

Telegram EDA

Final Project for Computational Social Science 2022

By Taras Kreshchenko
Teacher: Andrew Kurochkin

NaUKMA Subgroup #0

13/12/2022

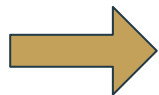
Outline

- Project introduction
- The workflow
- Final results
- Future work


Introduction

The idea: obtain Telegram messaging data, form research questions and try to answer them using this data.

The goal: come to interesting conclusions and create awesome visualisations.



Step 1: Data collection

- The data consists of dialogs, messages, and users from my Telegram, which has been obtained using Telegram API
- Make sure you don't forget your MFA password 
- Took 2 hrs 49 mins to download

219 MB of data, including:

- **672109 messages:**
 - 27361 sent
 - 644748 received
- **220 dialogs:**
 - Private chats
 - Group chats
 - Channels
- **1138 users**

Step 2: Column cleanup & preprocessing

- Making sure I understand the data perfectly
- Remove useless data, fill missing values, preprocess invalid values
- Took me over 4 hours
- Ended up with very pleasant data to work with, in the form of 5 dataframes: df, df_meta, df_combined, df_meta_users, df_meta_dialogs

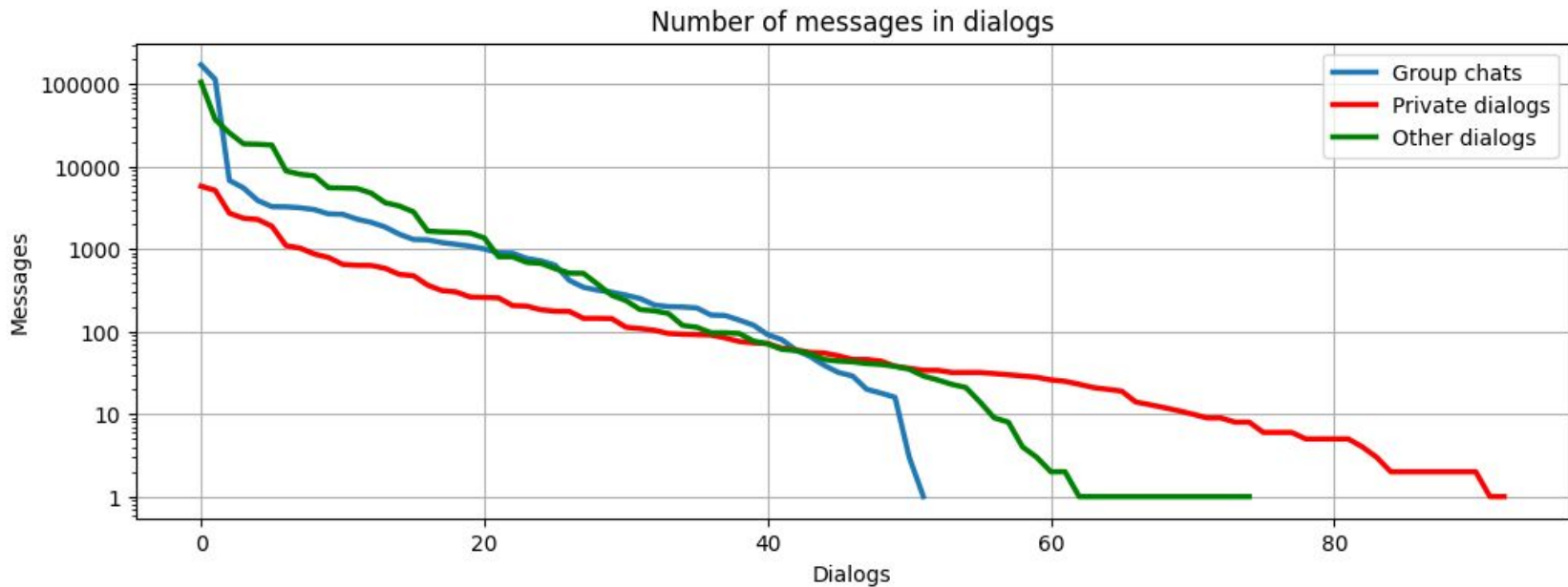
Step 3: Generating research ideas

- In the beginning I was going through each column trying to come up with ideas for possible correlations or patterns.
 - Some columns could have interesting data in them by themselves (message, first_name)
 - Some columns would have to be combined with other ones to reveal correlations (date, type, duration, dialog_id)

Research directions

I ended up with a bunch of different “research lines”:

- Comparing dialog types
- Change of something over time
- Tokenizing messages in order to find trends
- Analysing users (names, bots, character occurrence)
- Monthly trending cities



Q: Is there a difference in the number of messages in private chats, group chats, and channels?

A: There are a lot of private chats with 10+ messages, but generally group chats have the most average messages. Channel contain even more messages than group chats.

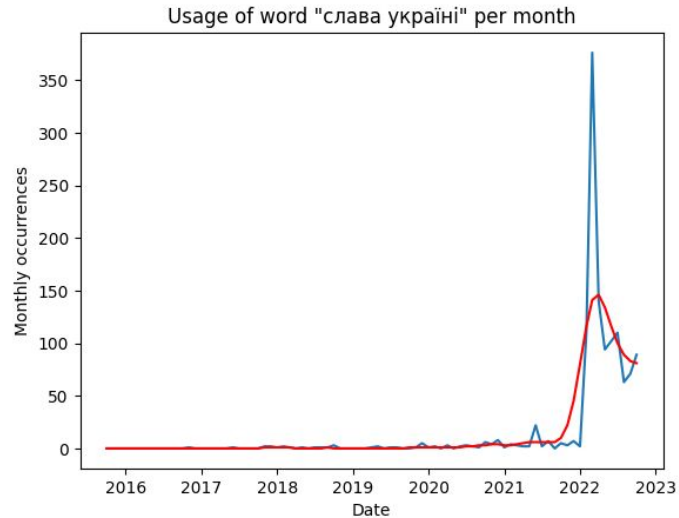
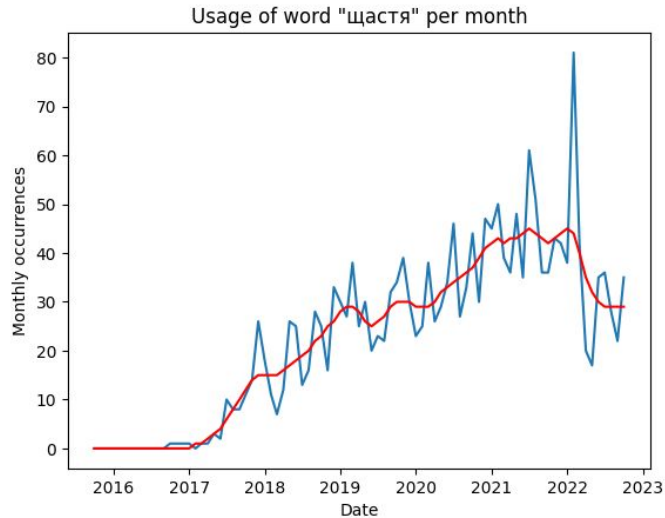
Note: "Other dialogs" includes channels and private chats with deleted users or users without a username.

