COVID-19 Twitter Data Analysis

March 22, 2020

1 Data

We consider tweets provided in [1]. The data and instructions for downloading the tweet IDs are available at https://github.com/echen102/COVID-19-TweetIDs. You will need a Twitter developer account and Twitter API tokens to 'hydrate' the tweets. Downloading the data takes a few days to accommodate Twitter's request rate limits.

There is a lot of information for each tweet and we are likely only interested in a subset of it. Each tweet has the following fields available:

```
['created_at', 'id', 'id_str', 'full_text', 'truncated', 'display_text_range', 'entities',
 'extended_entities', 'source', 'in_reply_to_status_id', 'in_reply_to_status_id_str',
 'in_reply_to_user_id', 'in_reply_to_user_id_str', 'in_reply_to_screen_name', 'user',
 'geo', 'coordinates', 'place', 'contributors', 'retweeted_status', 'is_quote_status',
 'retweet_count', 'favorite_count', 'favorited', 'retweeted', 'possibly_sensitive', 'lang']
```

Many of these fields contain multiple pieces of information as well.

```
'entities': ['hashtags', 'symbols', 'user_mentions', 'urls', 'media']
within entities, 'media': gives a list of linked media.
Each entry contains:
['id', 'id_str', 'indices', 'media_url', 'media_url_https', 'url', 'display_url',
 'expanded_url', 'type', 'sizes', 'source_status_id', 'source_status_id_str',
 'source_user_id', 'source_user_id_str']
```

The 'extended_entities' field contains more detailed media info including an 'additional_media_info' field with

```
['title', 'description', 'call_to_actions', 'monetizable', 'source_user']
```

The 'source_user' field and the 'user' field of the tweet additionally contain

```
['id', 'id_str', 'name', 'screen_name', 'location', 'description', 'url', 'entities',
 'protected', 'followers_count', 'friends_count', 'listed_count', 'created_at',
 'favourites_count', 'utc_offset', 'time_zone', 'geo_enabled', 'verified',
 'statuses_count', 'lang', 'contributors_enabled', 'is_translator',
 'is_translation_enabled', 'profile_background_color',
 'profile_background_image_url', 'profile_background_image_url_https',
 'profile_background_tile', 'profile_image_url', 'profile_image_url_https',
 'profile_banner_url', 'profile_image_extensions_alt_text',
 'profile_banner_extensions_alt_text', 'profile_link_color',
 'profile_sidebar_border_color', 'profile_sidebar_fill_color', 'profile_text_color',
 'profile_use_background_image', 'has_extended_profile', 'default_profile',
 'default_profile_image', 'following', 'follow_request_sent', 'notifications',
 'translator_type']
```

1.1 Plan to filter data:

The dataset is very large (60 million tweets) and growing. We will need to filter the data and also process as a stream or in parts.

- Limit to tweets that are in English (can use the 'lang' field to do this easily).
- Remove unnecessary information fields - what information are we most interested in. What would we like to keep.

References

[1] Emily Chen, Kristina Lerman, and Emilio Ferrara. COVID-19: The First Public Coronavirus Twitter Dataset. *arXiv e-prints*, page arXiv:2003.07372, March 2020.