Working with Telegram Data

Ishita Gopal

What is Telegram?

- A messaging platform with social media like features
  - developed by 2 Russian brothers who made VK
  - developed in response to repression they faced from the Russian Government

- Anecdotal evidence:
  - “Protest app” – enjoys popularity amongst the opposition
  - Enjoys popularity in authoritarian regimes
  - Governments have made various attempts to censor it

Project

- Activity in Belarusian protest groups which were active during the 2020-21 protest wave. Analyzing how it relates to on ground protests and repression events.

- There isn't much research on Telegram.
- Even fewer studies examine the type of political content on it.
- Research (Contentious Politics ∩ Social Media) suffers from
  - Platform Bias – most studies concentrate on Twitter

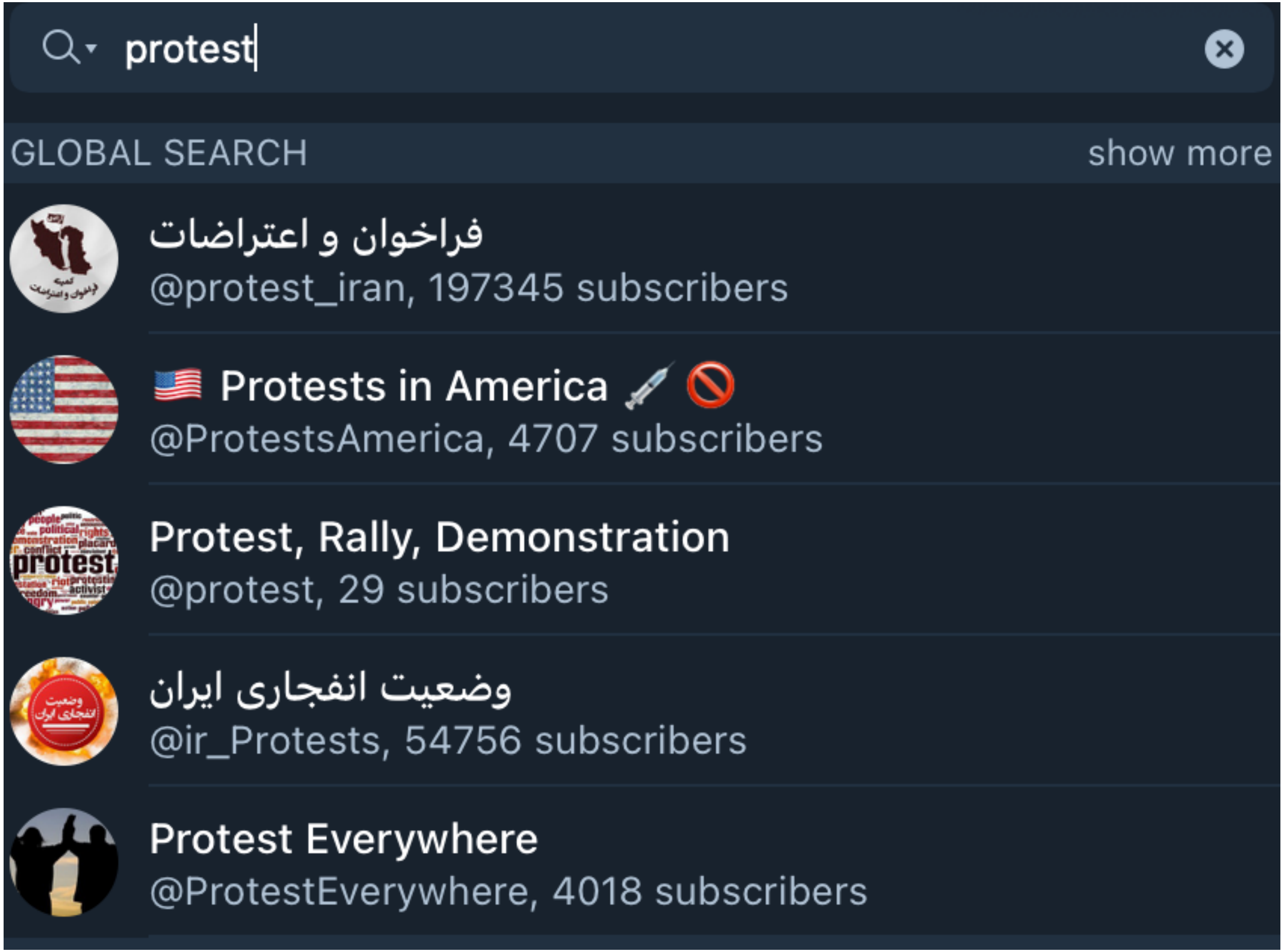
Affordance of the platform

Layout

1) 1 to 1 messaging
2) Many to many messaging (Group Chats)
    - Public or Private
    - members <= 200,000
3) 1 to many messaging (Channels)
    - Broadcasting pages
    - People who follow are called “subscribers”
    - Similar to Facebook Pages
4) Perceived security features

How to collect data?

- The keyword search is limited:
  - Returns about 5-10 channel/group suggestions.
  - Returns channels/group chats which have the search term in the “@username” or the “title” of the chat.
  - Searching for a topic will not always return relevant channels.
  - The exact ranking logic is not known.
  - Once you know the username, you go back in time.
  - Allows you to collect data from when the chat was created.



protest

GLOBAL SEARCH show more

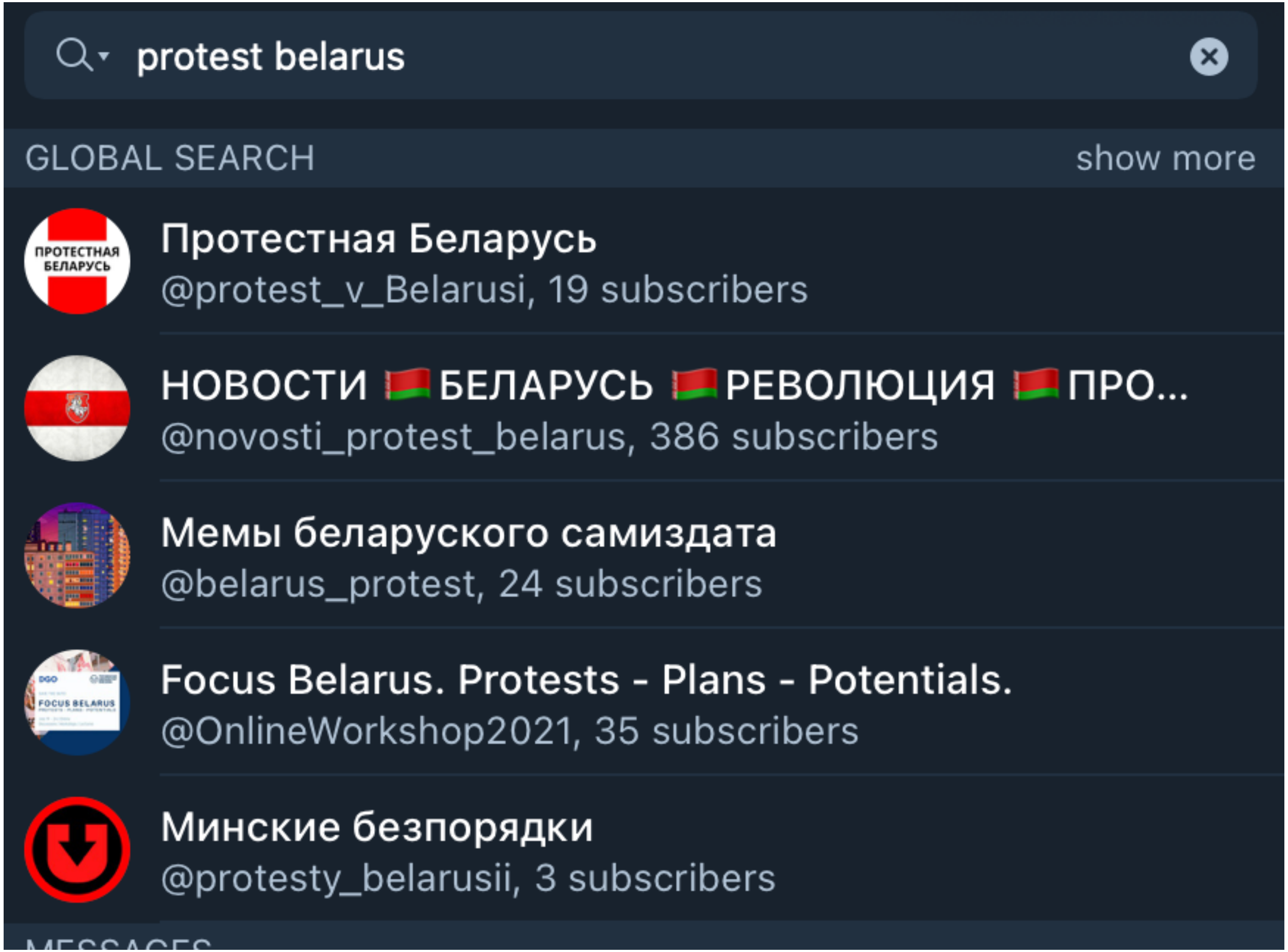
فراخوان و اعتراضات
@protest_iran, 197345 subscribers

🇺🇸 Protests in America 💉 🚫
@ProtestsAmerica, 4707 subscribers

Protest, Rally, Demonstration
@protest, 29 subscribers

وضعیت انفجاری ایران
@ir_Protests, 54756 subscribers

Protest Everywhere
@ProtestEverywhere, 4018 subscribers



protest belarus

GLOBAL SEARCH show more

Протестная Беларусь
@protest_v_Belarusi, 19 subscribers

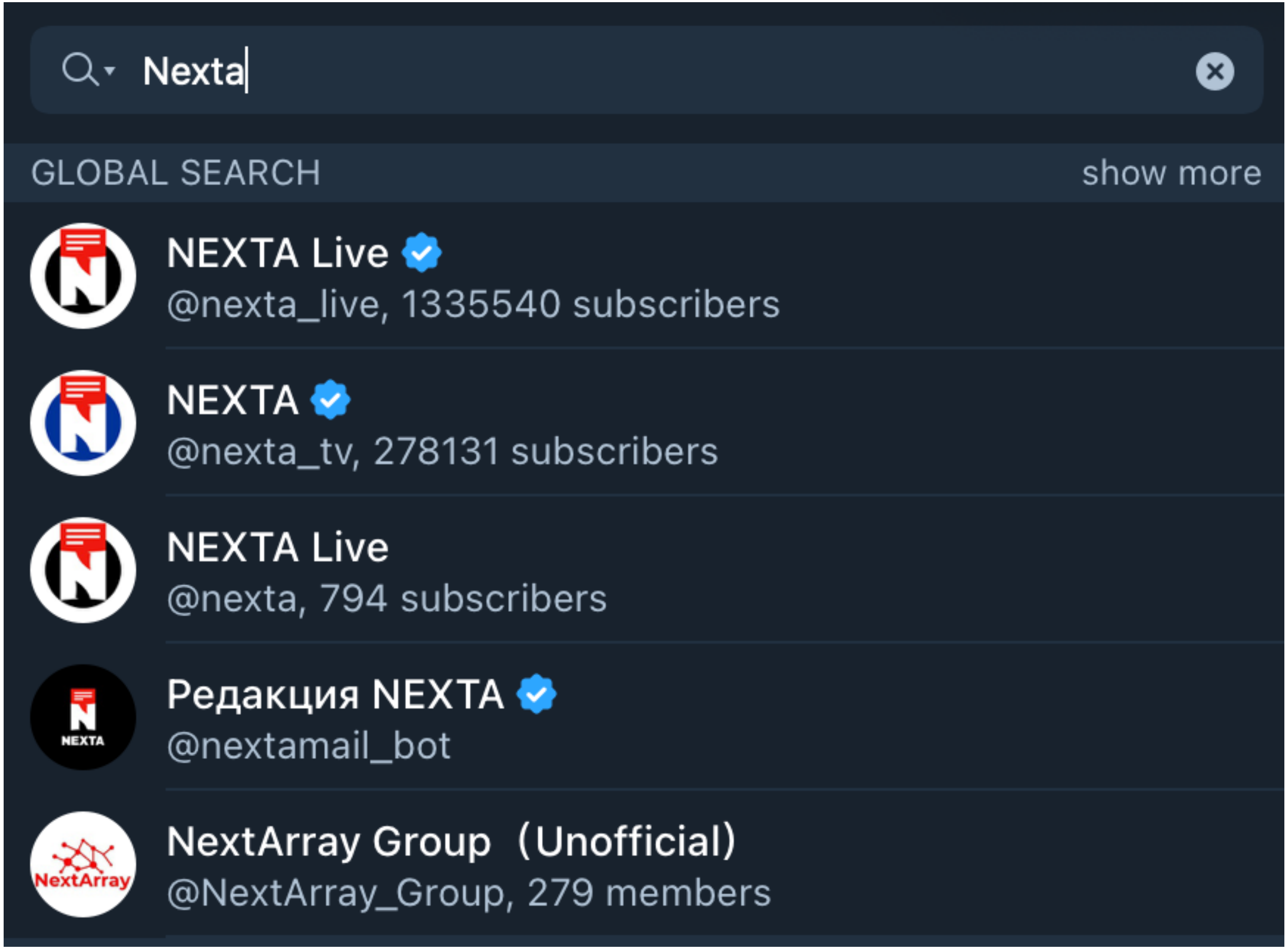
НОВОСТИ 🇧🇾 БЕЛАРУСЬ 🇧🇾 РЕВОЛЮЦИЯ 🇧🇾 ПРО...
@novosti_protest_belarus, 386 subscribers

Мемы беларуского самиздата
@belarus_protest, 24 subscribers

Focus Belarus. Protests - Plans - Potentials.
@OnlineWorkshop2021, 35 subscribers

Минские безпорядки
@protesty_belarusii, 3 subscribers

MESSAGES



Nexta

GLOBAL SEARCH show more

NEXTA Live ✓
@nexta_live, 1335540 subscribers

NEXTA ✓
@nexta_tv, 278131 subscribers

NEXTA Live
@nexta, 794 subscribers

Редакция NEXTA ✓
@nextamail_bot

NextArray Group (Unofficial)
@NextArray_Group, 279 members

Usernames

- Usernames are similar to Twitter handles and are unique
- Eg: @protest_iran
- Nexta Live (@nexta_live) was one of the most influential Telegram channels according to anecdotal evidence but does not show up unless explicitly searched.