Change of something over time

- The “date” column can be combined with basically anything to show how it changed as the time went.

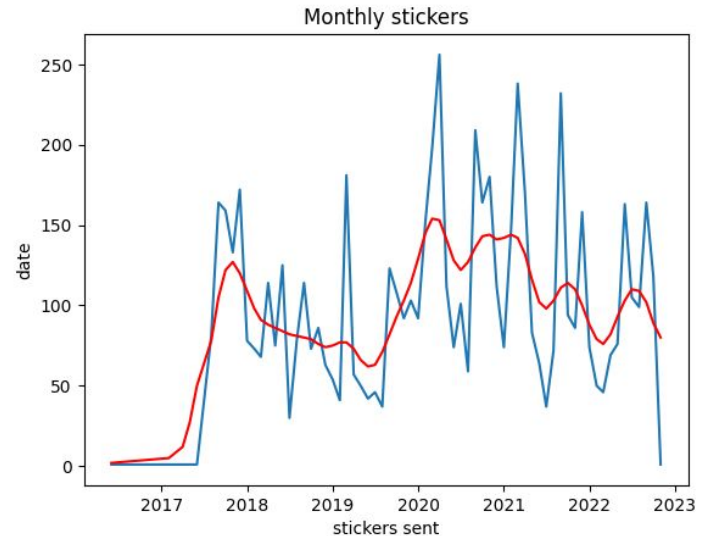
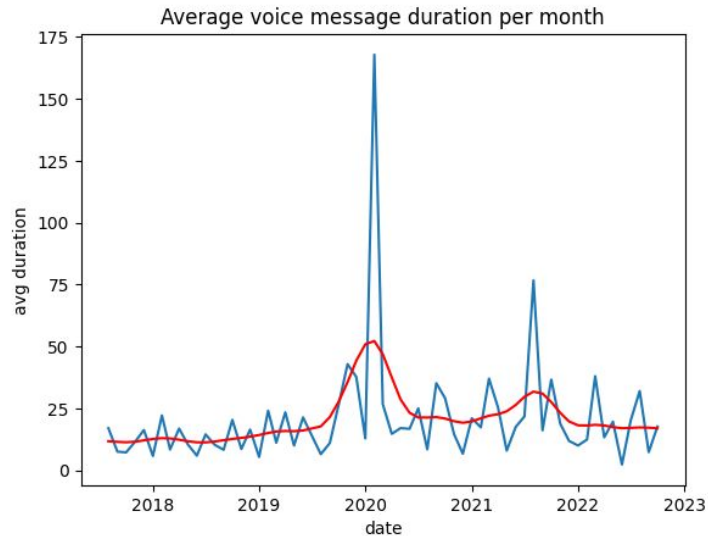
Change of a word/phrase popularity over time

- The “date” column can be combined with basically anything to show how it changed as the time went.



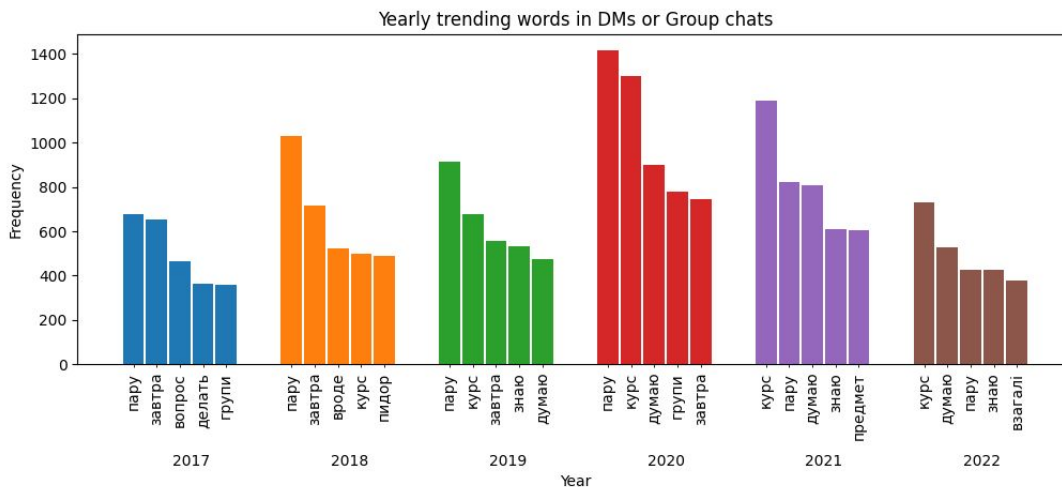
Change of message type popularity over time

- The “date” column can be combined with basically anything to show how it changed as the time went.



Tokenizing messages to find trends

- Split messages into words, then tokenize them
- For each encountered language (Ukrainian, Russian, English):
 - Remove stopwords
 - Apply stemming
- Calculate yearly trends

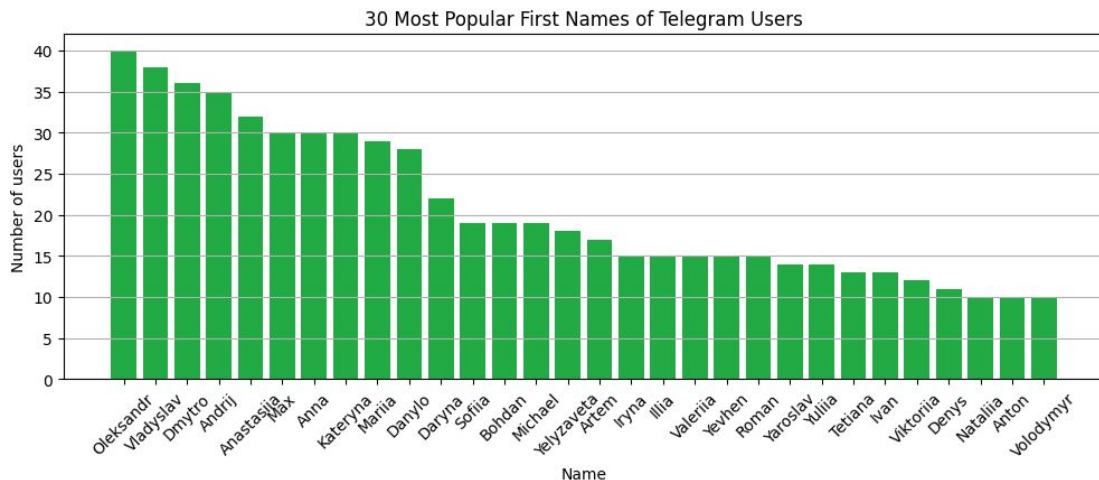


Analysing user data

- Most popular names
- Most used bots
- Letter distribution in different positions in names

