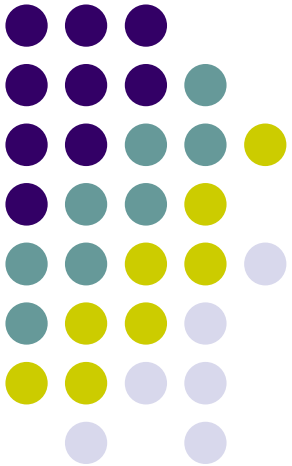
Big Data Analytics

Lab 3

Twitter geolocations



References

- ✓ Kruspe, Anna et al. (2021). Changes in Twitter geolocations: Insights and suggestions for future usage. *arXiv:2108.12251v1*

Geolocation data

Terminology

"*geolocated*": tweets containing explicit metadata about a geographic location they were posted from or are referring to

"*geotagging*": user action that causes this metadata to be attached

Since mid-2019 Twitter's policy radically changed with respect to geolocation availability

Motivations? Privacy

Geolocation data

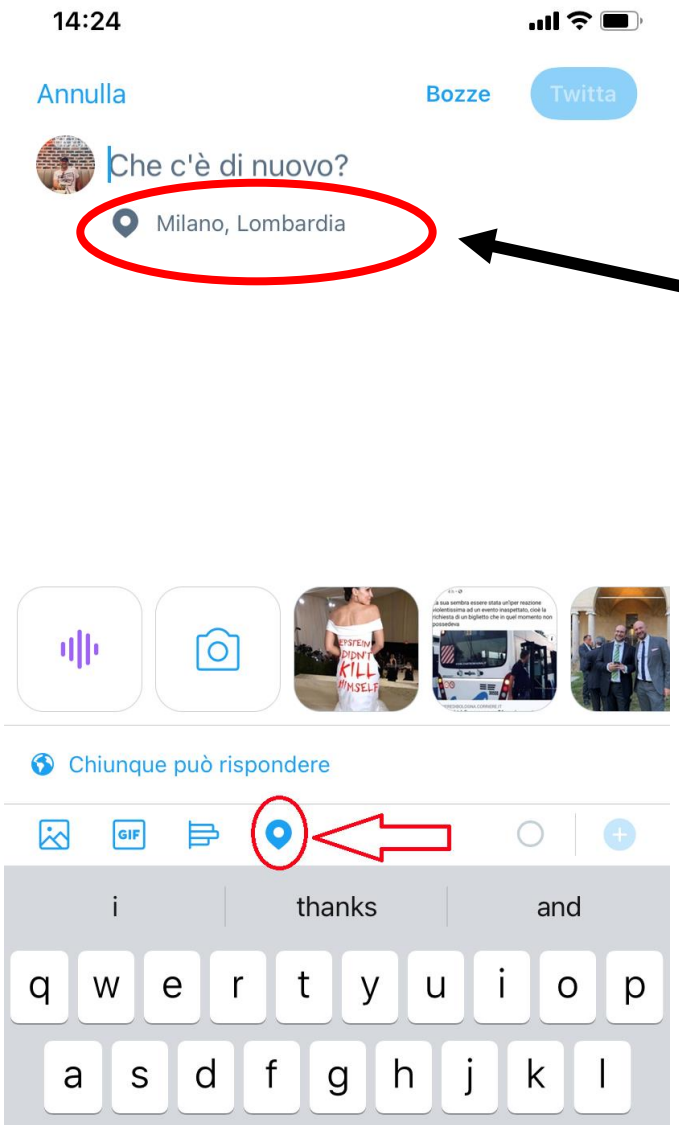
Which geo-data are then available (at least using `rtweet` 0.7.0 – using the 1.0.2. version, the geographical metadata are once again far less…)