- Identify relevant public channels or groups → use usernames of these channels/groups to collect data

Example: identifying chats

- Lukashenko won a fraud election on 9th August 2020

- This led to protests between August 2020 – March 2021

- Protests were organized using group chats and channels dedicated to the movement

- By 2021, 107 Telegram channels/groups were recognized as "extremist" and this list is growing

- I collected data from group chats which were listed as “extremist” and which still existed

- I was able to identify 35 such groups from different regions in Belarus

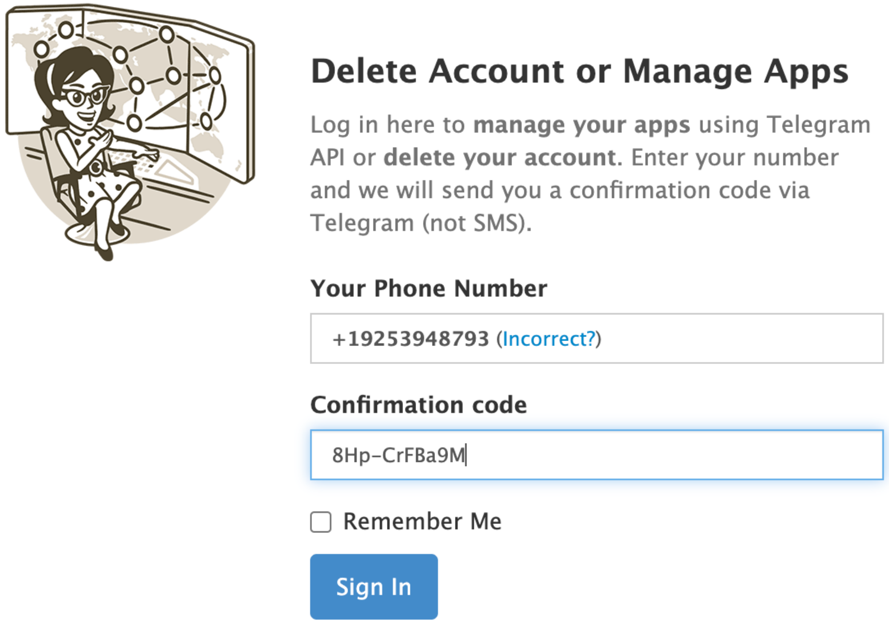
- Many of these groups were eventually abandoned but during the movement they had as many as 100K subscribers in a group.

Telegram's API

- 1. Download the Telegram app and sign up with your phone number.

- 2. Get API credentials: *api_id, api_hash* for authentication.

(https://my.telegram.org/ → https://my.telegram.org/apps)



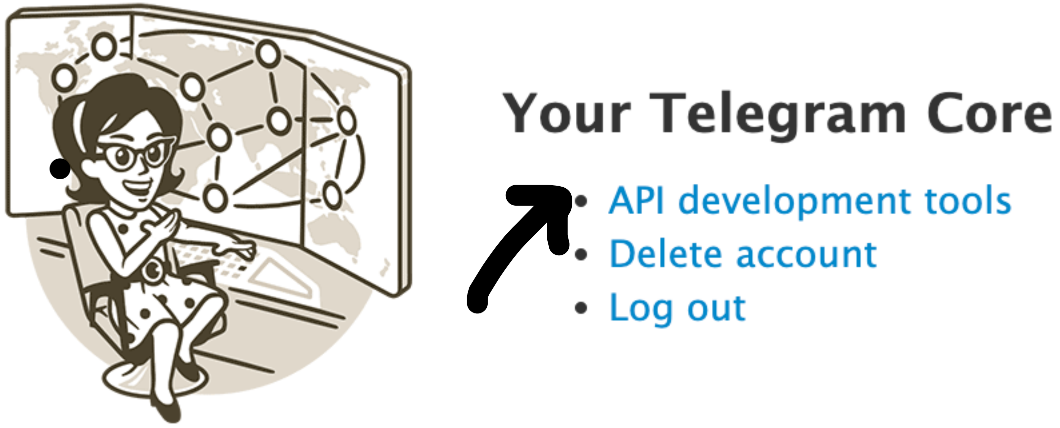
Delete Account or Manage Apps

Log in here to **manage your apps** using Telegram API or **delete your account**. Enter your number and we will send you a confirmation code via Telegram (not SMS).

Your Phone Number: +19253948793 (Incorrect?)

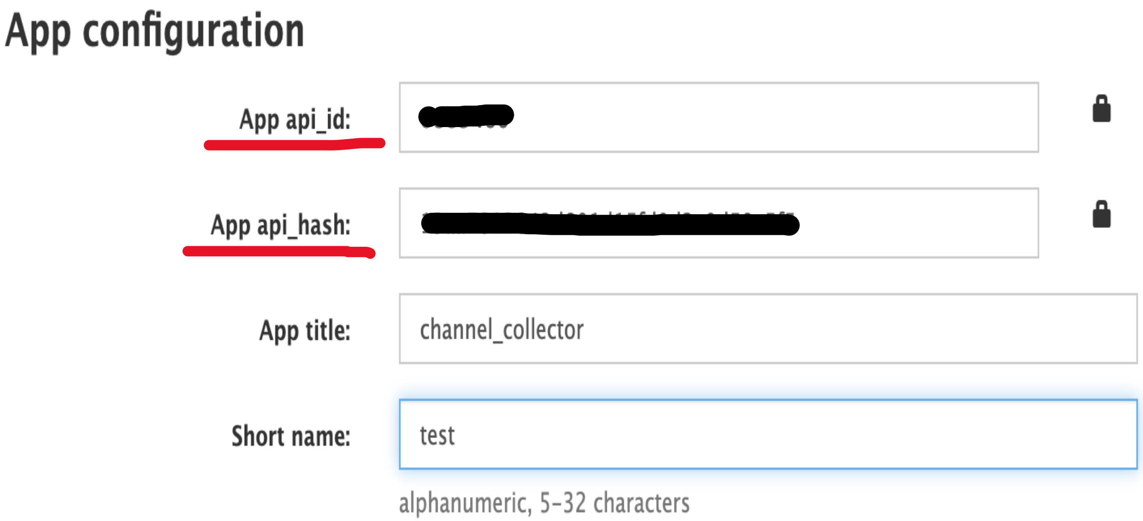
Confirmation code: 8Hp-CrFBa9M

□ Remember Me



Your Telegram Core

- API development tools
- Delete account
- Log out



App configuration

App api_id: [redacted]

App api_hash: [redacted]

App title: channel_collector

Short name: test

alphanumeric, 5-32 characters

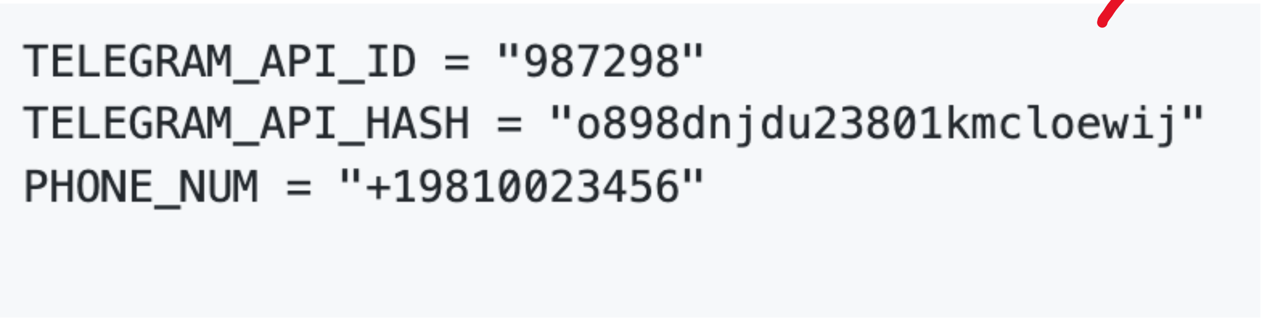
Store API credentials in a .env file

1.pip install python-dotenv

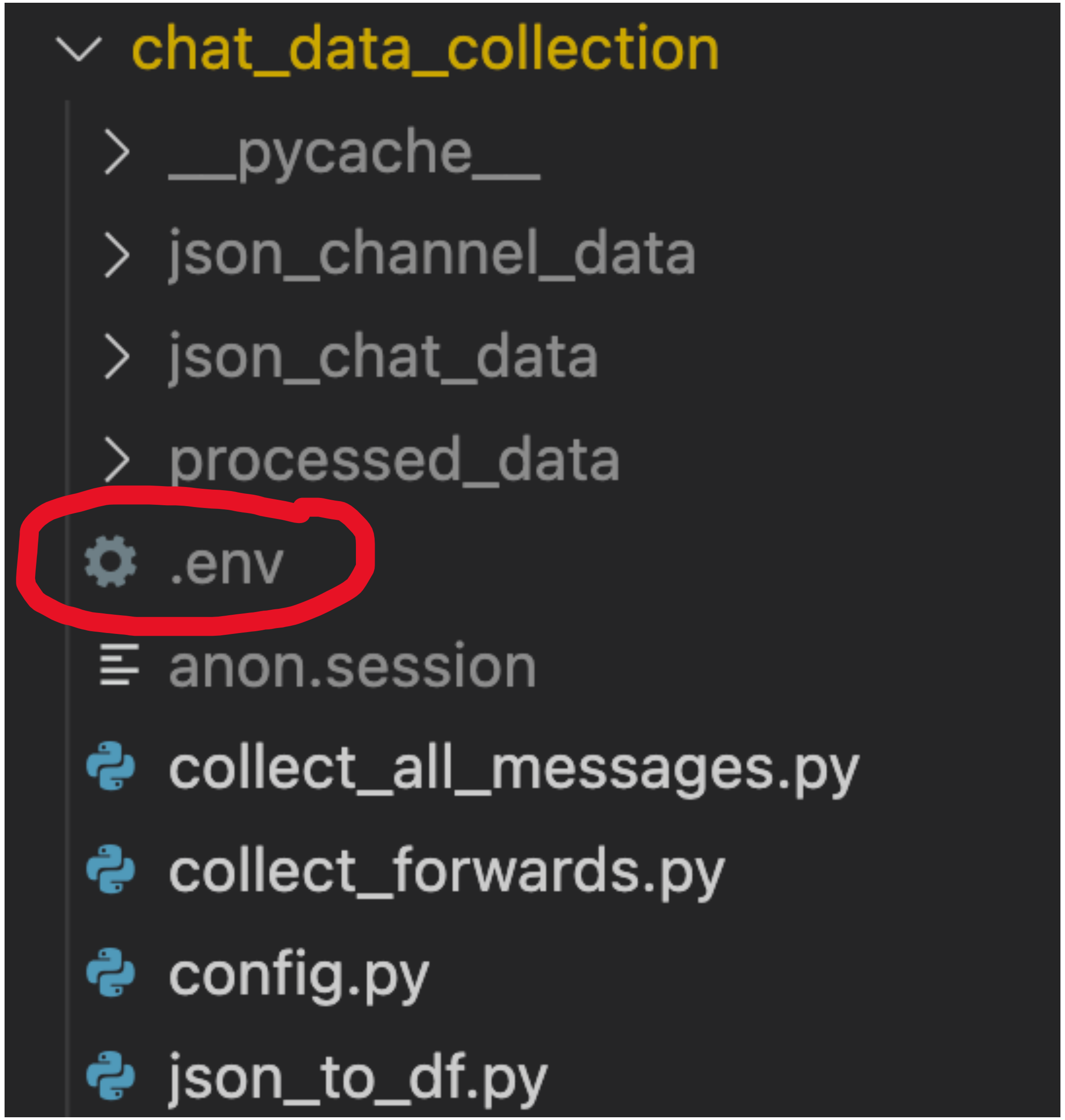
2.create a file with name .env in the working directory

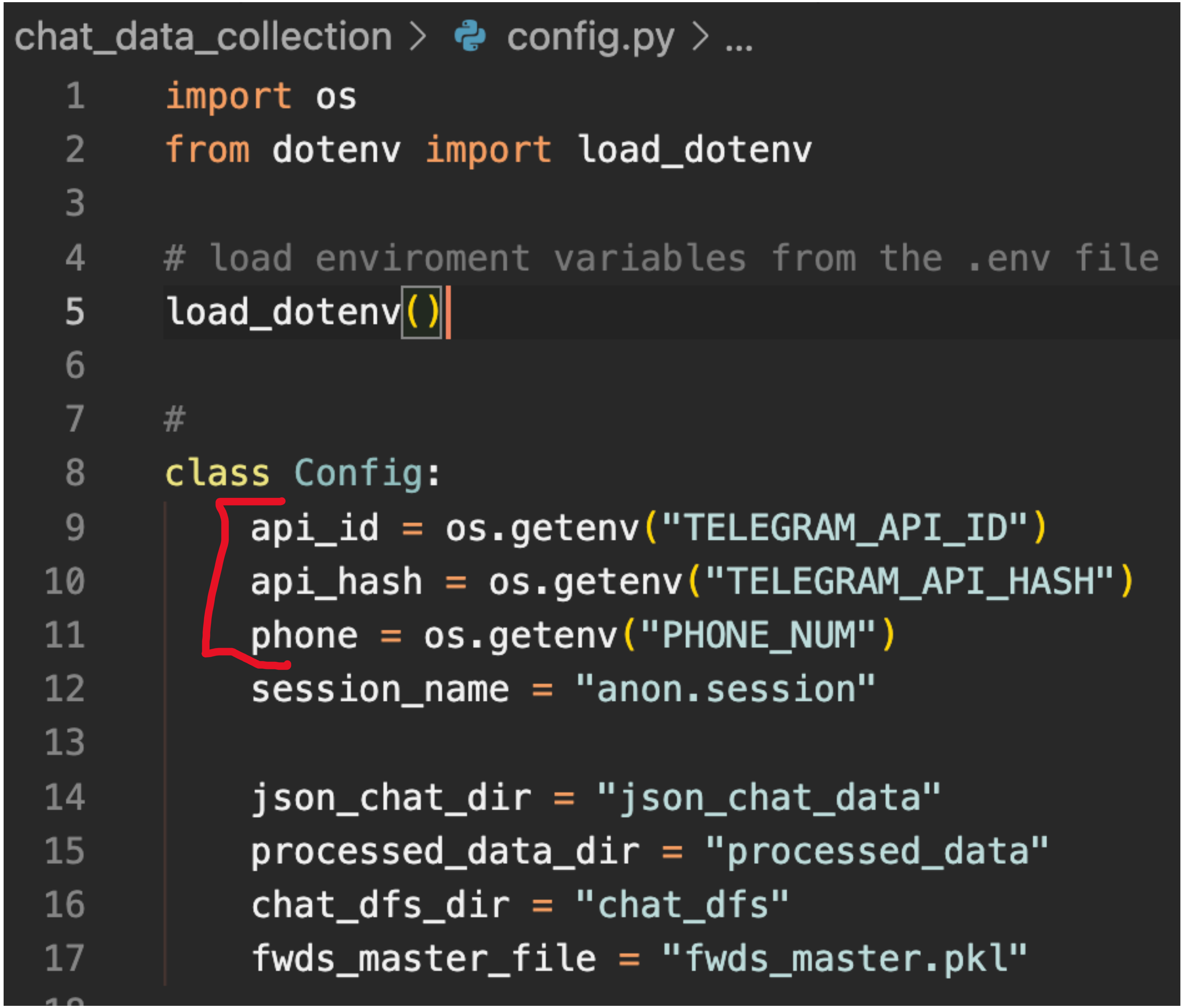
3.put your API key and hash in the following format

