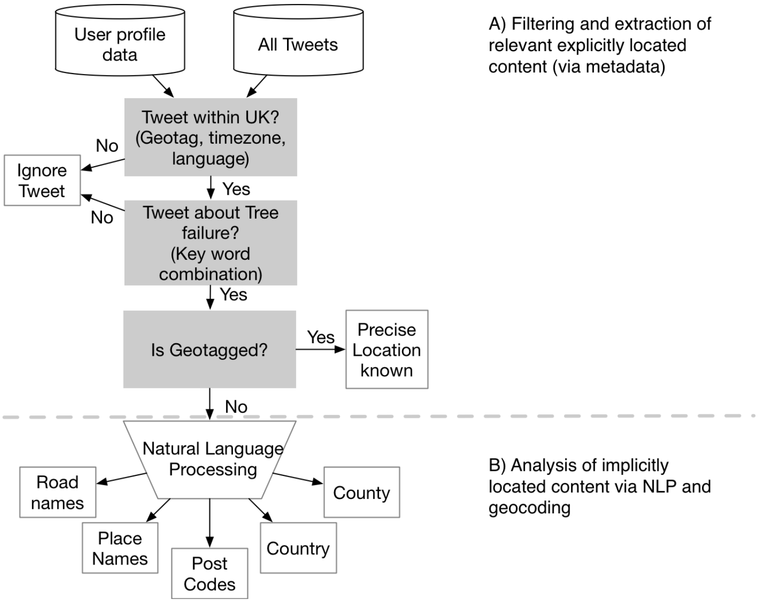

User profile data
All Tweets
A) Filtering and extraction of relevant explicitly located content (via metadata)
Tweet within UK? (Geotag, timezone, language)
No
Ignore Tweet
Yes
No
Tweet about Tree failure? (Key word combination)
Yes
Is Geotagged?
Yes
Precise Location known
No
Natural Language Processing
Road names
County
B) Analysis of implicitly located content via NLP and geocoding
Place Names
Post Codes
Country

Figure 1: A diagram showing the geoprocessing chain for explicitly and implicitly locating tweets.

Firstly, tweets originating outside the UK are removed. Where possible, geotags (Latitude and Longitude) are used to locate each tweet. Where this is not possible, a combination of spoken language, timezone settings and home location data taken from the users home profile are used to assess the likelihood of the tweet originating in the UK. Tweets considered to be within the UK are then evaluated for relevance to tree failure through key word searches e.g "Tree blocking" in combination with "Road" or "Path".

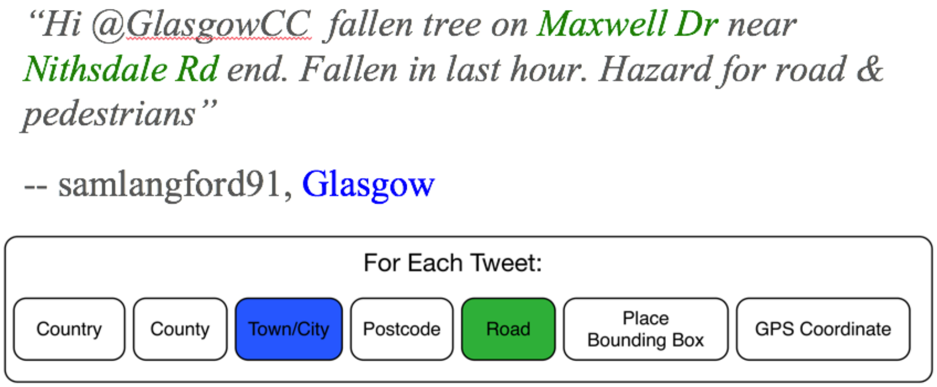
[3] See: http://www.tweepy.org/

The resulting tweets are then evaluated for location; those that are geotagged or geocoded can easily be located. Those that are not are processed using the NLP engine OpenNLP [4].

OpenNLP analyses the body (content) of the tweet and highlights words that are likely to be the names of “places” or “things” given the semantic rules that govern the english language. These identified places are then compared to a gazetteer of known entities, such as roads and town names as a means of locating the tweet. For each tweet any successful matches are recorded. The accuracy of the final location is dependant on the nature of features or places mentioned in the tweet, e.g Lancaster or the M6 motorway. Further spatial refinement is possible through triangulation of information (tweets) around a single event (i.e multiple reports of tree failure in Lancaster, some more locationally specific that others).

3 An Example

Locating the varying specificity of the location data identified using this process represents substantial challenge. Coordinates supplied in geotagged tweets are generally be considered precise (although not always representative). Most tweets are much less specific, sometimes only refering to a county, or road name. These tweets also typically match an area rather than a specific point. Similarly, multiple places may share names. Consider the following tweet:



“Hi @GlasgowCC fallen tree on Maxwell Dr near Nithsdale Rd end. Fallen in last hour. Hazard for road & pedestrians”

-- samlangford91, Glasgow

Figure 2: An example tweet with identified locatable entities.

