2020/5/30

I collected data using both Twitter search API and streaming API.

Dataset1 was collected by these keywords, and users could be identified by tweet languages. The command of search API was traversing all tweets containing keywords in the past week or so, without defining locations and languages.

1. Keywords

**['corona', '#coronavirus', '#COVID19Italia', '#Koronawirus', '#COVID19Pandemic', '新型コロナウィルス', 'コロナ','新型肺炎'].**

 I selected a part of keywords from all of the keywords mentioned before because I realized that if we want to collect tweets created as early as possible, too many keywords will affect to review the past tweets. The algorithm was like after traveling all of the tweets in the past week containing keyword 1, then finally starts to traverse all tweets of the past week containing keyword2…. So I choosed the main keywords.

 As of today, Dataset1 contains 124 kind of languages, a total of 5.3 million tweets, and 2.7 million users information. 2.3% of users' tweet was in Italian, Polish, and Japanese (The number of users: it\_user: 8,914, pl\_user: 4,881, ja\_user: 48,951).

2. Geolocation

 We knew that only a few tweets were geotagged. And yes, I think searching place names from the user's profile and text might improve this problem. However, I don't have enough confidence in defining suitable location names since I don't understand Italian and Polish.

In addition, the proportion of users in the three languages ​​is too small, which I didn't expect before.Therefore, I tried the Twitter streaming API**.**

Different from the search API, **streaming API** collects data in real-time. As we already understand the publicity of the users and geographic locations of three languages ​​from Dataset 1, I selected all of the keywords, three languages ​​, and geolocations of countries' longitude and latitude in this step. I haven't collected data in this way, it took some time but finally working.

We could get tweets and user information from a country area, regardless of whether users are publishing their location or not. In other words, we could collect tweets meeting the languages, location, and keywords conditions at the same time.

 e.p. Using the three languages, all of the keywords, and the latitude and longitude of Italy to collect data could ensure that the tweet was sent from Italy. And not only Italian tweets, but also from users who use Polish and Japanese in Italy. As Krzys said, they may be an interesting comparison group.

 Using the streaming API, I collected more than 10,000 tweets and user information from Italy in 3 hours. The same applies to Poland and Japan. (Concluding the data of tweet text, user\_id, language,user\_followers\_count,user\_location, created\_at..)

I think we could have more choices about the dataset now.  **In order to get suitable user ids, Twitter streaming API seems to work well.**And if we need, we could also add some conditions in filtering, such as tweet's retweet count, or user's follower count, and so on.

Finally, we could use the user id to collect their 3200 tweets, as well as those tweets and user information.