- Note: dotenv allows us to access private credentials from a secret file. These files don't show up in the file browsers.



```
TELEGRAM_API_ID = "987298"
TELEGRAM_API_HASH = "o898dnjdu23801kmcloewij"
PHONE_NUM = "+19810023456"
```



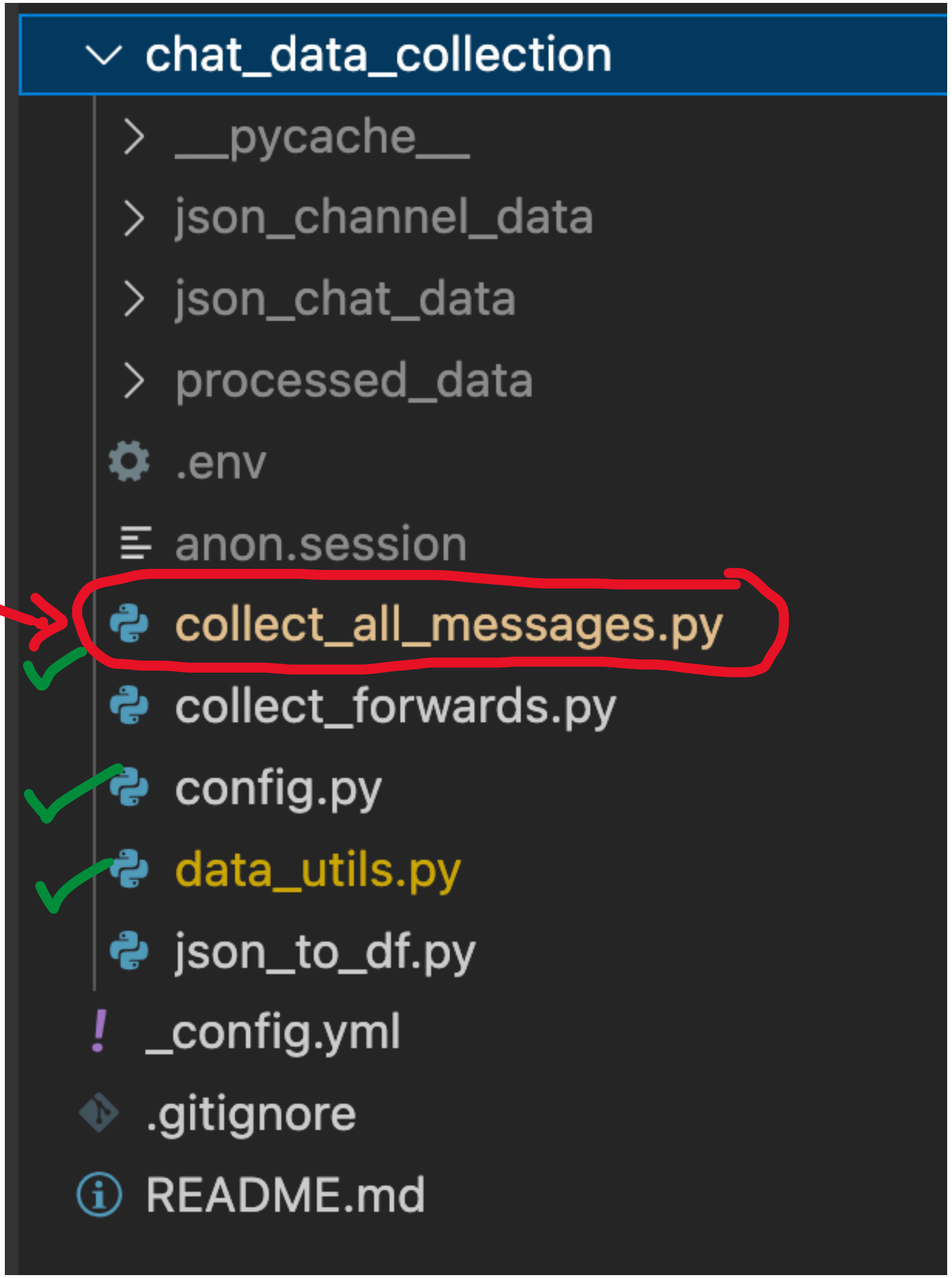


```python
import os
from dotenv import load_dotenv

# load enviroment variables from the .env file
load_dotenv()

#
class Config:
    api_id = os.getenv("TELEGRAM_API_ID")
    api_hash = os.getenv("TELEGRAM_API_HASH")
    phone = os.getenv("PHONE_NUM")
    session_name = "anon.session"

    json_chat_dir = "json_chat_data"
    processed_data_dir = "processed_data"
    chat_dfs_dir = "chat_dfs"
    fwds_master_file = "fwds_master.pkl"
```

- Execute collect_all_messages.py in the terminal.

- This will use directory locations and additional functions provided in config.py and data_utils.py

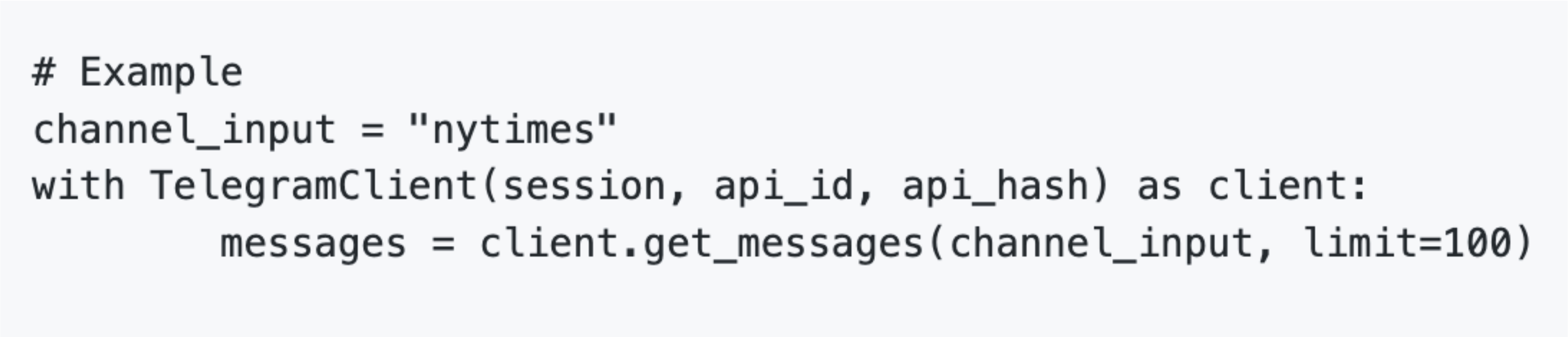
- Make sure you have all 3 of these scripts.



```
chat_data_collection
  __pycache__
  json_channel_data
  json_chat_data
  processed_data
  .env
  anon.session
  collect_all_messages.py ✓
  collect_forwards.py
  config.py ✓
  data_utils.py ✓
  json_to_df.py
_config.yml
.gitignore
README.md
```

- Use Telethon in Python : a wrapper for the Telegram's API.

- Use get_messages() method to collect messages which takes the channel/group username as an input.



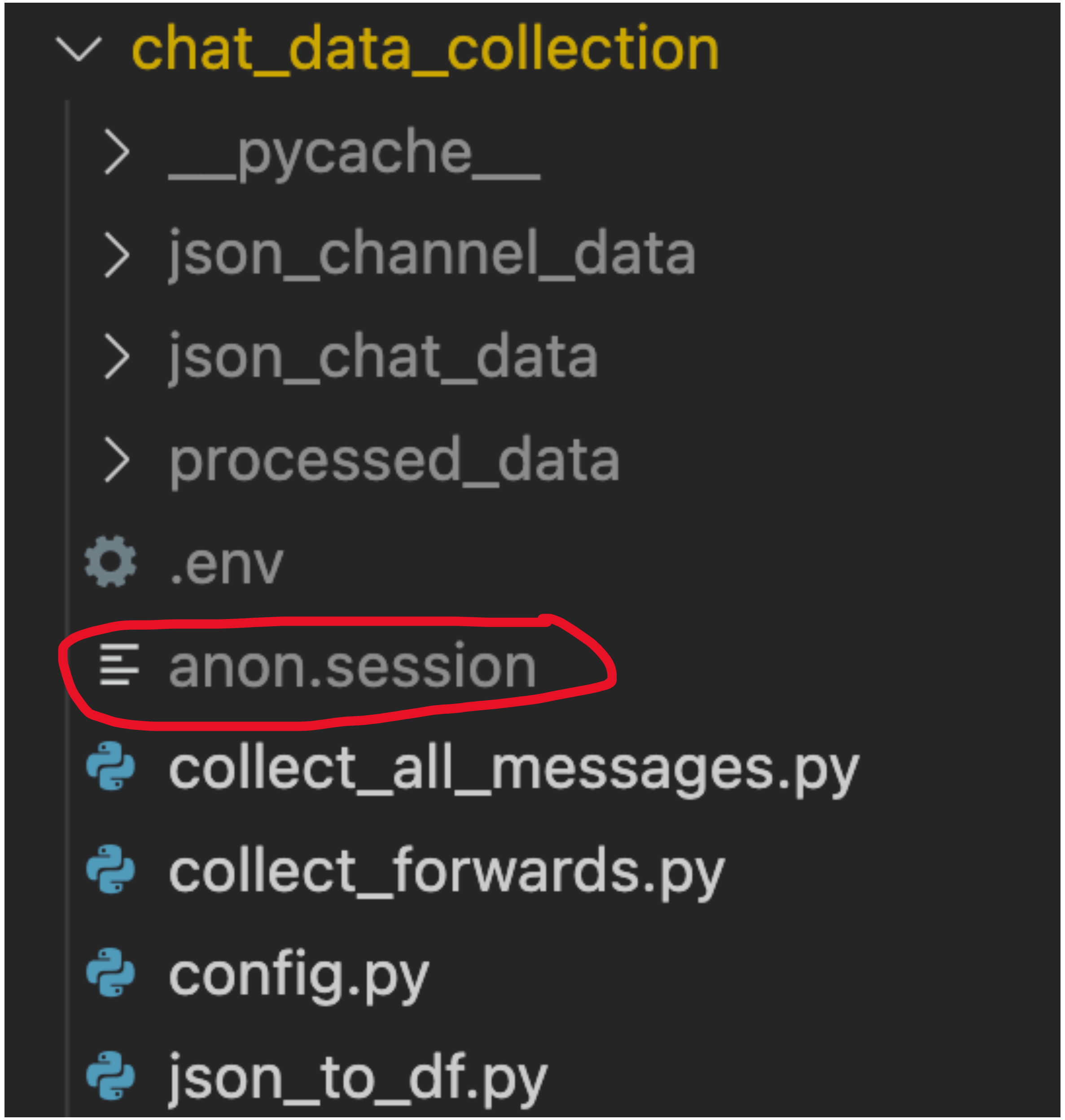
```
# Example
channel_input = "nytimes"
with TelegramClient(session, api_id, api_hash) as client:
        messages = client.get_messages(channel_input, limit=100)
```

What is a session file?

