Project

Graph database description

20th January 2023

In this document, a description of the nodes and relationships within the graph, including their distinct properties, will be presented.

Node types

1. User

The Twitter users that are found at least once in the raw data.

1. **favourites_count :** The number of Tweets this user has liked in the account's lifetime.
2. **followers_count :** The number of followers this account currently has.
3. **friends_count :** The number of users this account is following (AKA their "followings").
4. **id :** The integer representation of the unique identifier for this User.
5. **isVerified :** When true, indicates that the user has a verified account.
6. **listed_count :** The number of public lists that this user is a member of.
7. **name :** The name of the user, as they've defined it. Not necessarily a person's name.
8. **screen_name :** The screen name, handle, or alias that this user identifies themselves with. screen_names are unique but subject to change. Use id as a user identifier whenever possible.
9. **statuses_count :** The number of Tweets (including retweets) issued by the user.
10. **tweets_count :** Deprecated, should be deleted.

2. Tweets

All the tweets found in the raw data. Some properties were stripped because there was either no data, or were not useful to the project.

1. **annotation_annotated :** Boolean that indicates if a tweet was annotated or not.
2. **annotation_num_judgements :** Number of annotations given for a tweet
3. **annotation_postPriority :** Relative priority of the tweet with respect to the event it is about.
4. **created_at :** UTC time when this Tweet was created.
5. **favourite_count :** Indicates approximately how many times this Tweet has been liked by Twitter users.

6. **id :** The integer representation of the unique identifier for this Tweet.
7. **id_str :** The string representation of the unique identifier for this Tweet. Implementations should use this rather than the large integer in id.
8. **isTruncated :** Indicates whether the value of the text parameter was truncated, for example, as a result of a retweet exceeding the original Tweet text length limit of 140 characters. Truncated text will end in ellipsis, like this "..." Since Twitter now rejects long Tweets vs truncating them, the large majority of Tweets will have this set to false.
9. **is_quote_status :** Indicates whether this is a Quoted Tweet.
10. **possibly_sensitive :** This field indicates content may be recognized as sensitive. The Tweet author can select within their own account preferences and choose "Mark media you tweet as having material that may be sensitive" so each Tweet created after has this flag set.
    This may also be judged and labeled by an internal Twitter support agent.
11. **retweet_count :** Number of times this Tweet has been retweeted.
12. **text :** The actual UTF-8 text of the status.
13. **topic :** Topic of the tweet symbolized by the file it was found in. Is equivalent to the event it is about.

3. Hashtag

All the hashtags present in the tweets of the dataset. Will be used for the information retrieval system.

1. **id :** The hashtag in text form (excluding the **#** symbol).
2. **occurrences :** The number of times this hashtag was used in the dataset.

4. Event

The events of the TREC-IS dataset (can be considered as an annotation).

1. **eventType :** The type of event (*wildfire, flood, earthquake, ...*)
2. **id :** The event id (for example *fireColorado2012, floodColorado2013, costaRicaEarthquake2012, ...*)
3. **trecisid :** the file that corresponds to this id (the file contains all the tweets of this event)

5. PostCategory

The categories of the tweets. This information is obtained thanks to the annotations given in the TREC-IS dataset and is not available for all the tweets.

1. **id :** The category (for example : *Official, News, Discussion, FirstPartyObservation, Advice, ...*)

Relationship types

1. `REPLY_TO`
   ↳ *`Tweet`* → *`Tweet`*

Links a reply to the source tweet (*reply* → *original tweet*)

2. `RETWEETED`
   ↳ *`Tweet`* → *`Tweet`*

Links a tweet to the retweeted tweet (*tweet* → *retweeted tweet*)

3. `HAS_HASHTAG`
   ↳ *`Tweet`* → *`Hashtag`*

Links a tweet to the hashtags used in it.

4. `HAS_CATEGORY`
   ↳ *`Tweet`* → *`PostCategory`*

Links a tweet to its categories (annotation).

5. `IS_ABOUT`
   ↳ *`Tweet`* → *`Event`*

Links a tweet to the event it is about.

6. `REPLY_TO`
   ↳ `Tweet` → `User`

Links a tweet to the user it is replying to.

7. `MENTIONS`
   ↳ *`Tweet`* → *`User`*

Links a tweet to the user(s) it mentions.

8. `POSTED`
   ↳ *`User`* → *`Tweet`*

Links a user to all the tweets he posted (tweets of the dataset)

9. RETWEETS

↳ *User* → *User*

Links a user to all the user they retweeted at least once

1. **times :** Number of times this user retweeted the other user.

10. REPLIED_TO

↳ *User* → *User*

Links a user to all the users they replied to.

1. **times :** Number of times this user replied to the other user.

11. MENTIONS

↳ *User* → *User*

Links a user to all the users they mentioned at least once.

1. **times :** Number of times this user retweeted the other user.

12. TALKS_ABOUT

↳ *User* → *Event*

Links a user to all the events they talked about at least once.