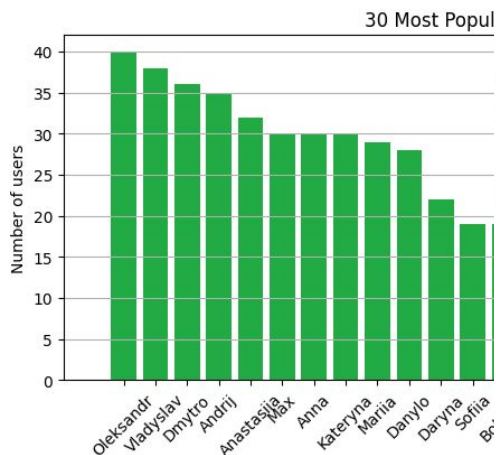
Analysing user data

- Most popular names
- Most used bots
- Letter distribution in different positions in names



Analysing user data

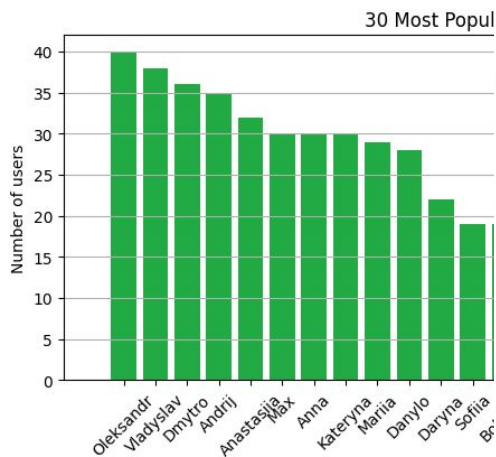
- Most popular names
- Most used bots
- Letter distribution in different positions in names



	first_name	last_name	username	phone
user_id				
232117096	Sublime Bot	None	SublimeBot	None
210944655	Combot	None	combot	None
468253535	True Mafia	None	TrueMafiaBot	None
912522878	FI Post Bot	None	fipostbot	None
84210004	PollBot	None	PollBot	None
1480974941	Meest ПОШТА bot	None	meestposhtomat_bot	None
430091003	KMAScheduler	None	KMASchedulerBot	None
502273726	UptimeRobot	None	officialuptimerobot	None
5202184607	STOP Russian War	None	stop_russian_war_bot	None
702126376	Уведомления Gogetlinks	None	gglinfo_bot	None

Analysing user data

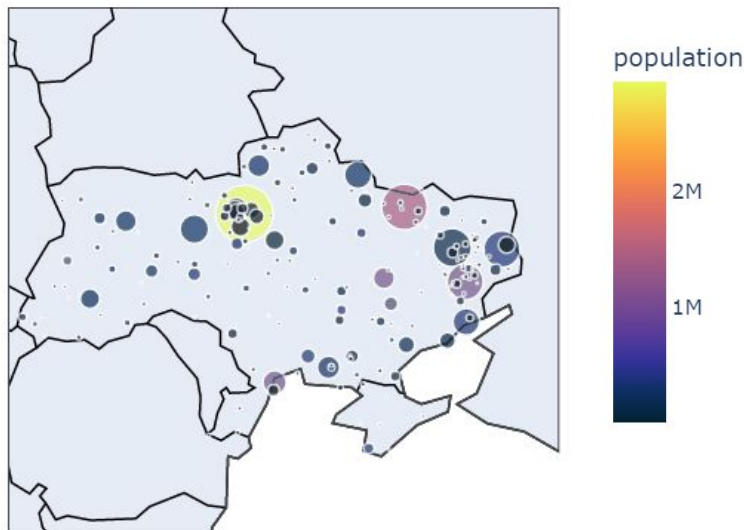
- Most popular names
- Most used bots
- Letter distribution in different positions in names



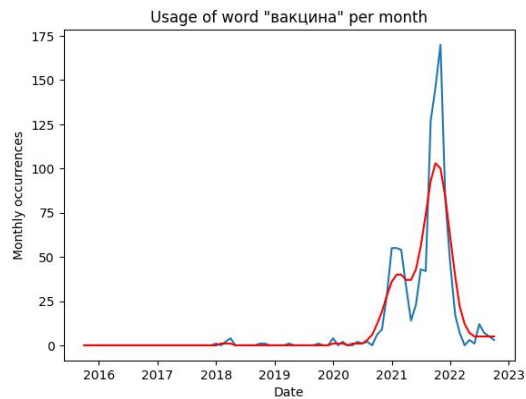
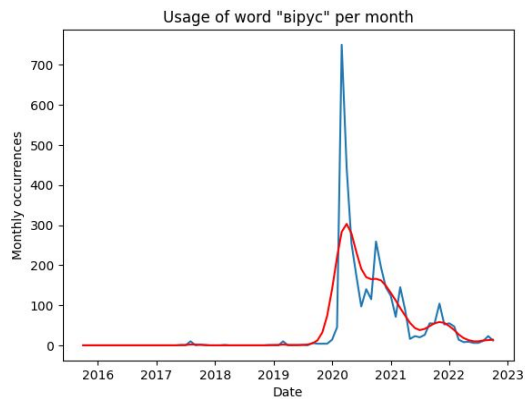
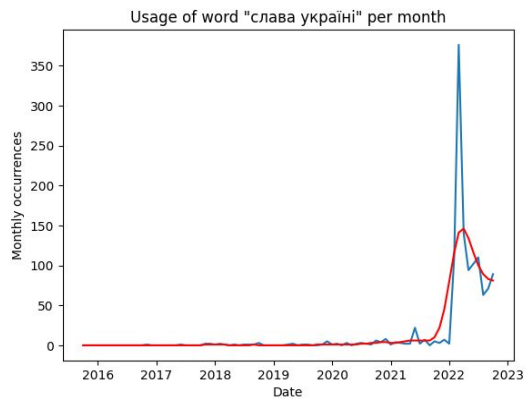
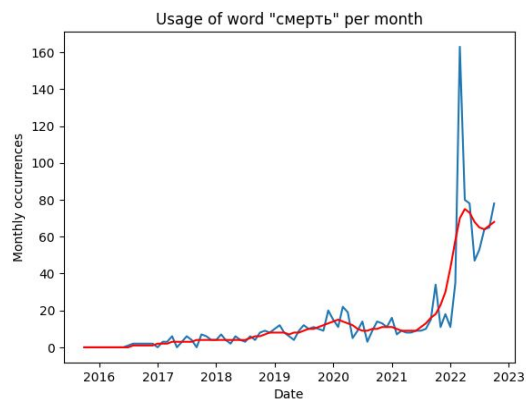
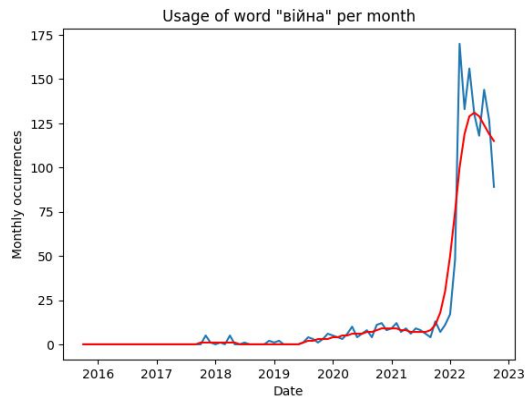
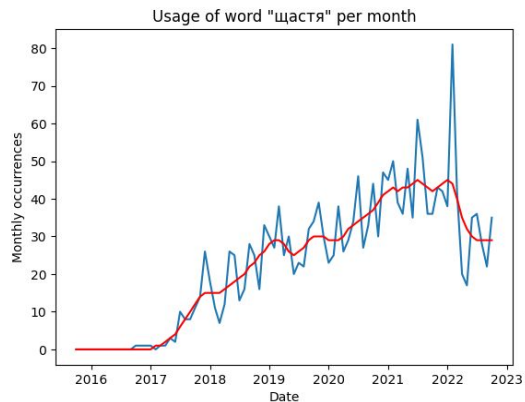
	first_name	last_name		0	1	2	3	4	5	6	
user_id			m	0.077329	0.018453	0.050967	0.022847	0.035149	0.037307	0.030663	0.01
232117096	Sublime Bot	None	a	0.065026	0.222320	0.090510	0.092267	0.119508	0.113740	0.098912	0.12
210944655	Comboto	None	d	0.058875	0.010545	0.028998	0.045694	0.014938	0.020928	0.015826	0.04
468253535	True Mafia	None	s	0.056239	0.017575	0.053603	0.045694	0.050088	0.057325	0.043521	0.04
912522878	FI Post Bot	None	k	0.051845	0.013181	0.045694	0.039543	0.041301	0.048226	0.063304	0.05
84210004	PollBot	None
1480974941	Meest ПОШТА bot	None	7	0.000000	0.000879	0.000879	0.000000	0.002636	0.000000	0.002967	0.00
430091003	KMAScheduler	None	4	0.000000	0.000000	0.000879	0.006151	0.005272	0.004550	0.005935	0.00
502273726	UptimeRobot	None	6	0.000000	0.000000	0.000879	0.000879	0.001757	0.000910	0.001978	0.00
5202184607	STOP Russian War	None	9	0.000000	0.000000	0.000000	0.001757	0.004394	0.000910	0.001978	0.00
702126376	Уведомления Gogetlinks	None	8	0.000000	0.000000	0.000000	0.000000	0.001757	0.001820	0.000989	1