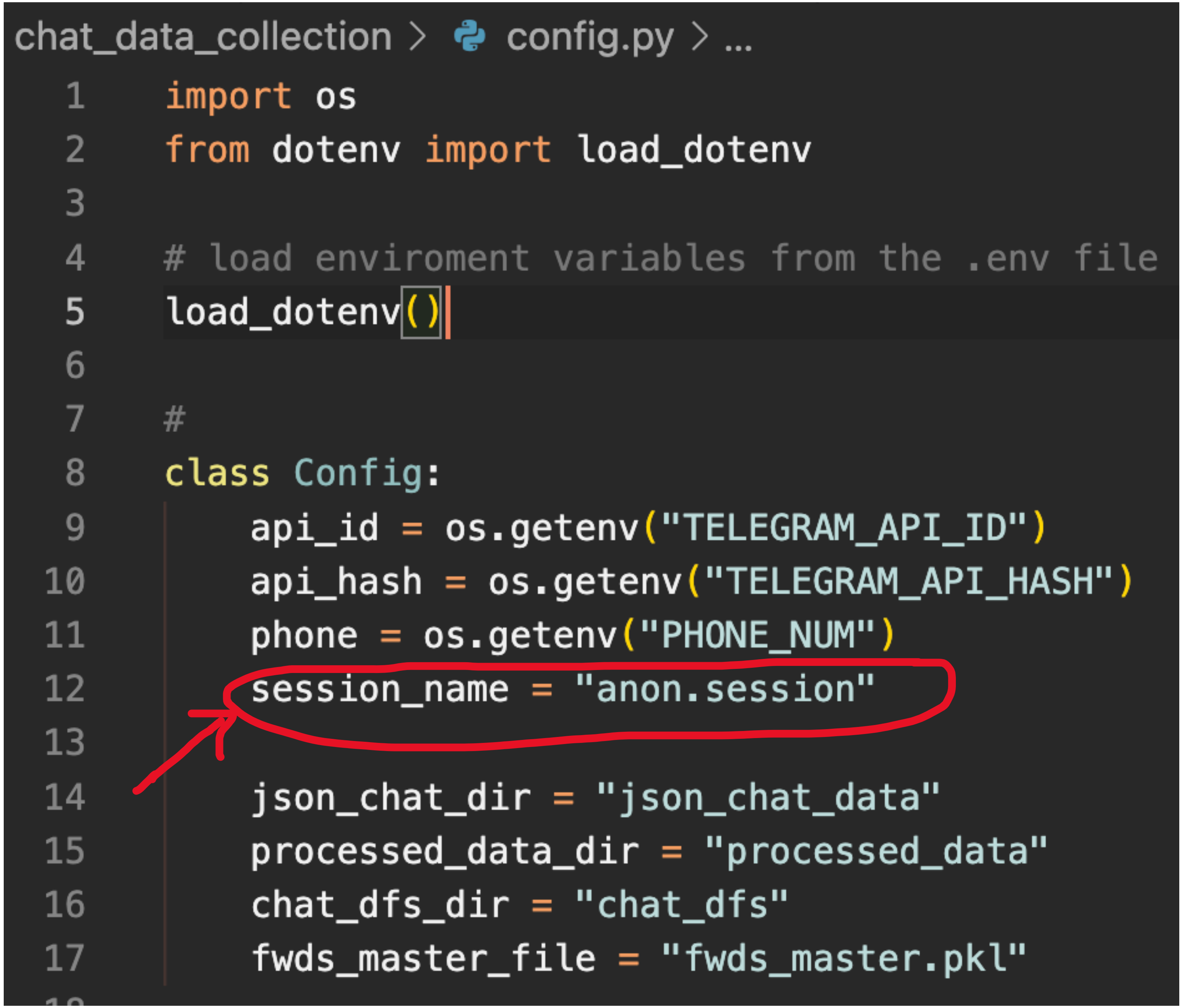
1. These files contain enough information for you to login without re-sending the code each time you make a request

2. Can be named anything

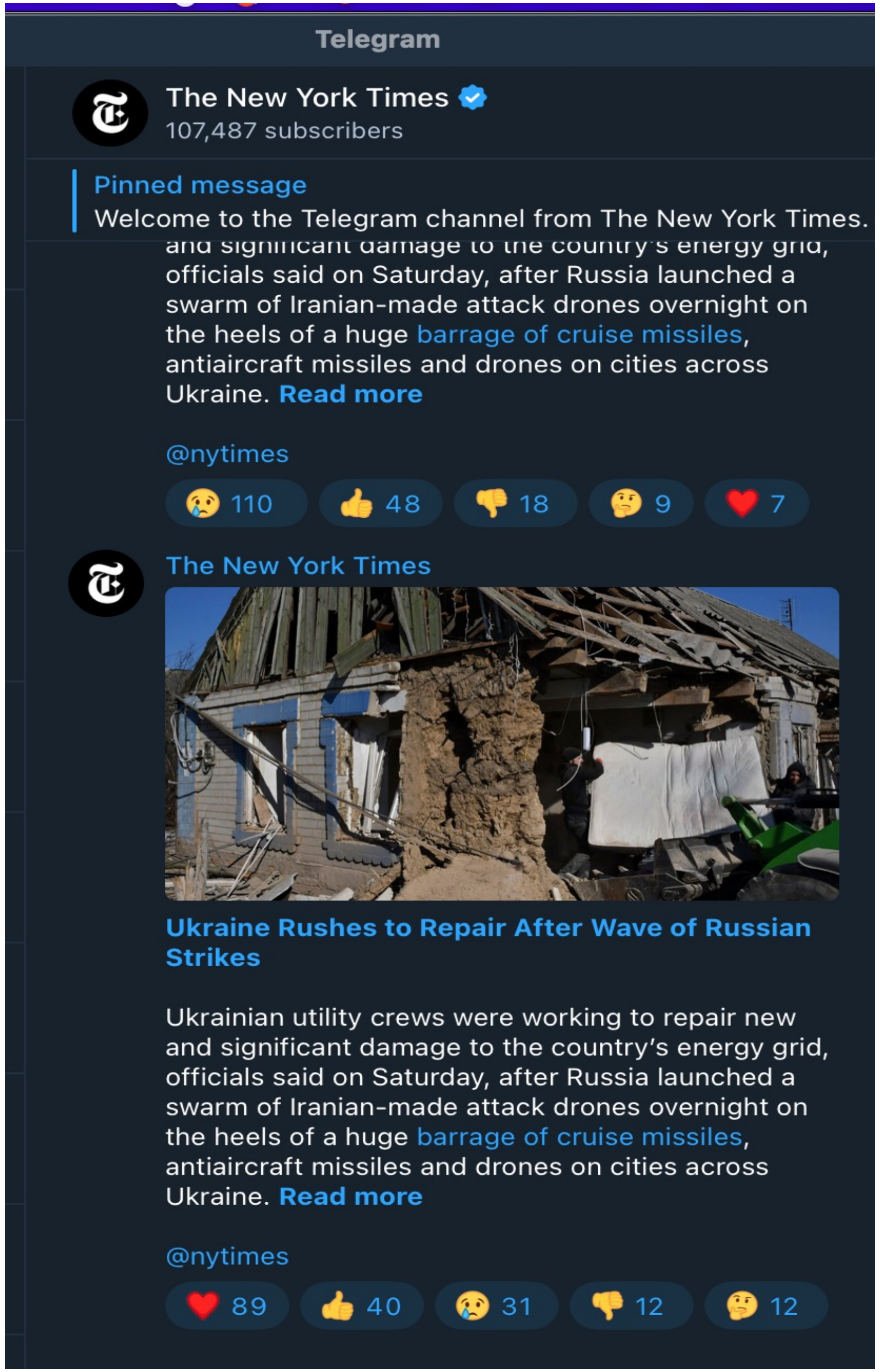
3. File will be created in the working directory



```
chat_data_collection
    __pycache__
    json_channel_data
    json_chat_data
    processed_data
    .env
    anon.session
    collect_all_messages.py
    collect_forwards.py
    config.py
    json_to_df.py
```



```python
import os
from dotenv import load_dotenv

# load enviroment variables from the .env file
load_dotenv()

#
class Config:
    api_id = os.getenv("TELEGRAM_API_ID")
    api_hash = os.getenv("TELEGRAM_API_HASH")
    phone = os.getenv("PHONE_NUM")
    session_name = "anon.session"

    json_chat_dir = "json_chat_data"
    processed_data_dir = "processed_data"
    chat_dfs_dir = "chat_dfs"
    fwds_master_file = "fwds_master.pkl"
```



Telegram

The New York Times
107,487 subscribers

Pinned message
Welcome to the Telegram channel from The New York Times.

and significant damage to the country's energy grid, officials said on Saturday, after Russia launched a swarm of Iranian-made attack drones overnight on the heels of a huge barrage of cruise missiles, antiaircraft missiles and drones on cities across Ukraine. Read more

@nytimes

😢 110 👍 48 👎 18 🤔 9 ❤️ 7

The New York Times

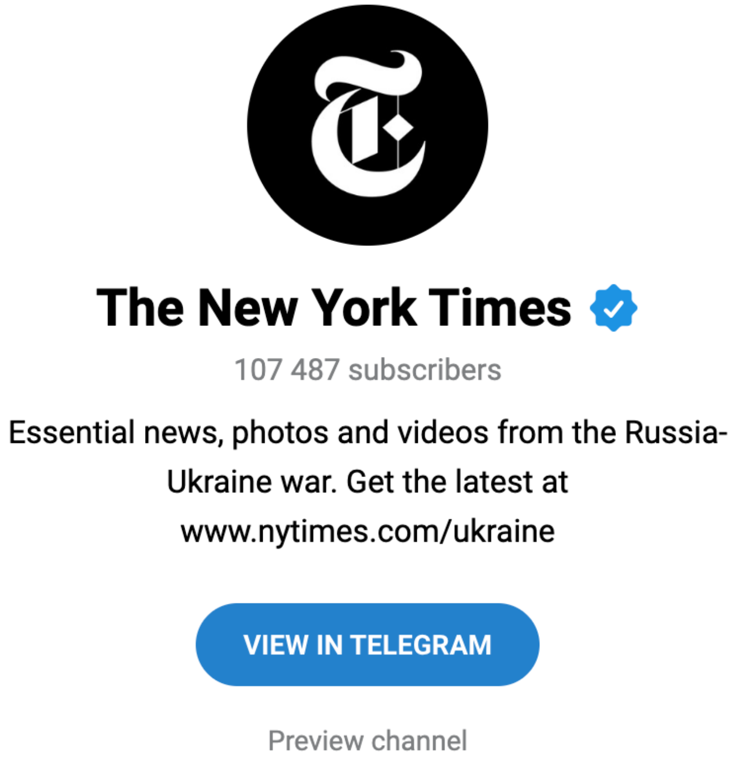
Ukraine Rushes to Repair After Wave of Russian Strikes

Ukrainian utility crews were working to repair new and significant damage to the country's energy grid, officials said on Saturday, after Russia launched a swarm of Iranian-made attack drones overnight on the heels of a huge barrage of cruise missiles, antiaircraft missiles and drones on cities across Ukraine. Read more

@nytimes

❤️ 89 👍 40 😢 31 👎 12 🤔 12

- Ex: https://t.me/nytimes



The New York Times

107 487 subscribers

Essential news, photos and videos from the Russia-Ukraine war. Get the latest at www.nytimes.com/ukraine

Collect all messages

1. You will need to
   1. provide a channel/group username as the input argument.
   2. The type of chat –channel or group chat

2. The below example will collect all the messages from New York Time's telegram channel (viewable at t.me/nytimes) and save the json.

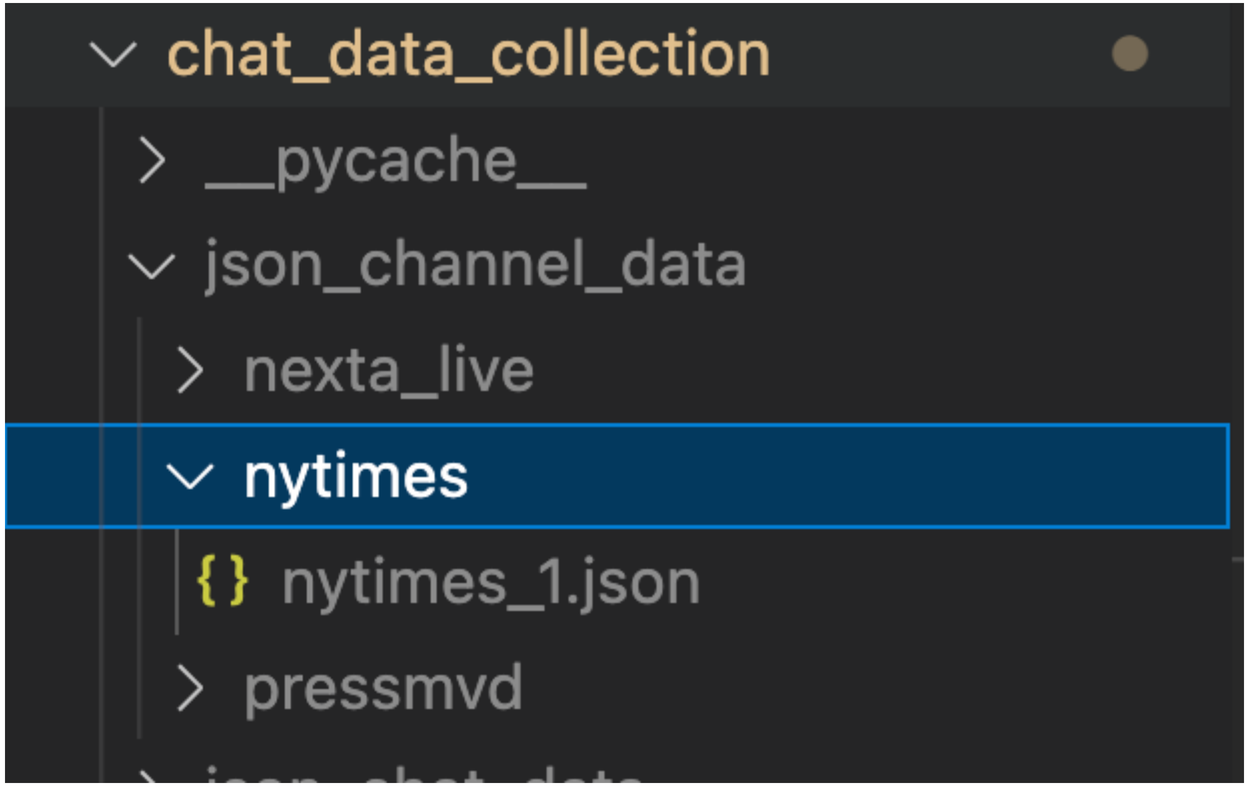
3. The first time, you will be prompted to input your phone number and authenticate by providing the code sent to you on Telegram app.

4. Each json file will contain a maximum of 10000 messages and will be suffixed by the message id of the last post collected.

5. There will be multiple json files if there are more than 10000 messages to collect.

4.The script creates a folder with the same name as the chat and stores all json files in it.