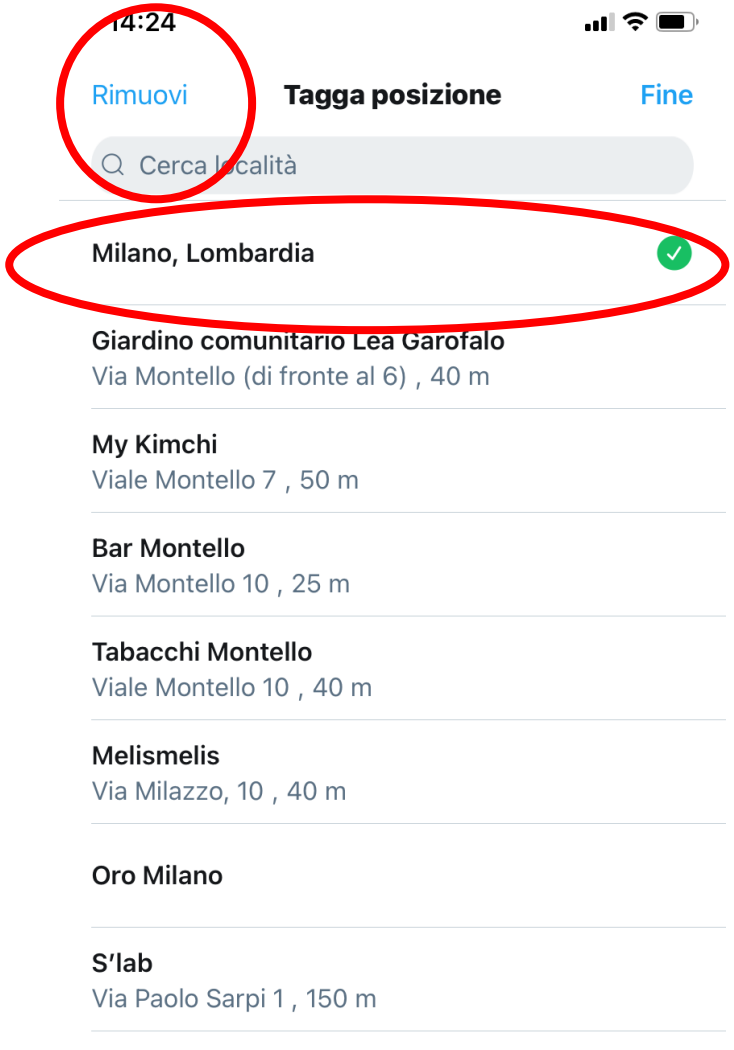
`place` attributes: the place attribute serves to assign a pre-defined geographic entity to a post

Twitter offers **users** the option to select this entity from a list of those found nearby (within a radius of roughly 200m) when sending a tweet

These entities may be countries, cities, neighborhoods, points of interest (POI), etc.

Geolocation data



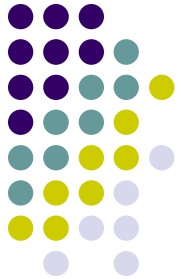


Geolocation data

place's sub-fields are then automatically filled using information from geolocation services

Among those subfields you have *bbox_coords* that contains a set of coordinates spanning a polygon

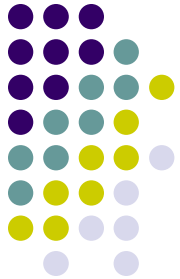
Geolocation data



coords_coords and geo_coords attributes:
originally (pre-2019) they were containing the longitude-latitude values of the tweet (provided the users allowed the geotagging option on her smartphone)

Nowadays refer basically to two possibilities: a) a user is employing a very old version of Twitter software on her smartphone; b) the tweet is a cross-post from third-party sources (typically a post on Instagram), and the coordinates reported on Twitter are those picked on Instagram

Geolocation data



However note one important point in the latter case: in this case the coordinates are not anymore representative of the user's geolocation from which the post was sent, but of some pre-defined location selected by the user, which may be very different from their physical location

In the case of native Twitter posts, these locations (via *bbox_coords*) will at least be somewhere close to the GPS location of the device (around 200m radius), whereas in Instagram, they may be anywhere in the world (as selected by the Instagram user)

Geolocation data

In general, the percentage of geolocated tweets out of all tweets is low at 1-2%

How to increase it? We can take advantage of the text either included in the tweets or in users' profiles (30/40% of profiles contain some form of geolocations) via for example a Named Entity Recognition approach (or by paying the Enterprise API…)