Trending cities

- I found a dataset of Ukrainian populated places, exported from OpenStreetMap. It contained over 200k cities. (https://data.humdata.org/dataset/hotosm_ukr_populated_places)
- Tokenized both Telegram messages and city names in the dataset.
- Dropped:
 - Cities with population smaller than 1000
 - Cities with names that are too similar to other common words (Українськ, Войниха, Танкове, Березне, etc)
- Used Plotly to generate animated visualizations
- Bigger cities have brighter colours
- This took me over 8 hours

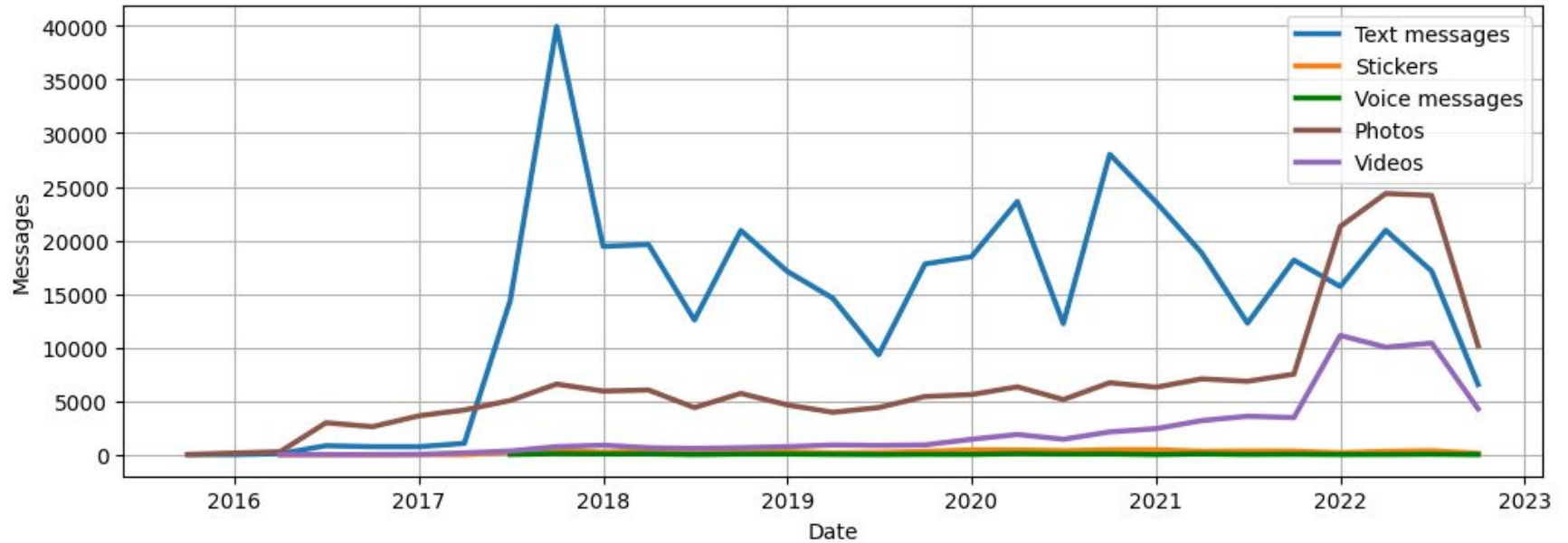


Final Results



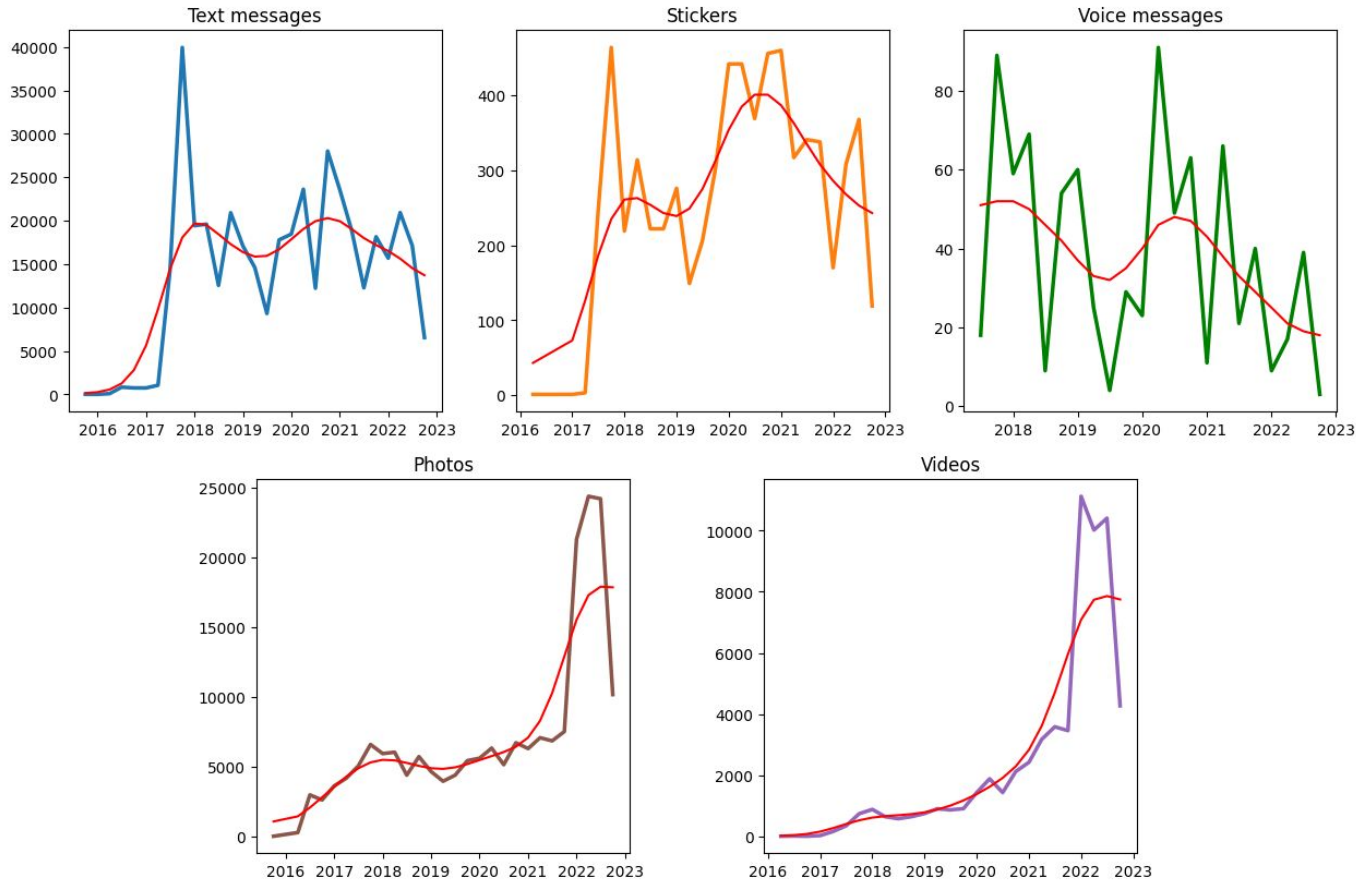
How the popularity of a word/phrase changed over time

The distribution of message types

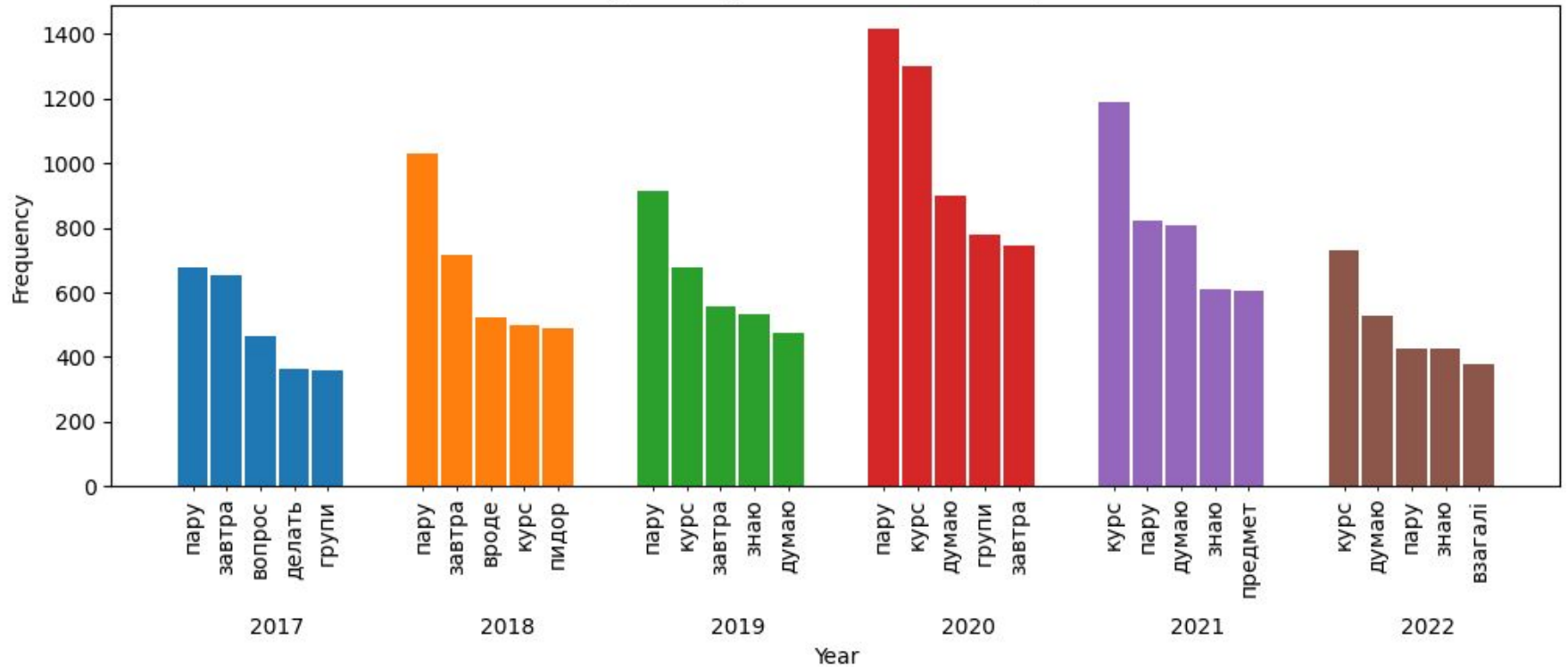


Period = per quarter (3 months)