5.The chat folder is stored within the "json_chat_data" (see config.py).



```
∨ chat_data_collection
  > __pycache__
  ∨ json_channel_data
    > nexta_live
    ∨ nytimes
      {} nytimes_1.json
    > pressmvd
  > json_chat_data
```

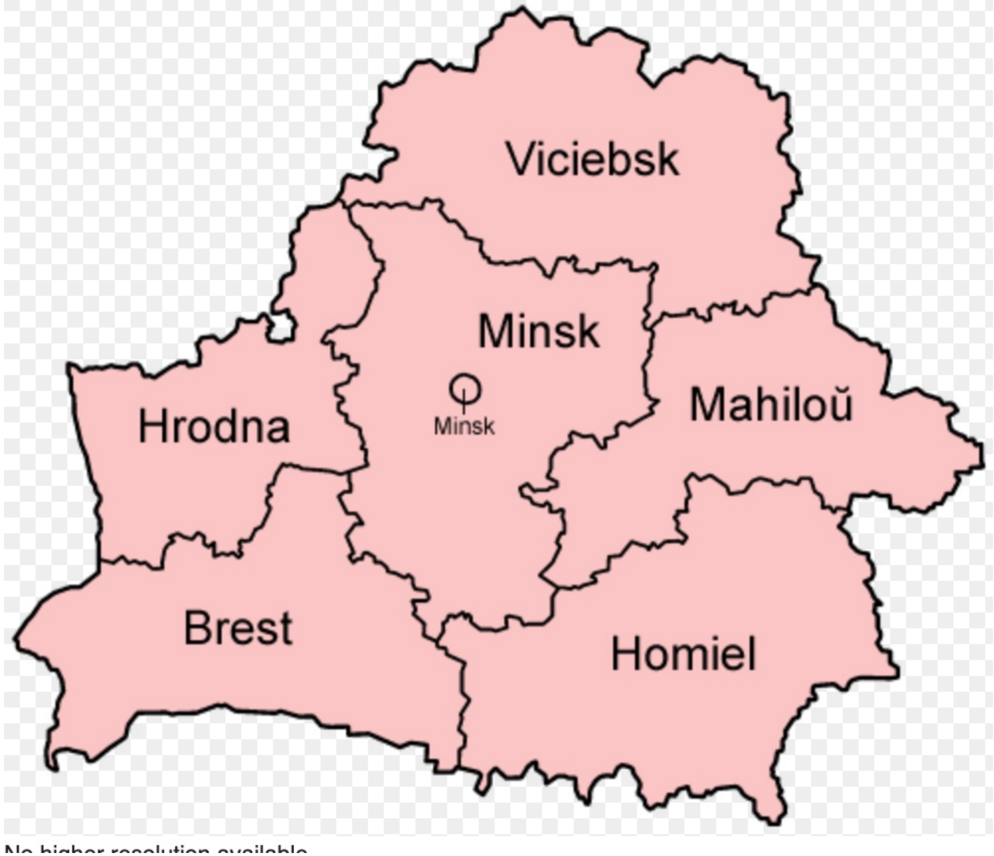
Limitations

- Getting a random sample is difficult
- Need to make decisions on how to define a "document"
  - The length of message can vary widely Eg: in my data the range was 1-40,000 characters
  - The output is a conversation in group chats

Advantages

- Its free!☺
- You can get longer histories
- You can study a platform that is being increasingly used to organize protests (especially useful where Twitter is not widely used)
- The data returned is very similar to response which Twitter returns.

Project: Initial Data

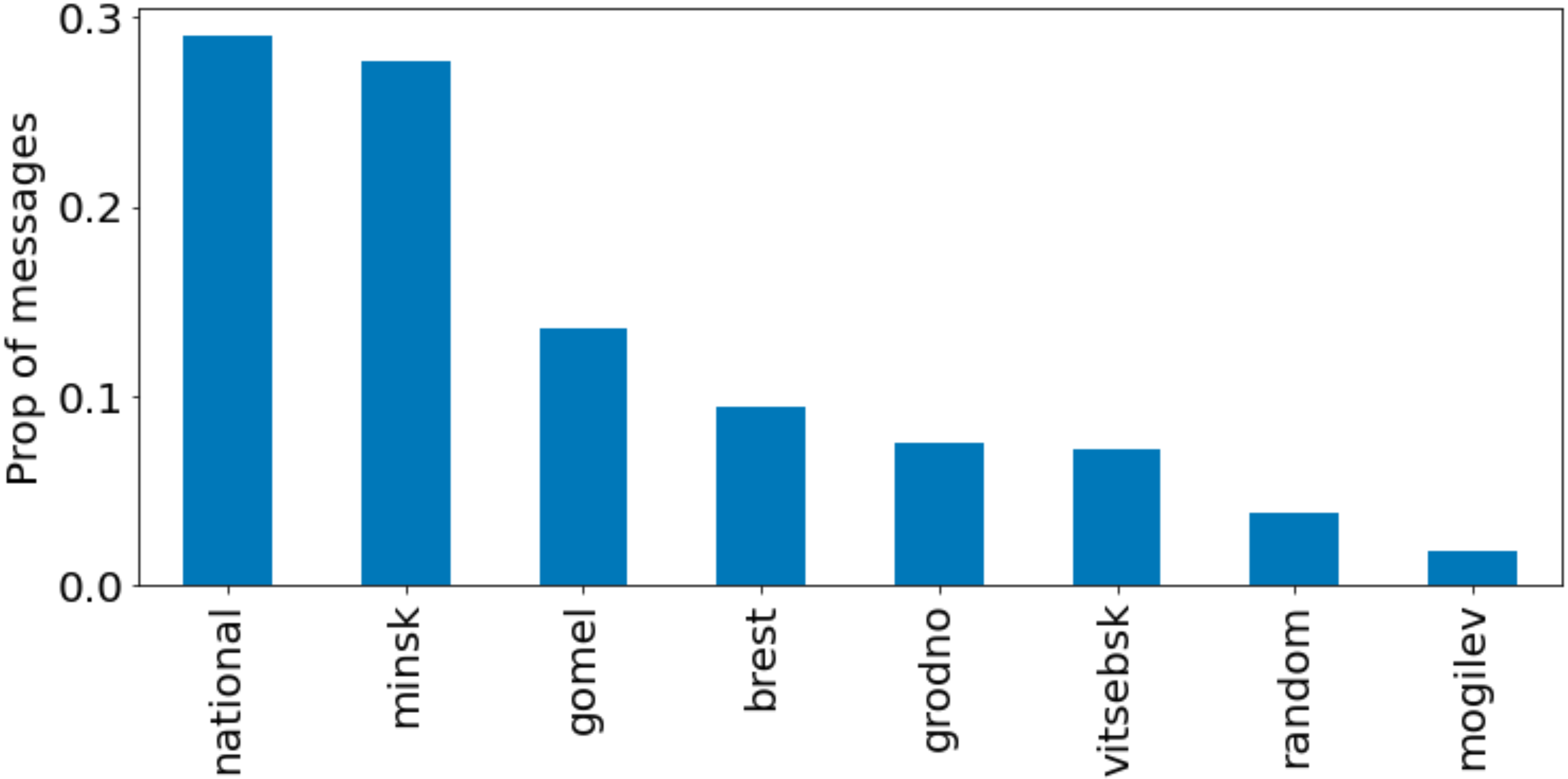


- There are 6 main administrative regions in Belarus.

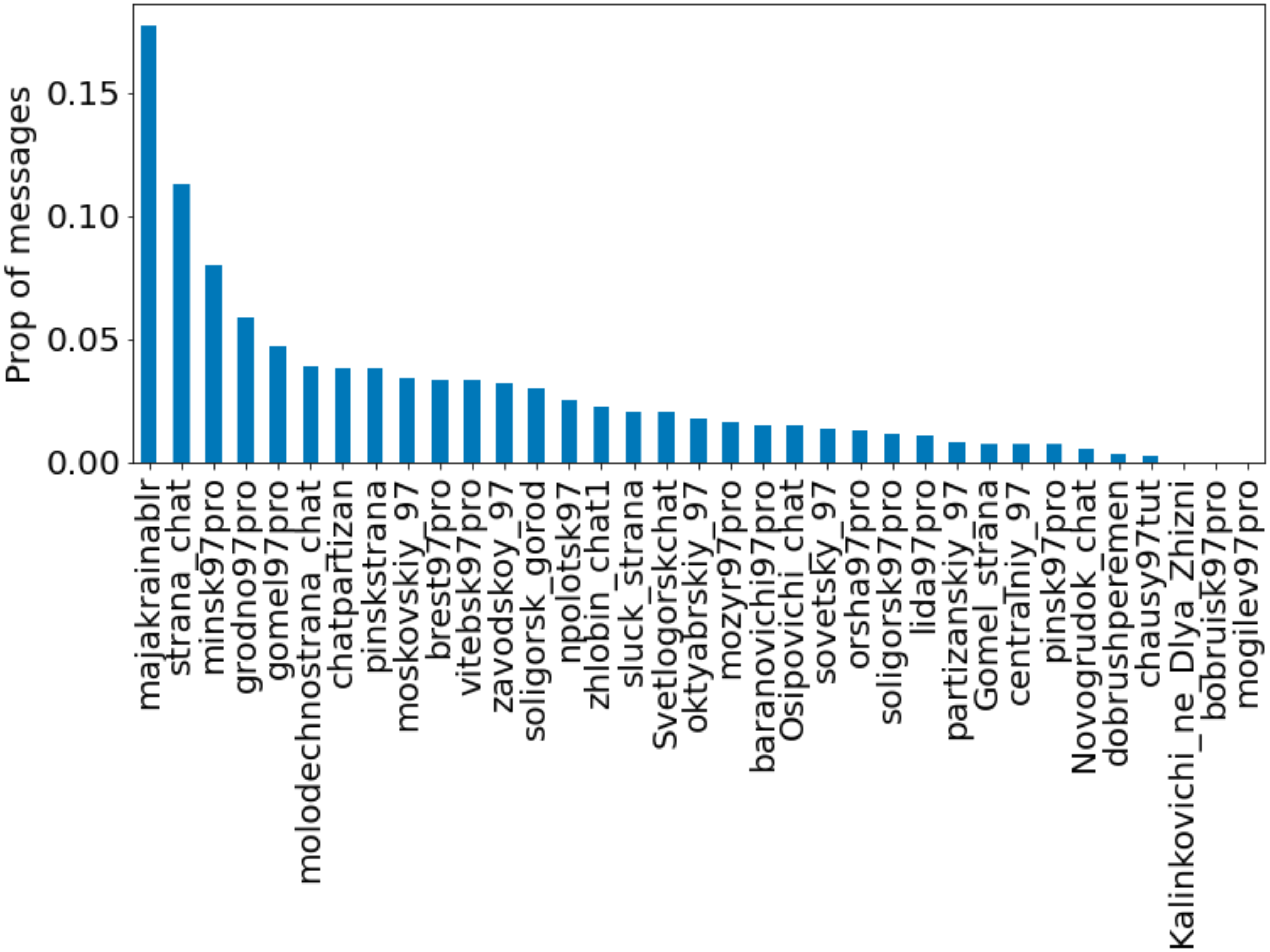
  The capital, Minsk, is in the Minsk Region.

- I have 2 Million messages from 35 group chats spanning:
- Region specific group chats: @minsk97pro (dedicated to the Minsk region, @gomel97pro (dedicated to the Gomel)
- City specific group chats: @lida97pro (a city in Grodno region)
- National level group chats: @strana_chat

Volume of messages by region:



Volume of messages by group chat:

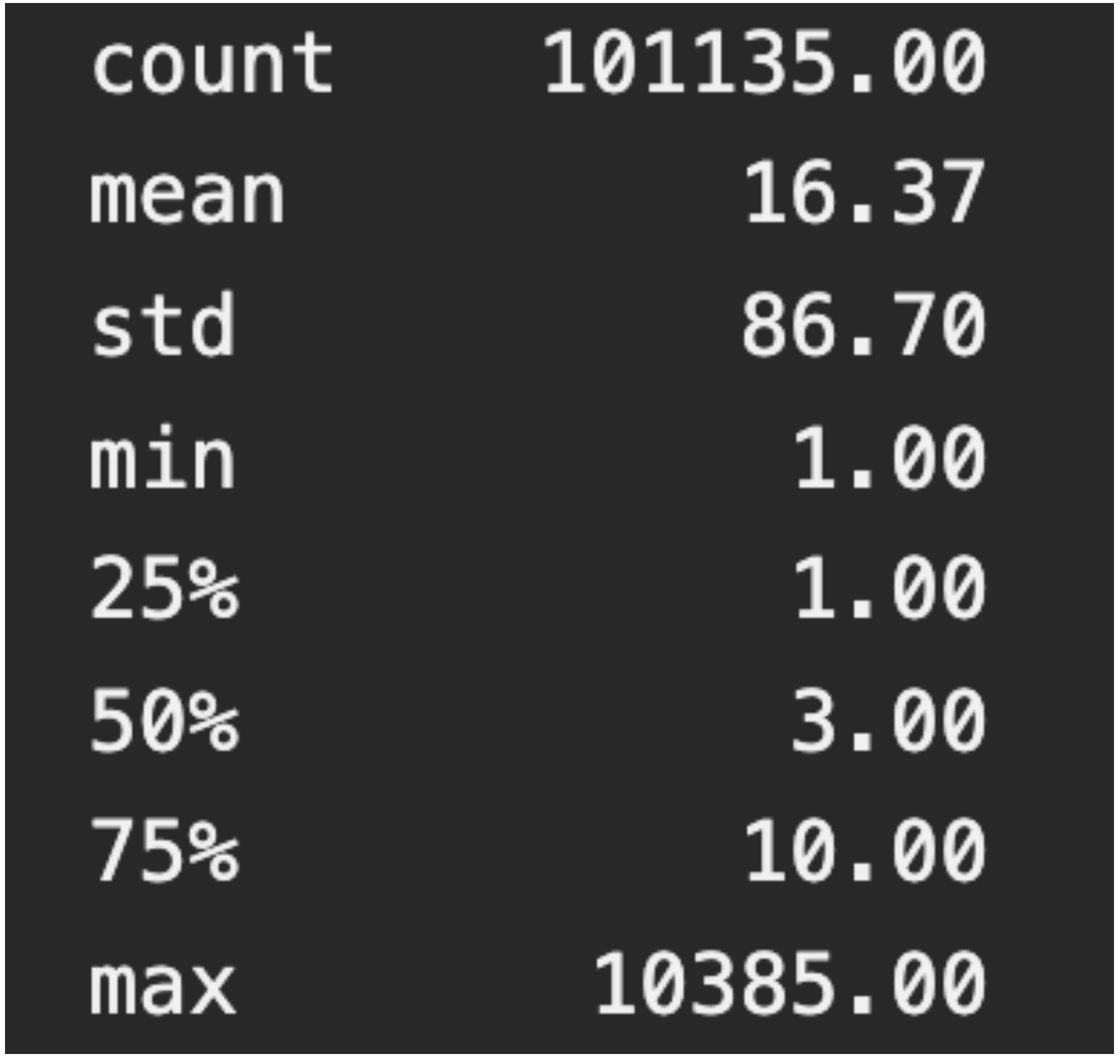


- Why the suffix Pro97?
- **Charter 97** is a declaration calling for democracy in Belarus.
- “devotion to the principles of independence, freedom and democracy, respect to the human rights, solidarity with everybody, who stands for elimination of dictatorial regime and restoration of democracy in Belarus.”

