1.collect Twitter data about Covid from Japan, Poland and Italy,

2.use Twitter handles of all users who posted these tweets to collect further data on the last 3200 tweets (posted from these accounts regardless of their hashtags).

Q:Are you able to collect tweets about Covid from these three countries during the next two weeks?

Is it somehow problematic to identify the geo-localization of Twitter accounts?

if yes, do you think it is feasible to collect all Tweets and then select those written in Italian, Japanese or Polish?

Data Collection

1. collect Italian, Japanese, Polish tweets using related keywords

2. identify accounts by geo-location

3. collect & tract the identified user’s last 3200 tweets

4. select Italian, Japanese, Polish tweets

Italian, Japanese, Polish geo-located, language identified

"it" "ja" "pl"

/Users/jiaqizheng/Desktop/my\_research/TwitterResearch/DB\_Backup/postgre\_sql/em

pty\_twitter\_db.sql.gz

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Latitude | Longitude | Southwest |  | Northeast |  |
|  | 41.871941 |  |  |  |  |  |
| Poland (PL) |  |  |  |  |  |  |
| Japan (JP) |  |  |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Italy (IT) | Poland (PL) | Japan (JP) |
| Latitude | 41.871941 | 51.919437 | 36.204823 |
| Longitude | 12.567380 | 19.145136 | 138.252930 |
| Southwest | Naples (NA) | Wroclaw Wrocław | Okinawa |
|  | 40.833333, 14.25 | 51.107883, 17.038538 | 26.5, 128 |
| Northeast | Milan | Łomża | Hokkaido |
|  | 45.464664, 9.188540 | 53.176389, 22.073056 | 43.2203、142.8635 |
|  |  |  |  |

installed geopandas

# LOCATIONS are the longitude, latitude coordinate corners for a box that restricts the geographic area from which you will stream tweets. The first two define the southwest corner of the box and the second two define the northeast corner of the box.

Geo-Location

List of extreme points of J, I, P

日本の東西南北端の経度緯度

区分 場所 世界測地系

経度 緯度

最北端 北海道　択捉島 148°45′08″ 45°33′26″

[148.4508, 45.3326]

最南端 東京都　沖ノ鳥島 136°04′11″ 20°25′31″

[136.0411, 20.2531]

最西端 沖縄県　与那国島 122°55′57″ 24°27′05″

[122.5557, 24.2705]

最東端 東京都　南鳥島 153°59′12″ 24°16′59″

[153.5912, 24.1659]

<Italy>

Northernmost point: Westliches Zwillingsköpfl, Predoi, Alto Adige at 47°5′N 12°11′E

[47.083333, 12.183333]

Southernmost point on the mainland: Capo Spartivento, Calabria at 37°56′N 16°3′E; on

Lampedusa, Sicily: Punta Pesce Spada, at 35°29′N 12°36′E

[35.483333, 12.6]

Westernmost point: Rocca Bernauda, Bardonecchia, Piedmont at 45°6′N 6°37′E

[45.1, 6.616667]

Easternmost point: Capo d'Otranto, Otranto, Apulia at 40°6′N 18°31′E

[40.1, 18.516667]

<Poland>

• Northernmost point: 54°50′N 18°04′E

A beach in Jastrzębia Góra, near the town Władysławowo, Pomeranian Voivodeship, marked by the "Northern Star" obelisk, near Cape Rozewie on the Baltic coast

[54.833333, 18.066667]

• Southernmost point: 49.00238°N 22.84710°E

Wołosate ridge, near mount Opołonek, Eastern Beskids mountains, Subcarpathian Voivodeship

[49.00238, 22.8471]

• Easternmost point: 50.86852°N 24.14585°E

Bug River, near Zosin, Hrubieszów County, Lublin Voivodeship

[50.86852, 24.14585]

• Westernmost point: 52.83827°N 14.12298°E

Oder River, near Osinów Dolny, West Pomeranian Voivodeship

[52.83827, 14.12298]

Covid-19 related keywords

'corona', 'coronavirus', 'covid', 'covid19','sarscov2',

'covid-19', 'corona virus', 'quarantena', 'autoisolamento', 'auto-isolamento',

'iorestoacasa', 'stateacasa', 'COVID19Italia', 'redditodicittadinaza',

'eurobond', 'coronabond', 'restiamoacasa', 'preghiamoinsieme',

'NoMes', '#milanononsiferma', 'bergamononsiferma', 'l’italianonsiferma',

'abbraccciauncinese', 'iononsonounvirus', 'iononmifermo', 'aperisera',

'covidunstria', 'italiazonarossa', 'bergamoisrunning', 'chiudetetutto',

'apritetutto', 'CuraItalia', 'ciricordiamotutto', 'oggisciopero',

'chiudiamolefabbriche', 'iononrinuncioalletradizioni', 'andràtuttobene',

'INPSdown', 'percheQuando', 'cercareDi', 'ringraziarevoglio',

'600euro', 'CineINPS', 'COVID19Pandemic'

