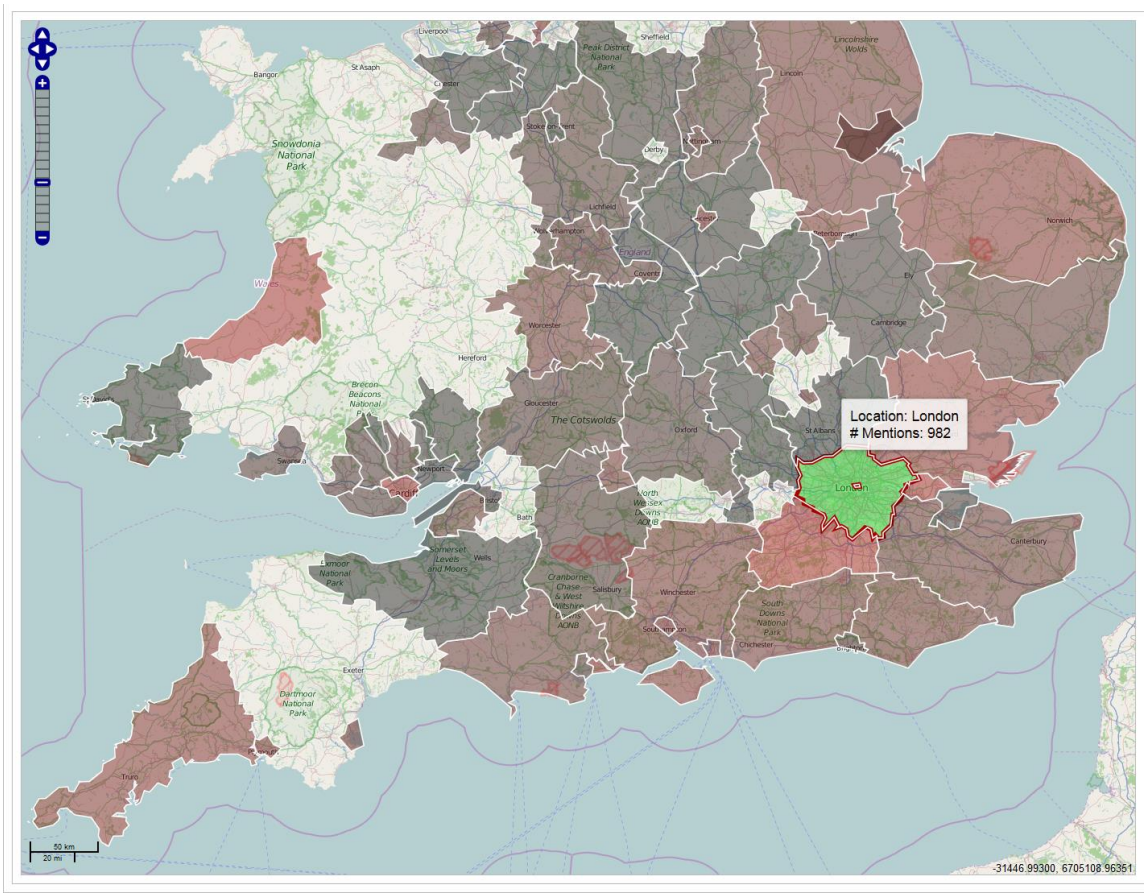
2nd Dec 2016

Scalable Python Geoparse Library Released - geoparsepy

Geoparsing is the process of extracting a location from text. It's a very useful thing for many applications including social media analysis. Manual inspection of metadata from 1000's of Tweets to find a location reference is not very practical - automatic geoparsing is often the solution!



Why would you need to extract text from a Tweet when it's got a geotag?

Typically only about 1.5% of tweets have a geotag, and of this 1.5% about half are actually not real geotags at all, but rather are default coordinates from the user's Twitter profile home location. Also a geotag tells you where someone is when they Tweet, not what they are referring to in their message. A user with a geotag in the USA might tweet something really important about Paris, but the geotag will still be in the USA. This makes relying only on the geotag a bad idea.

Are there free HTTP geocoding services that do this already?

Yes there are! Services such as Google Geocoding API, OpenStreetMap Nominatim and Bing Maps API allow you to post a textual phrase and get back a likely location that matches it, along with a longitude and latitude coordinate.

These services are great but they have limited geoparsing throughput and strict rate limits on the numbers of requests per day. They also expect well parsed text inputs and do not make use of any contextual text around a mention of a location.

What if you want to scale up and geoparse 100,000's of social media posts in real-time?

Rate limited HTTP geocoding services are of no use here. You need a map gazetteer or database hosted locally to achieve this level of performance.

One option is to download a gazetteer such as the geonames and use this to lookup location data. The problem with gazetteers is that they usually only provide region level data coupled with point geometry. Real locations need a mixture of point, line and polygon geometry to be represented fully. This means any geoparsing with a gazetteer will only be at a region level and will miss out on important contextual geographic information that might help to disambiguate locations more accurately. Remember there are over ten places called 'London' in the world - you need context to know which one is being referred to!

The other option is to install and use a map database. geoparsepy uses a local OpenStreetMap planet database for geoparsing. This allows it to access the full location geometry, and consequently out-perform [1] [2] existing gazetteer-based geoparsing approaches. In addition, because we have the planet's geography at our disposal we can geoparse at region, street and building levels. This makes geoparsepy enabled solutions a powerful and scalable proposition.

Where can we get it to try it out?

geoparsepy is a free Python library for geoparsing hosted on the Python Package Index (PyPi). It was created by the University of Southampton IT Innovation Centre over a 5 year period under EU FP7 projects TRIDEC (grant agreement number 258723) and REVEAL (grant agreement number 610928). geoparsepy can be used for research, education or evaluation purposes for free. A commercial license is available free of charge on request also.

Acknowledgement

The work presented in this article is part of the research and development in the TRIDEC project (grant agreement number 258723) and REVEAL project (grant agreement 610928), supported by the 7th Framework Program of the European Commission.

About the author

Stuart E. Middleton is a senior research engineer at the University of Southampton IT Innovation Centre. His main research interests are social media, sensor systems, data fusion and semantics. Stuart has a PhD in Computer Science from the University of Southampton.

@stuart_e_middle @IT_Innov

http://www.it-innovation.soton.ac.uk http://users.ecs.soton.ac.uk/sem/

REVEAL project, @RevealEU

http://revealproject.eu/

https://pypi.python.org/pypi?:action=display&name=geoparsepy

References

[1] Middleton, S.E. Middleton, L. Modafferi, S. 2014. *Real-Time Crisis Mapping of Natural Disasters Using Social Media. Intelligent Systems*, IEEE, vol.29, no.2, 9-17, DOI:10.1109/MIS.2013.126

[2] Middleton, S.E. Krivcovs, V. 2016. *Geoparsing and Geosemantics for Social Media: Spatio-Temporal Grounding of Content Propagating Rumours to support Trust and Veracity Analysis during Breaking News*, ACM Transactions on Information Systems (TOIS), 34, 3, Article 16 (April 2016), 26 pages. DOI=10.1145/2842604