Those who post:

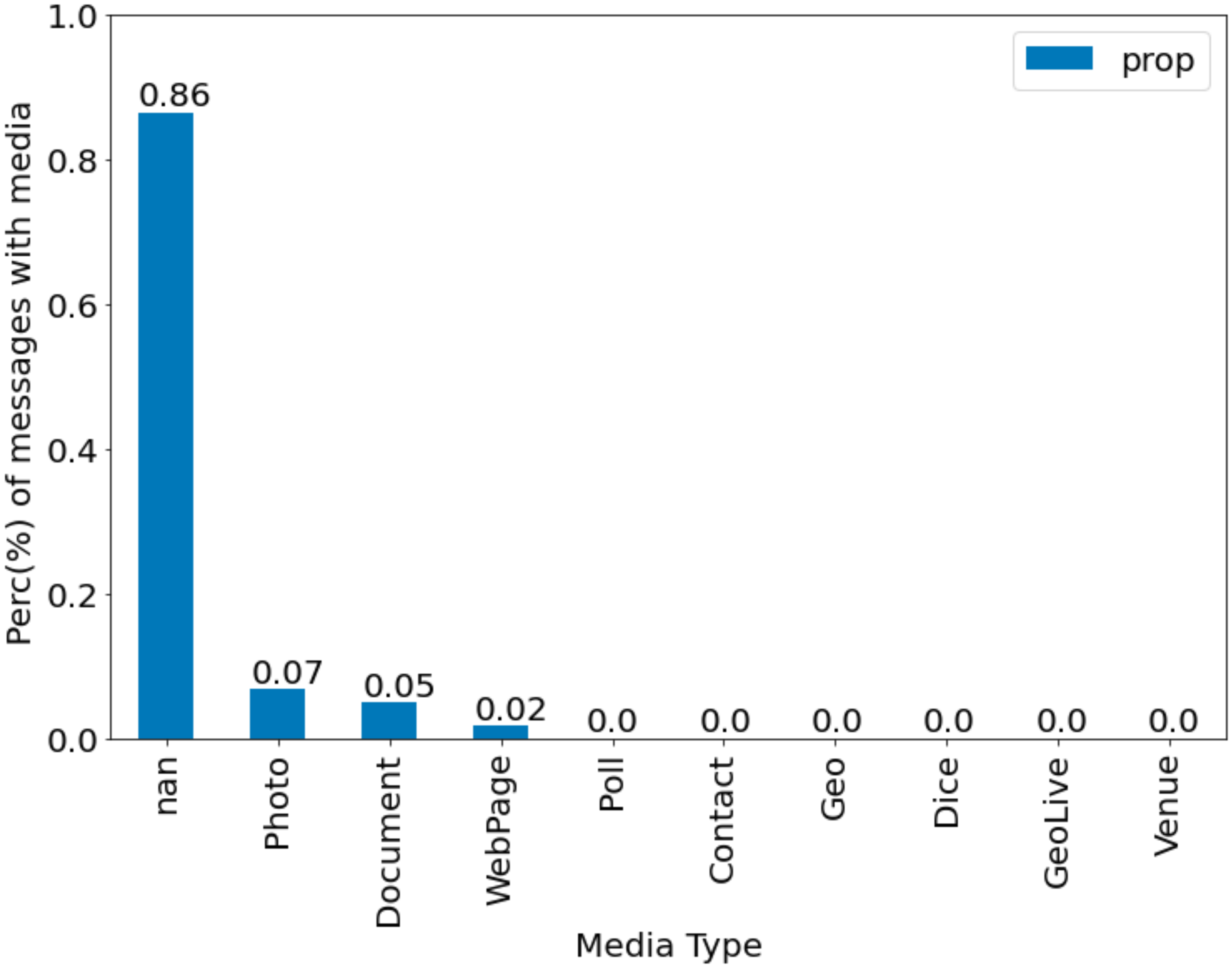
- There are 101135 accounts which have posted at least once.

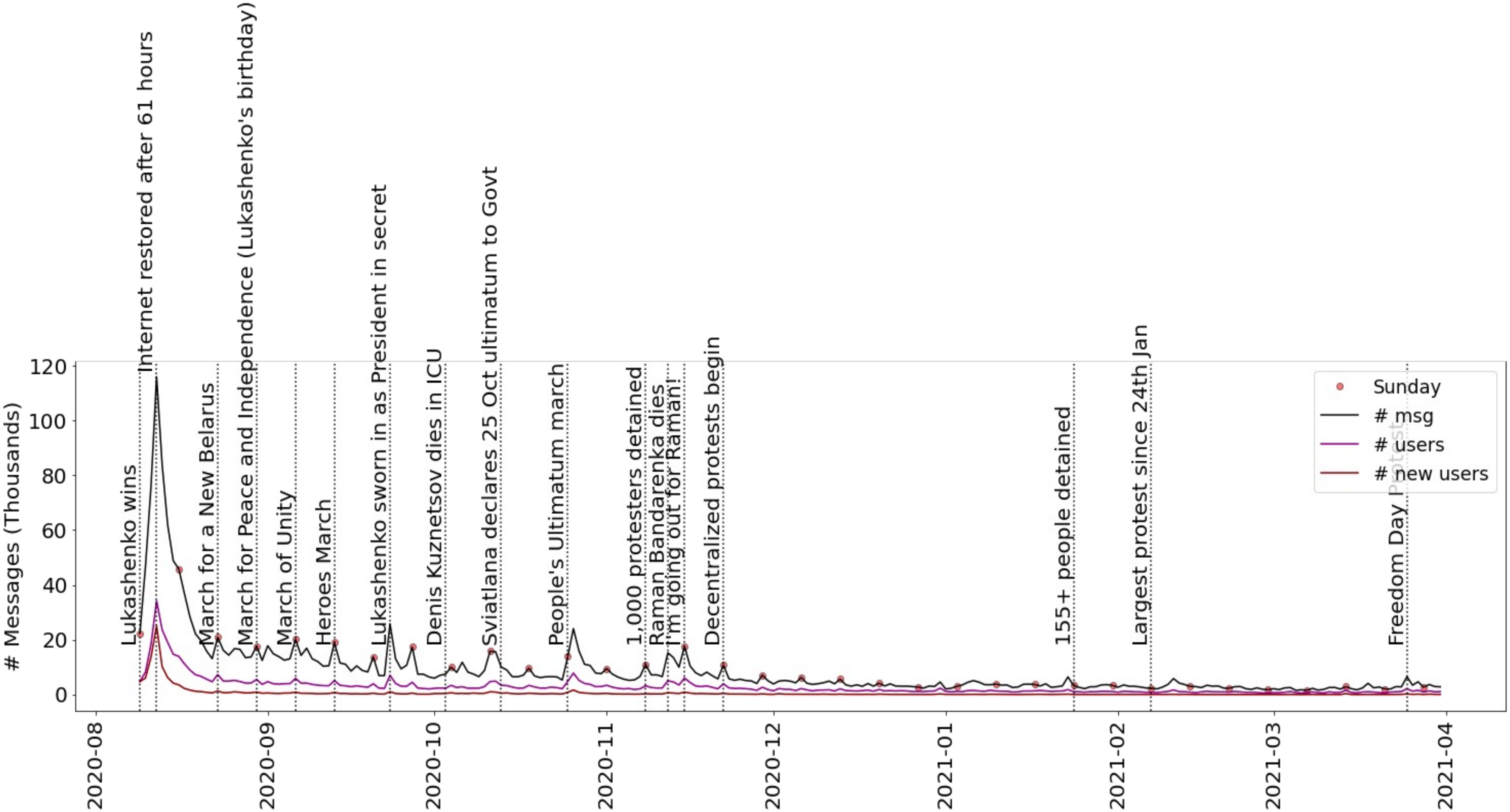
- 50% of users have
- Posted 3 times or less.



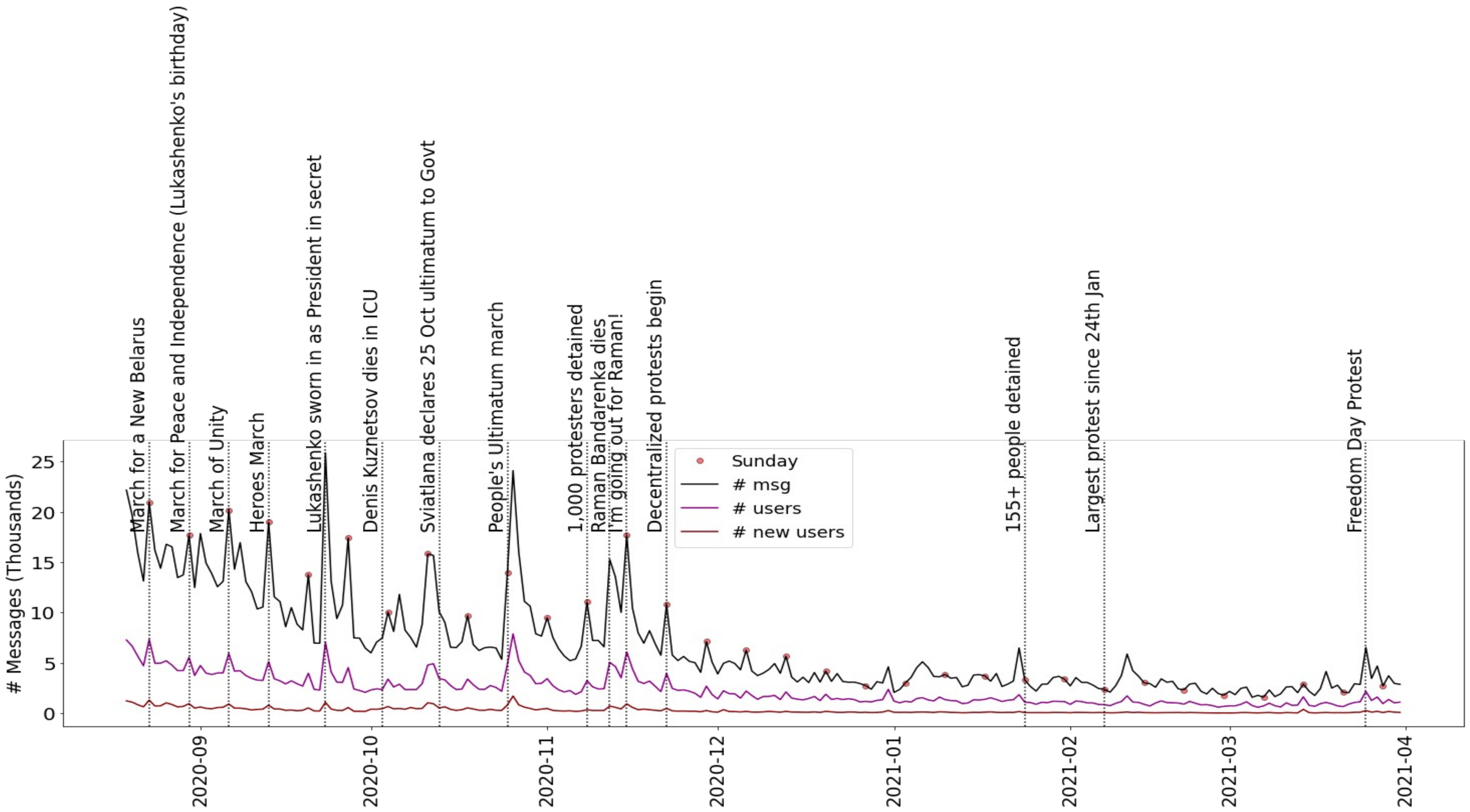
```
count     101135.00
mean          16.37
std           86.70
min            1.00
25%            1.00
50%            3.00
75%           10.00
max        10385.00
```

Media in messages:

