computational social media

assignment #2

25.03.2022

**The assignment:
hands-on exercise with Twitter data**

Goals

1. Students are asked to collect a small set of public tweets and perform a descriptive analysis, in the topic of climate change.

2. Through the exercise, students will (a) gain practical skills on how to collect Twitter data; (b) understand collective Twitter practices (using hashtags, sharing URLs, retweeting); and (c) reflect about the use of Twitter data in the context of world events.

Instructions

1. You can use the scripts that were made available on moodle.

2. Using this and the tutorial on how to get Twitter data, obtain one sample between 10,000-15,000 tweets that contain any combination of hashtags in the list posted on moodle. Feel free to expand the list with other hashtags you think are missing. The sample is a general, non-geo-localized sample. Please start this process as early as possible and be mindful of Twitter rate limits.

3. Before using the dataset, look on Twitter for the handles of (a) 20 media organizations that you know are tweeting about the climate change topic (e.g., @nytimes) and 20 NGO or government public officials (e.g., @UNFCCC, @350) who are also tweeting about this topic.

4. Compute the following descriptive statistics on the dataset:
   - Percentage of tweets that contain URLs
   - Percentage of tweets that are (or contain) retweets.
   - Distribution of languages declared in the tweet metadata (%EN, %FR,....)
   - Table of the 30 most frequent hashtags in the following format:
     [rank, hashtag, frequency]. Example: [1, #climatechange, 2500]

Instructions

5. Using the Twitter handles you chose in step 3, compute the following statistics:
   - Percentage of tweets <u>directly</u> generated by all the 20 **media** accounts together.
     e.g.: 3% of tweets were produced by the 20 media accounts altogether
   - Percentage of tweets <u>directly</u> generated by the 20 **NGO/gov.** accounts.
     e.g.: 5% of tweets were produced by the 20 NGO/government accounts.
   - Percentage of tweets generated by all the 20 **media** accounts that appear as <u>retweets</u> in the sample.
   - Percentage of tweets generated by all the 20 **NGO/gov.** accounts that appear as <u>retweets</u> in the sample.

6. Discuss in a couple of lines the results of points 4 & 5. Are any of the results unexpected? Why?

7. Input a selection of your results and discussion via the questionnaire that is available on moodle. You should also submit your Jupyter Notebook via moodle so we can check your code. The aggregated results will be discussed in class.

Logistics and deadline

In case of questions, contact Lakmal by email asap.
lakmal.meegahapola@epfl.ch

Deadline to submit assignment: Thu 14.04.2022, 7pm
        - please submit your assignment even if it is not complete
        - late assignments will not be given any credit

questions?

daniel.gatica-perez@epfl.ch