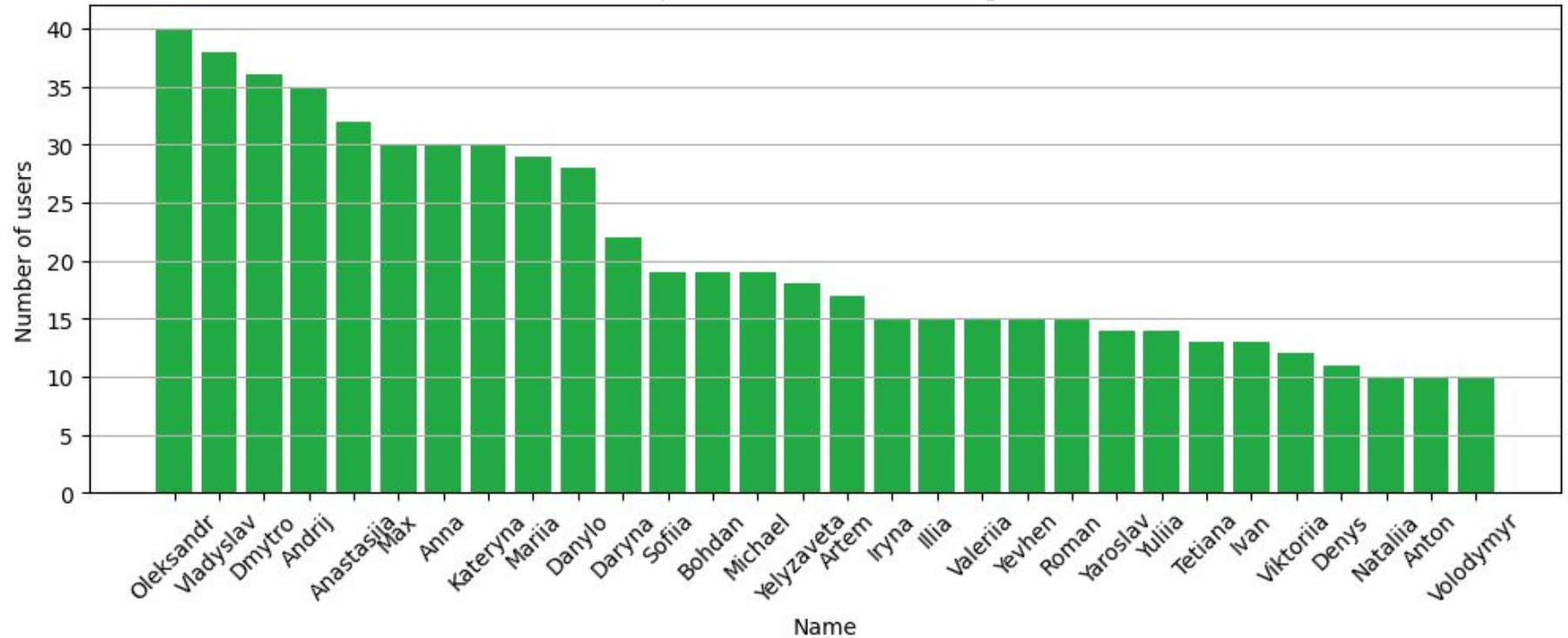
Message type usage trends (quarterly)



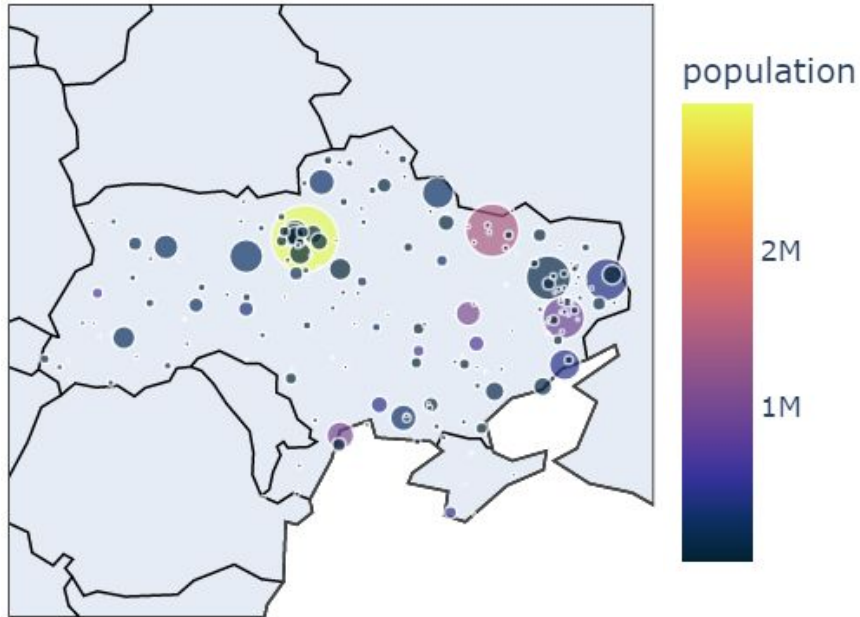
Yearly trending words in DMs or Group chats



30 Most Popular First Names of Telegram Users



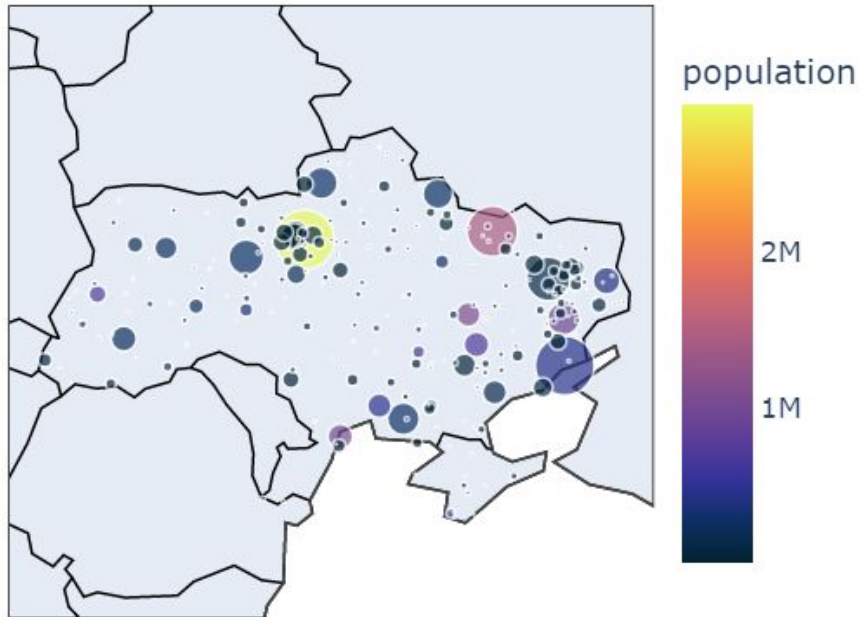
Most talked about Ukrainian cities during the full-scale invasion of Ukraine



February (02/2022)

- At the start the most mentioned cities were big cities, such as Kyiv, Kharkiv, Chernihiv and Sumy.