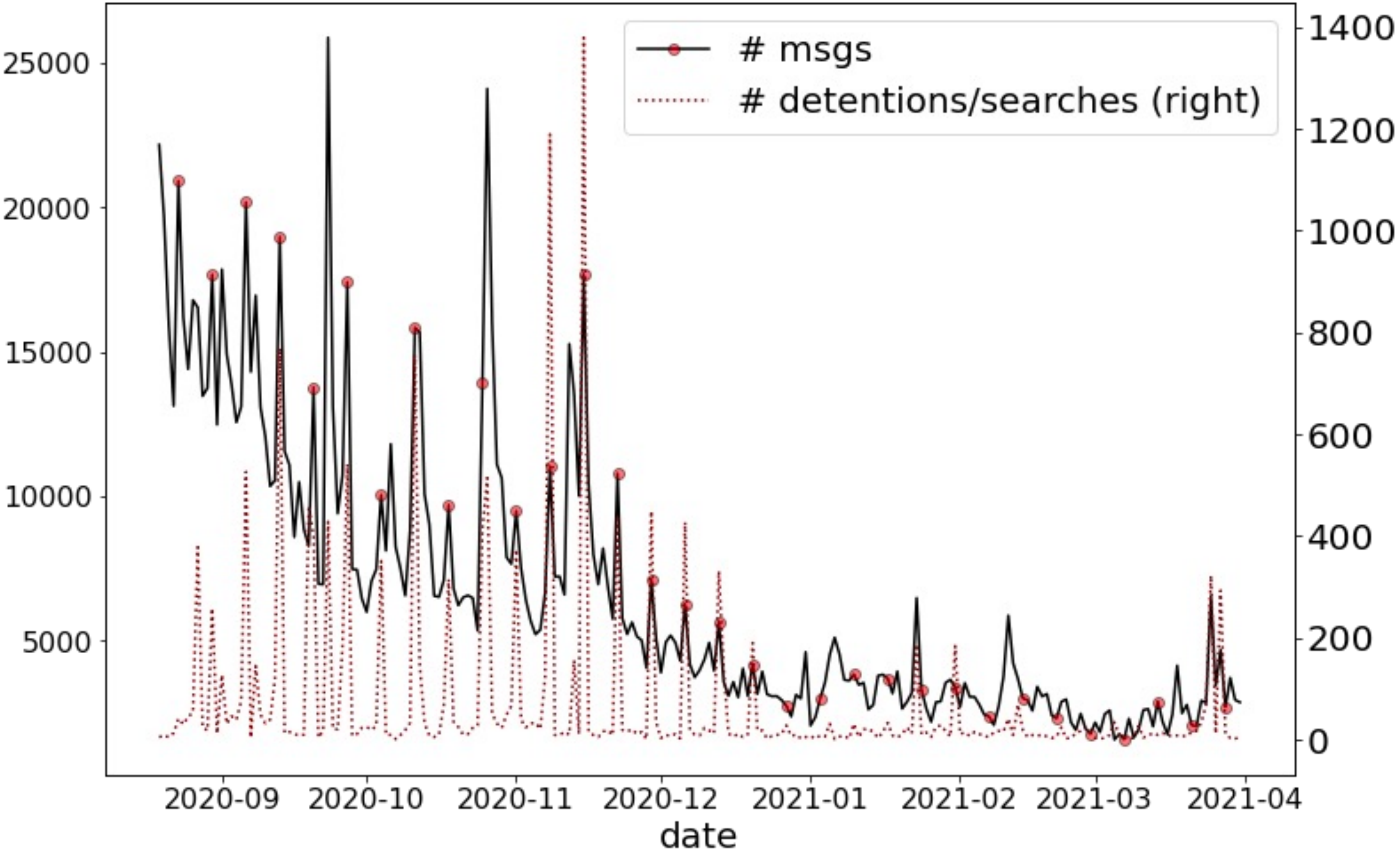


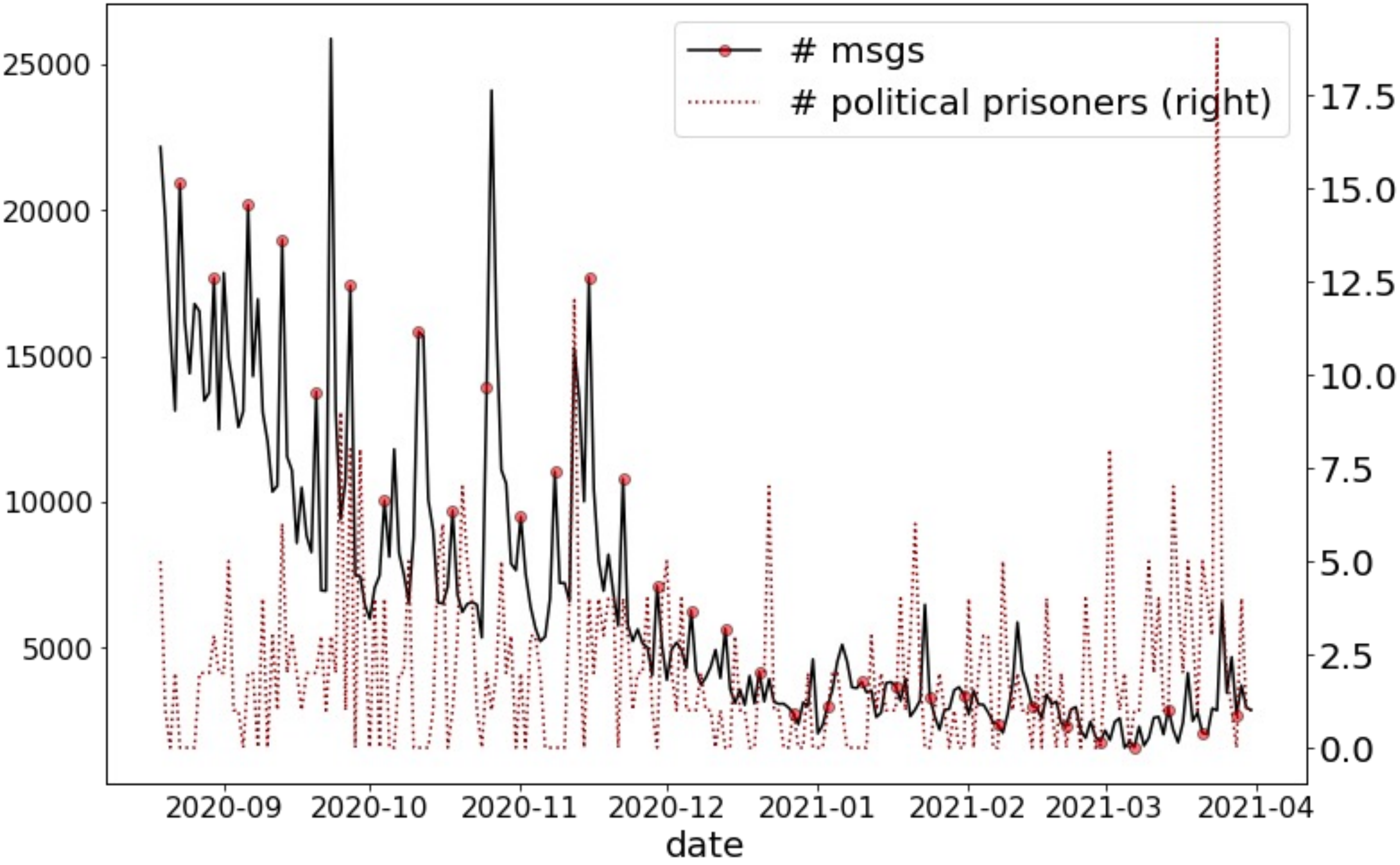


Lukashenko wins
Internet restored after 61 hours
March for a New Belarus
March for Peace and Independence (Lukashenko's birthday)
March of Unity
Heroes March
Lukashenko sworn in as President in secret
Denis Kuznetsov dies in ICU
Sviatlana declares 25 Oct ultimatum to Govt
People's Ultimatum march
1,000 protesters detained
Raman Bandarenka dies
I'm going out for Raman!
Decentralized protests begin
155+ people detained
Largest protest since 24th Jan
Freedom Day Protest
# Messages (Thousands)
0
20
40
60
80
100
120
2020-08
2020-09
2020-10
2020-11
2020-12
2021-01
2021-02
2021-03
2021-04
Sunday
# msg
# users
# new users


March for a New Belarus
March for Peace and Independence (Lukashenko's birthday)
March of Unity
Heroes March
Lukashenko sworn in as President in secret
Denis Kuznetsov dies in ICU
Sviatlana declares 25 Oct ultimatum to Govt
People's Ultimatum march
1,000 protesters detained
Raman Bandarenka dies
I'm going out for Raman!
Decentralized protests begin
155+ people detained
Largest protest since 24th Jan
Freedom Day Protest
Sunday
# msg
# users
# new users
# Messages (Thousands)
25
20
15
10
5
0
2020-09
2020-10
2020-11
2020-12
2021-01
2021-02
2021-03
2021-04

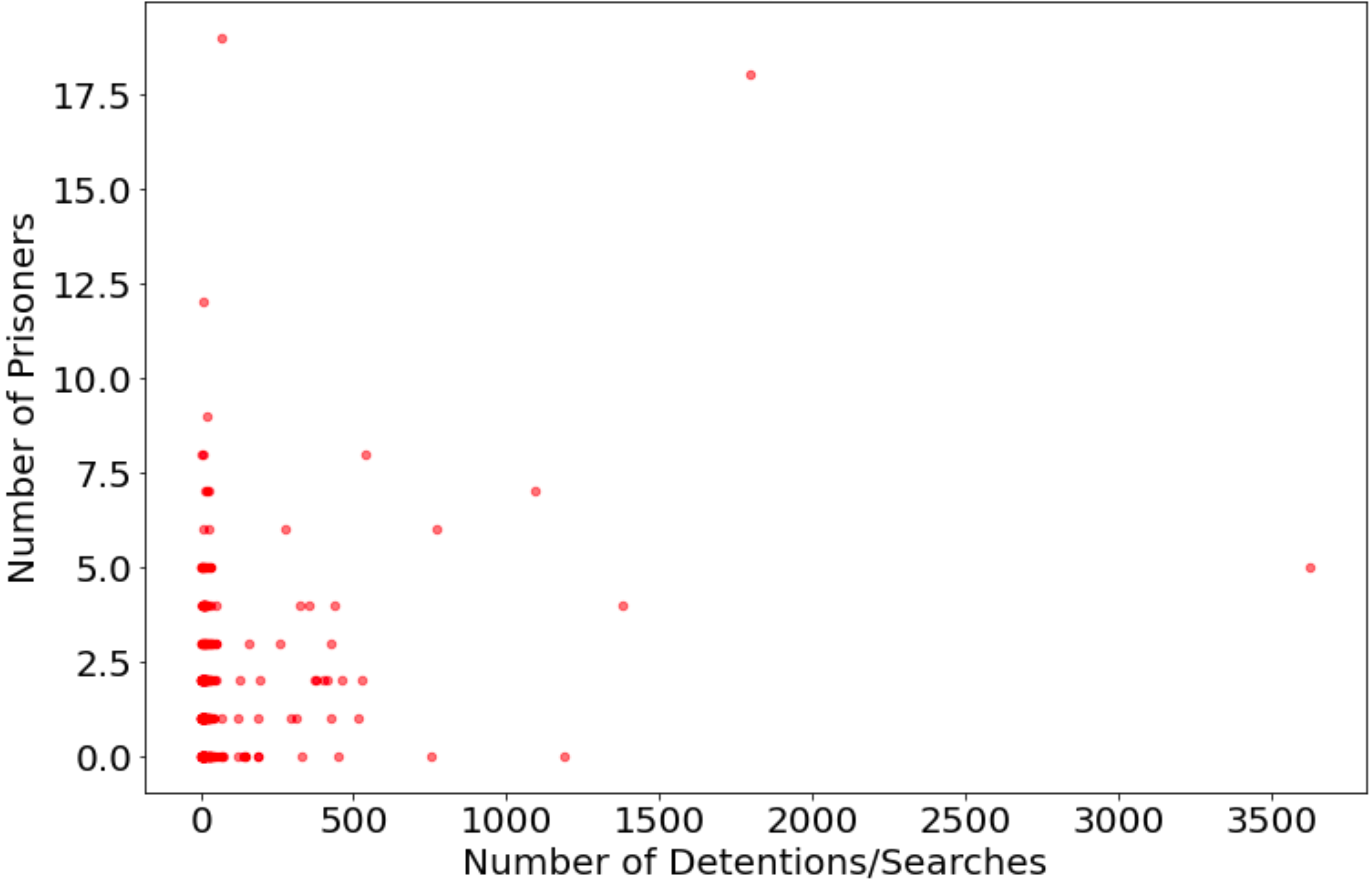
Repression data from *Viasna* a human rights group in Belarus



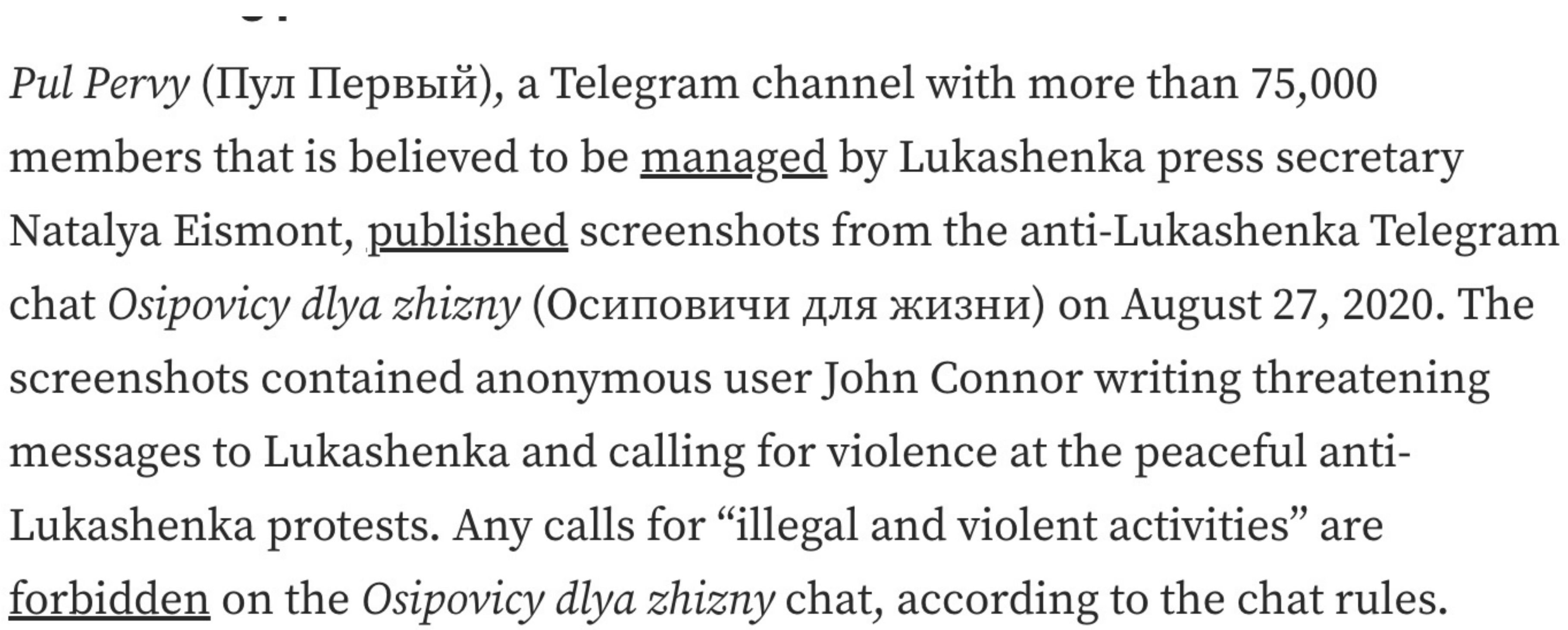


# msgs
# political prisoners (right)
25000
20000
15000
10000
5000
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
2020-09
2020-10
2020-11
2020-12
2021-01
2021-02
2021-03
2021-04
date