Reference

IEEE Dataport

<https://ieee-dataport.org/open-access/corona-virus-covid-19-tweets-dataset>

“corona”, "coronavirus", "covid", "covid19","sarscov2"

Twitter Developer

<https://developer.twitter.com/en/docs/labs/covid19-stream/filtering-rules>

<https://developer.twitter.com/en/docs/labs/covid19-stream/overview>

<https://github.com/echen102/COVID-19-TweetIDs>

Excellent Project

病例集可视化

<https://github.com/CSSEGISandData/COVID-19>

Large Arabic Twitter Dataset on COVID-19

<https://www.researchgate.net/publication/340541349_Large_Arabic_Twitter_Dataset_on_COVID-19>

<https://towardsdatascience.com/from-streaming-data-to-covid-19-twitter-analysis-using-aws-lambda-kinesis-firehose-and-b71b71279335>

(3)

‘#2019nCoV’, ‘#codvid\_19’, ‘#codvid19’, ‘#conronaviruspandemic’, ‘#coronaflu’, ‘#coronaoutbreak’, ‘#coronapandemic’, ‘#Coronapanik’, ‘#coronavid19’, ‘#コロナ’

‘#コロナウィルス’, ‘#新型コロナウイルス’, ‘#新型肺炎’ (4)

<Methodology >

<https://www.medrxiv.org/content/10.1101/2020.05.04.20090993v1.full.pdf>

We primarily analyed quantuative digital footprint data on the Internent from…to…

including their representativeness.

As a main keyword in investigated platforms

we chose a colloquial term “Koronawirus” / “Coronavirus” due to its high penetration in the society.

Other related keywords in use are much less popular, expect for Wikipedia, where the medical term SARS-CoV-2 was frequently chosen.

For

Keywords

According to

**1. Collect all COVID-related tweets by keywords (use Twitter search API)**

**→Have done the test, if the keywords are no problem, could start at any time.**

According to the related references or dataset, I choose these as the primary keywords in investigated platforms.

|  |  |
| --- | --- |
| Keywords | RFs |
| '#corona', '#coronavirus', '#covid', '#covid19', '#sarscov2', '#covid-19' | (1) |
| '#corona virus', ‘#Koronawirus’, ‘#Coronavirus’ | (2) |
| '#COVID19Italia', '#COVID19Pandemic' | (3) |
| ‘#2019nCoV’, ‘#codvid\_19’, ‘#codvid19’, ‘#conronaviruspandemic’, ‘#coronaflu’, ‘#coronaoutbreak’, ‘#coronapandemic’, ‘#Coronapanik’, ‘#coronavid19’, ‘#コロナ’, ‘#コロナウィルス’, ‘#新型コロナウイルス’, ‘#新型肺炎’ | (4) |

References

(1) IEEE Dataport <https://ieee-dataport.org/open-access/corona-virus-covid-19-tweets-dataset>

(2) Andrzej Jarynowski, Wojta-Kempa, Belik (2020) Trends in Perception of COVID-19 in Polish Internet. doi: <https://doi.org/10.1101/2020.05.04.20090993>

(3) http://twita.di.unito.it/dataset/40wita

(4) Twitter COVID-19 Stream. <https://developer.twitter.com/en/docs/labs/covid19-stream/filtering-rules>

**2. Collect & Track tweets by geographic location (using Twitter streaming API)**

**→I'm writing this code, still need some modifies for now.**

By one purely anecdotal estimate, somewhere between 1% of Twitter users have set up precise location tracking and 10% have set up tracking to a wider area, such as a city.

The Tweets could be tracked by location, language, and text by passing the three arrays shown below into the stream.filter() method of Tweepy, which is a Python API for Twitter. The location is a rectangle whose first two coordinates (longitude and latitude) are the bottom left corner and the last two are the top right corner.

Therefore, I collected the extreme points of Italy, Poland and Japan.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Italy | Poland | Japan |
| Northernmost | [47.083333, 12.183333] | [54.833333, 18.066667] | [148.4508, 45.3326] |
| Southernmost | [35.483333, 12.6] | [49.00238,  22.8471] | [136.0411, 20.2531] |
| Easternmost | [45.1, 6.616667] | [50.86852, 24.14585] | [122.5557, 24.2705] |
| Westernmost | [40.1, 18.516667] | [52.83827, 14.12298] | [153.5912, 24.1659] |

In this way, we could collect tweets that meet the language, location and keywords conditions at the same time.

I’ll try both of them to select the suitable data.