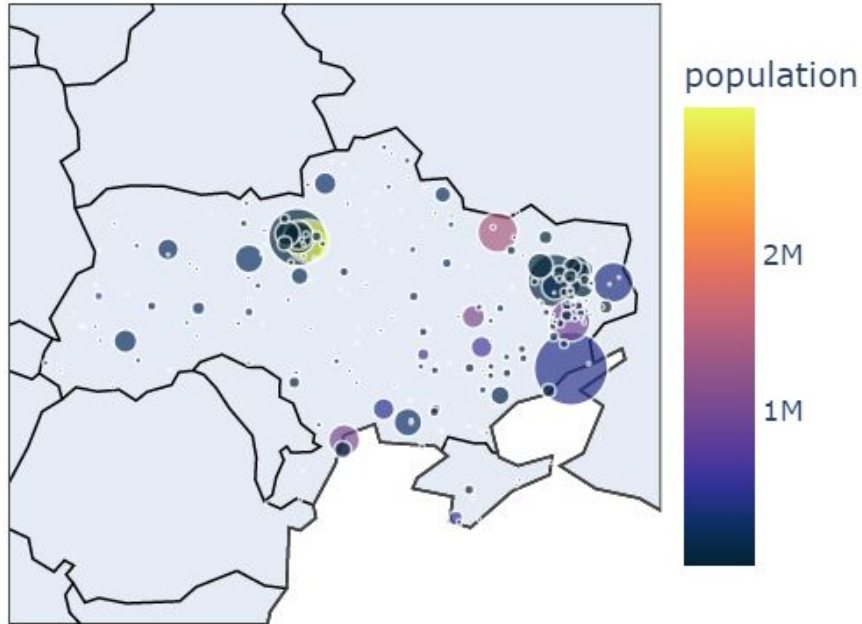
Most talked about Ukrainian cities during the full-scale invasion of Ukraine



March (03/2022)

- Mariupol becomes the main discussion topic.

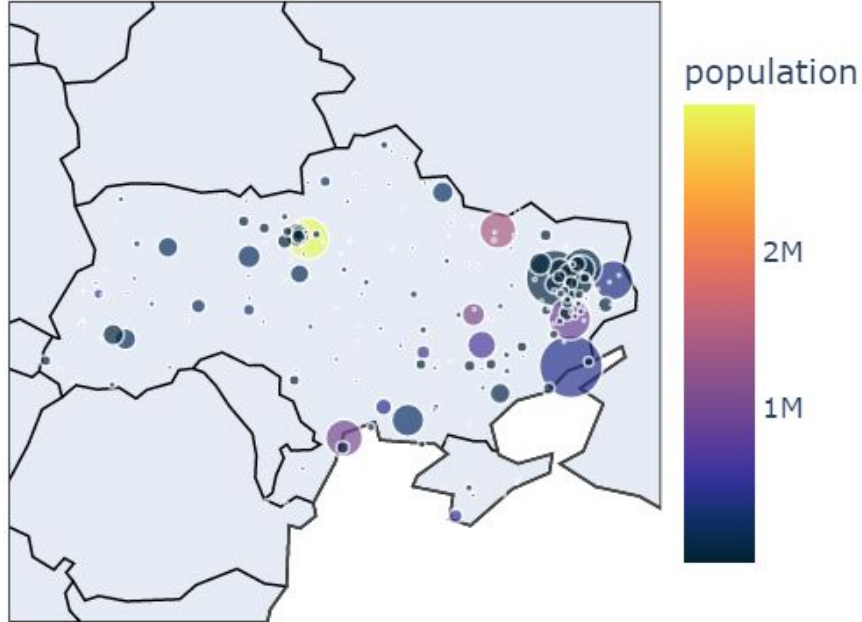
Most talked about Ukrainian cities during the full-scale invasion of Ukraine



April (04/2022)

- Small towns around Kyiv, such as Bucha, Irpin, and Hostomel, are now the main topic.
- Mariupol gains even more attention.

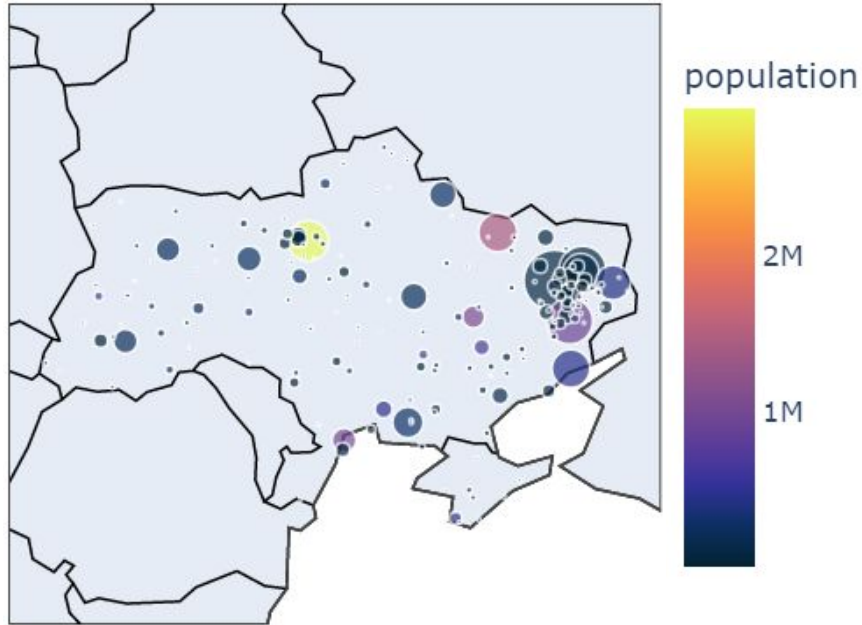
Most talked about Ukrainian cities during the full-scale invasion of Ukraine



May (05/2022)

- An increase in mentions of Eastern cities like Severodonetsk, Lysychansk, Bakhmut, and Slovyansk, as well as Odesa and Kherson at the South.

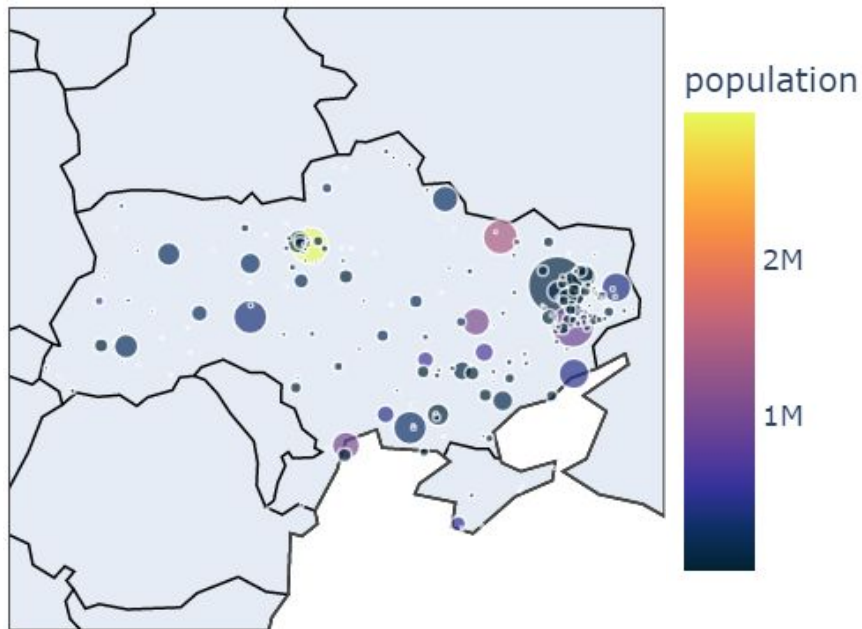
Most talked about Ukrainian cities during the full-scale invasion of Ukraine



June (06/2022)

- Even bigger emphasis on the East, but also Kremenchuk has a lot of mentions (Russian missile attack of a mall).