Correlation between Detentions and political imprisonment is low = 0.25



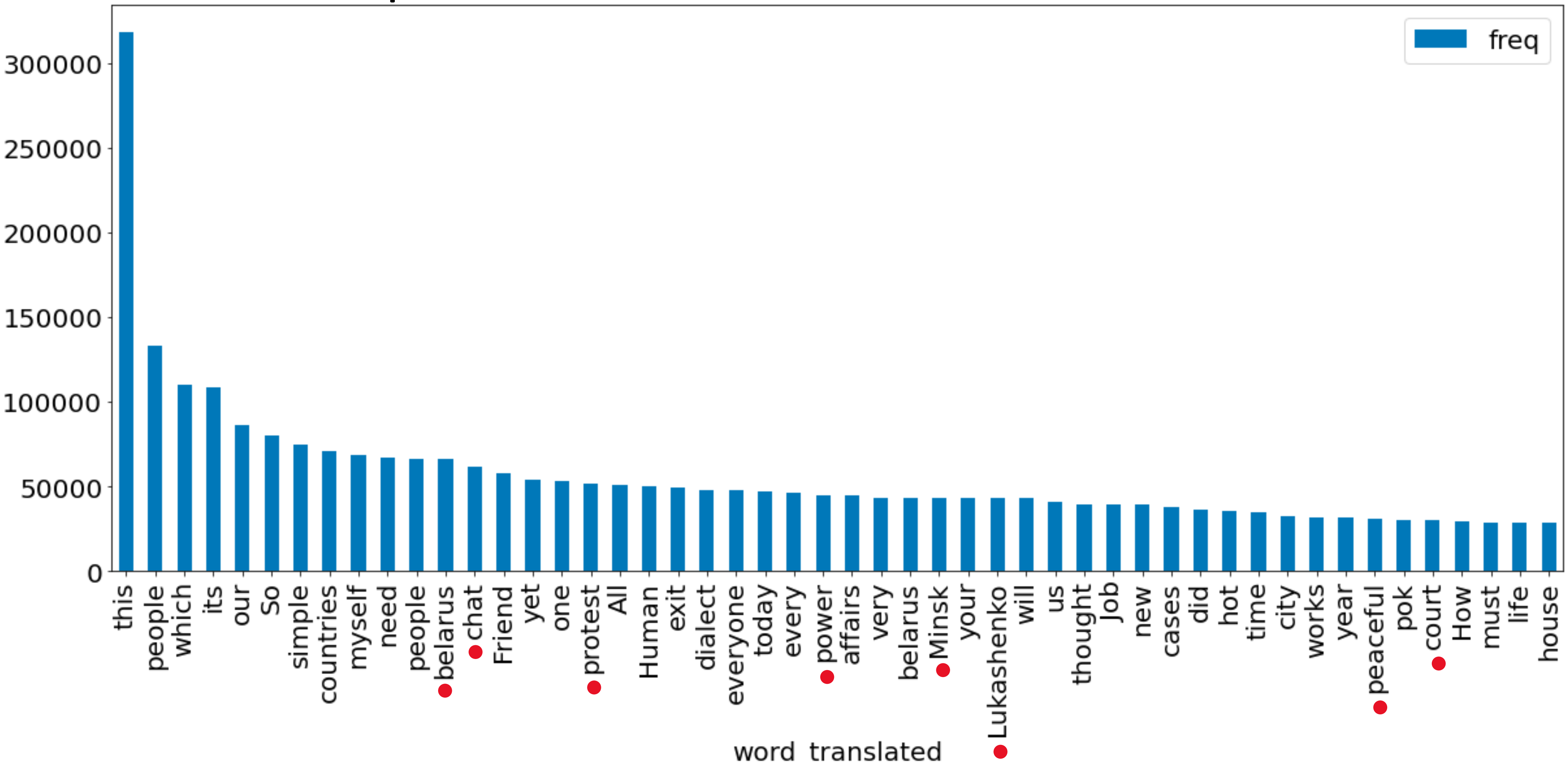
Government Surveillance Example:



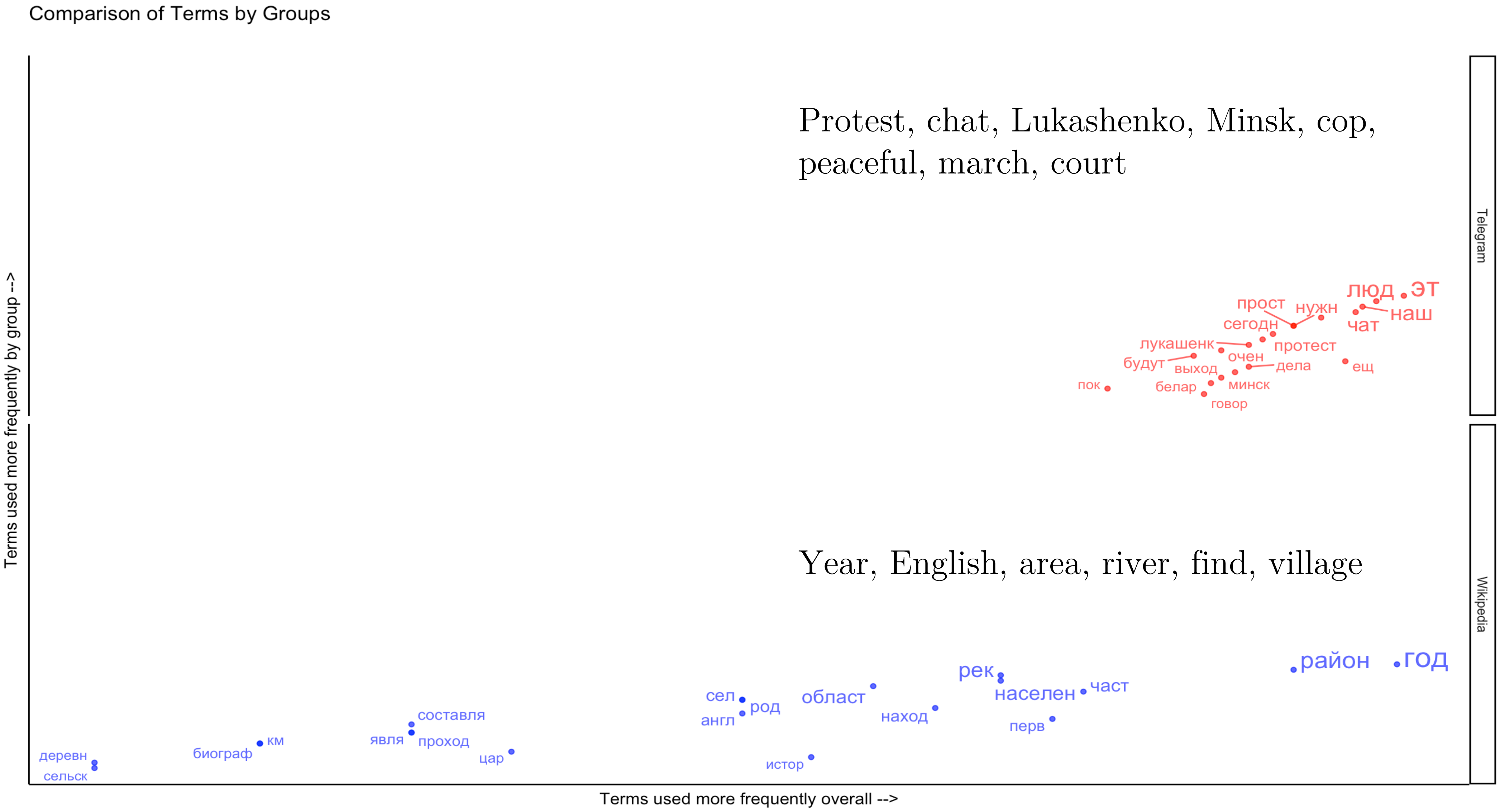
Pul Pervy (Пул Первый), a Telegram channel with more than 75,000 members that is believed to be managed by Lukashenka press secretary Natalya Eismont, published screenshots from the anti-Lukashenka Telegram chat *Osipovicy dlya zhizny* (Осиповичи для жизни) on August 27, 2020. The screenshots contained anonymous user John Connor writing threatening messages to Lukashenka and calling for violence at the peaceful anti-Lukashenka protests. Any calls for “illegal and violent activities” are forbidden on the *Osipovicy dlya zhizny* chat, according to the chat rules.

Does political repression (eg: imprisonment) increase before focal events (planned Sunday protests) in response to discussions in these group chats?

Most frequent terms



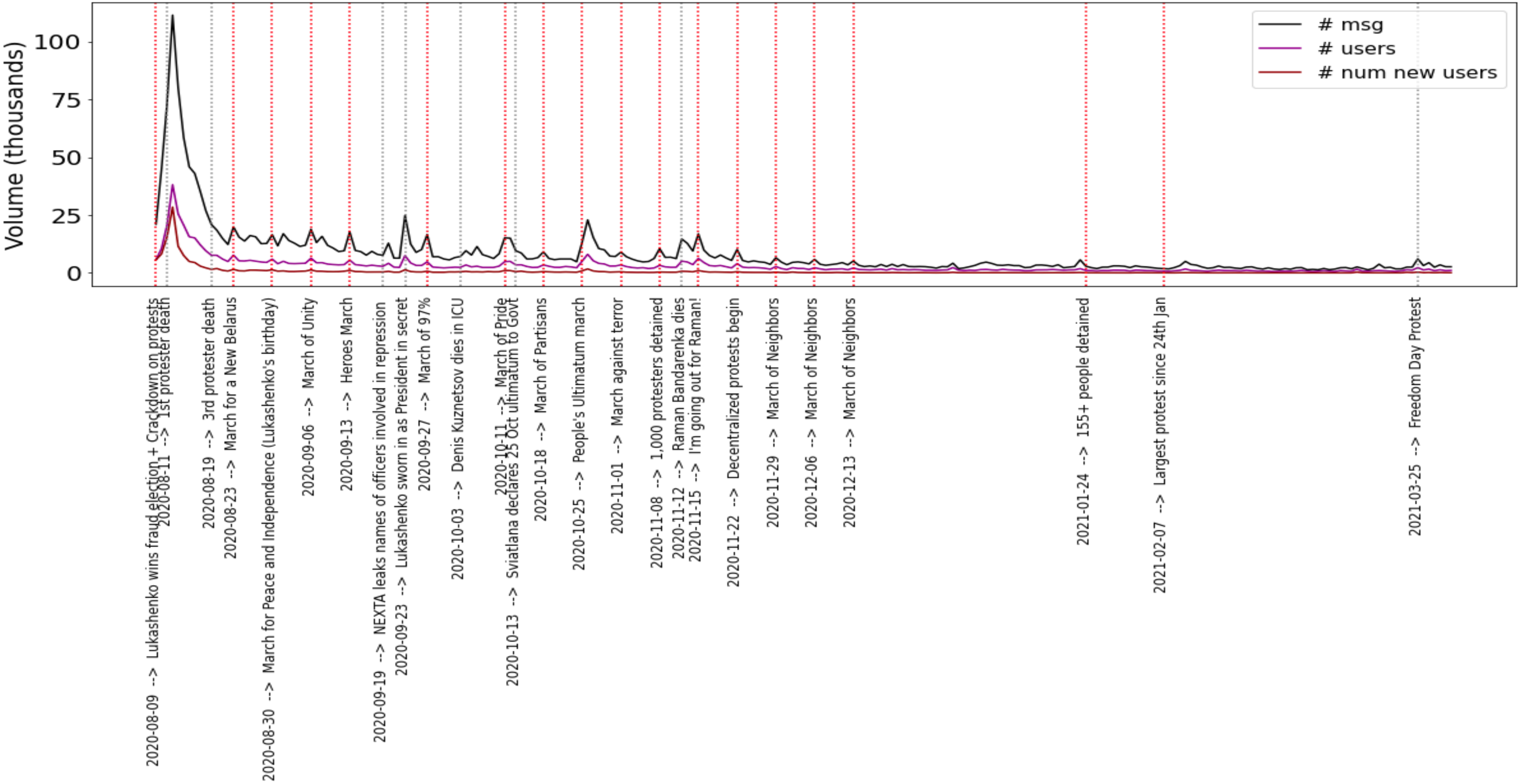
Discriminating Words – FW Statistic



Protest, chat, Lukashenko, Minsk, cop, peaceful, march, court

Year, English, area, river, find, village

Group-specific terms are ordered by Fightin' Words statistic (Monroe, et al. 2008)


Volume (thousands)
100
75
50
25
0
# msg
# users
# num new users
2020-08-09 --> Lukashenko wins fraud election + Crackdown on protests
2020-08-11 --> 1st protester death
2020-08-19 --> 3rd protester death
2020-08-23 --> March for a New Belarus
2020-08-30 --> March for Peace and Independence (Lukashenko's birthday)
2020-09-06 --> March of Unity
2020-09-13 --> Heroes March
2020-09-19 --> NEXTA leaks names of officers involved in repression
2020-09-23 --> Lukashenko sworn in as President in secret
2020-09-27 --> March of 97%
2020-10-03 --> Denis Kuznetsov dies in ICU
2020-10-11 --> March of Pride
2020-10-13 --> Sviatlana declares 25 Oct ultimatum to Govt
2020-10-18 --> March of Partisans
2020-10-25 --> People's Ultimatum march
2020-11-01 --> March against terror
2020-11-08 --> 1,000 protesters detained
2020-11-12 --> Raman Bandarenka dies
2020-11-15 --> I'm going out for Raman!
2020-11-22 --> Decentralized protests begin
2020-11-29 --> March of Neighbors
2020-12-06 --> March of Neighbors
2020-12-13 --> March of Neighbors
2021-01-24 --> 155+ people detained
2021-02-07 --> Largest protest since 24th Jan
2021-03-25 --> Freedom Day Protest