In this example, three locatable entities are mentioned: “Glasgow” as a city name, and “Maxwell Drive” and “Nithsdale Road” as road names. As a human interpreter it is possible to say that most likely the tweet references a junction between these two roads in Glasgow. However a computer lacks common sense and local knowledge and considers these as separate named entities with no apparent hierarchy or relation. Therefore a proxy such as minimum distance between all occurrences of named entities can be used to determine the most probable location. Figure 3 shows this case as the junction of Nithsdale Road and Maxwell Drive in Glasgow.

[4] See: https://opennlp.apache.org/

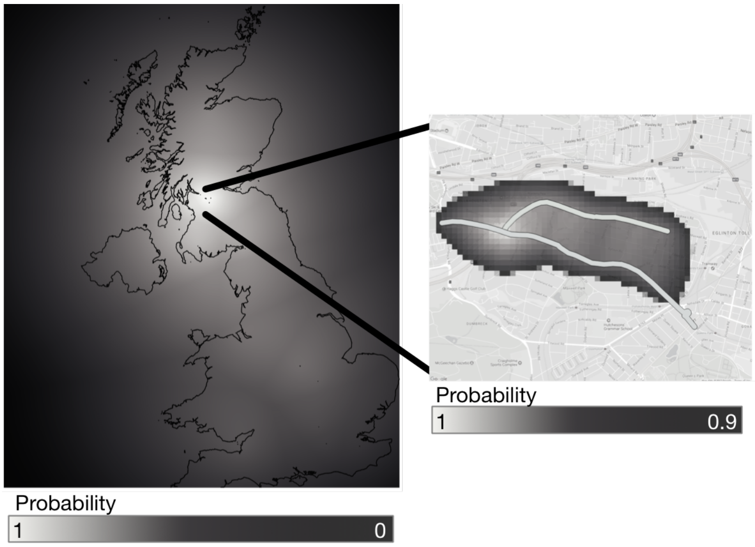


Figure 3: Resulting probability maps. Left: Whole UK map, right: close-up map with reduced probability limits.

This process is automatically repeated for each tweet, and results in probability maps for each tweet respectively. To date this process has yielded approximately 1400 tree related tweets in addition to those already geotagged or geocoded, and therefore extends the validation database with which TREEFALL can be compared.

4 Conclusion and Future work

Although built with focus on tree failure, this novel methodology can be applied to other areas of interest within the research community and facilitates access to the less explicit location data contained within twitter or other social media platforms.

Further work will focus on improvements to the methodology (short hand and abbreviation recognition), event identification through spatio-temporal triangulation of tweets (e.g. tracking tree failure events over the course of a major storm) and the development of a rule base to exploit fuzzy geographical references (.e.g “near the A6”, or “close to Lancaster”)

5 Biography

David Gullick is a Senior Research Associate at the Lancaster Environment Centre, Lancaster University, UK. His work focuses on Environmental Modelling, and the analysis of spatial data across the field.

Joseph Richardson is a Master student studying computer science at Lancaster University with a keen interest in the environmental sciences. His work focuses on text analysis and natural language processing.

Duncan Whyatt is a Senior Lecturer at Lancaster University with a background in Geography & Computer Science. He uses Geographic Information Systems (GIS) to visualise, integrate and analyse spatial data from a variety of different sources across the spectrum of Human and Physical Geography, Ecology and Environmental Science.

References

Gonzalez, R., Figueroa, G., & Chen, Y.-S. (2012). Tweolocator: a non-intrusive geographical locator system for twitter. In *Proceedings of the 5th acm sigspatial international workshop on location-based social networks* (pp. 24–31).

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211–221.

Huck, J., Whyatt, D., & Coulton, P. (2012). Challenges in geocoding socially-generated data.

Ikawa, Y., Enoki, M., & Tatsubori, M. (2012). Location inference using microblog messages. In *Proceedings of the 21st international conference on world wide web* (pp. 687–690).

Kinsella, S., Murdock, V., & O'Hare, N. (2011). I'm eating a sandwich in glasgow: modeling locations with tweets. In *Proceedings of the 3rd international workshop on search and mining user-generated contents* (pp. 61–68).

Krumm, J., & Horvitz, E. (2015). Eyewitness: Identifying local events via space-time signals in twitter feeds. In *Proceedings of the 23rd sigspatial international conference on advances in geographic information systems* (p. 20).

Matsuo, S., Shimoda, W., & Yanai, K. (2017). Twitter photo geo-localization using both textual and visual features. In *Multimedia big data (bigmm), 2017 ieee third international conference on* (pp. 22–25).

Morstatter, F., Lubold, N., Pon-Barry, H., Pfeffer, J., & Liu, H. (2014). Finding eyewitness tweets during crises. *arXiv preprint arXiv:1403.1773*.