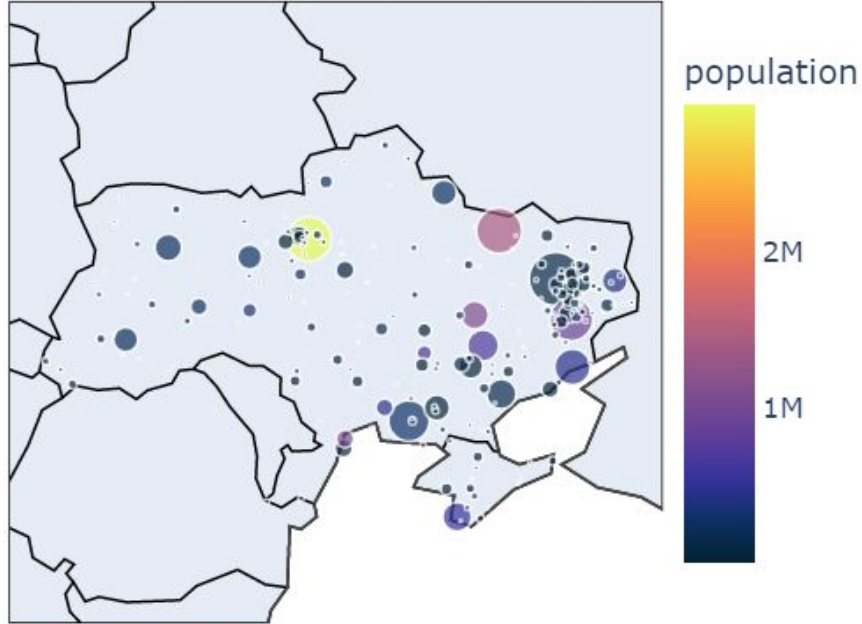
Most talked about Ukrainian cities during the full-scale invasion of Ukraine



July (07/2022)

- A big increase in the number of mentions of Vinnytsya, but also cities in Dnipro and Zaporizhzhya region.

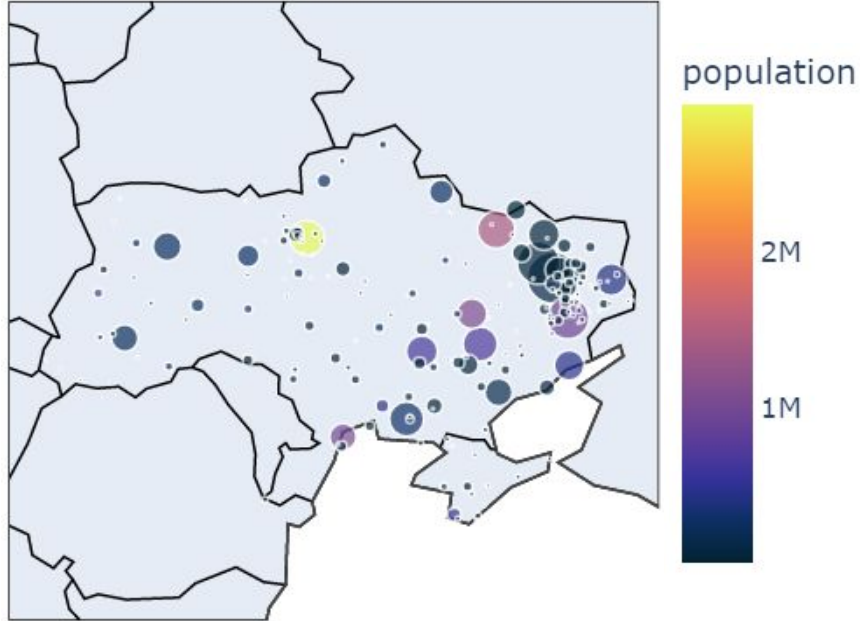
Most talked about Ukrainian cities during the full-scale invasion of Ukraine



August (08/2022)

- Even more talk about the Southern cities, but also worth noting that Sevastopol gained some attention as well.

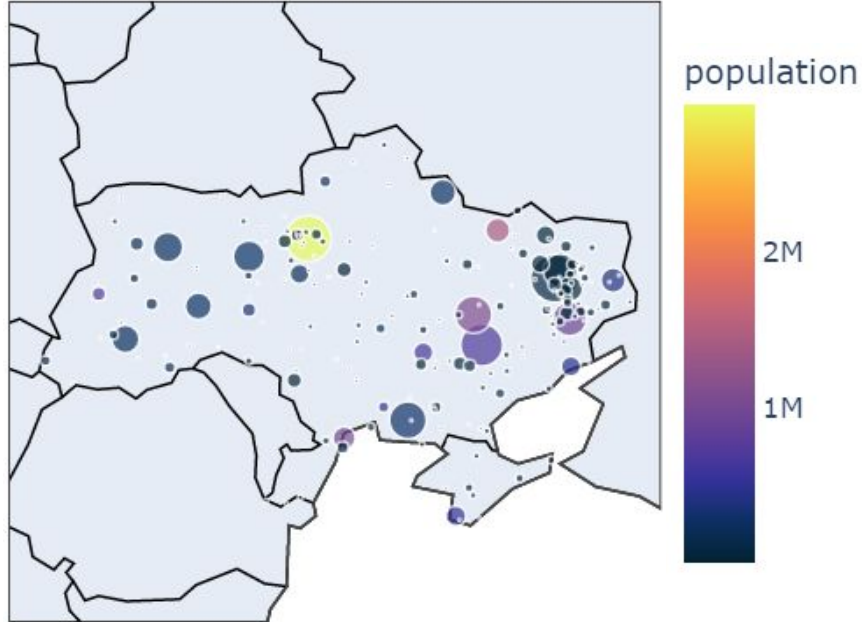
Most talked about Ukrainian cities during the full-scale invasion of Ukraine



September (09/2022)

- People started talking a lot about the Ukrainian counteroffensive in Kharkiv region.

Most talked about Ukrainian cities during the full-scale invasion of Ukraine

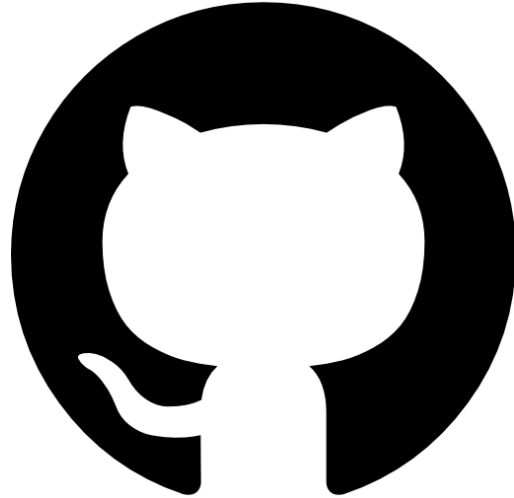


October (10/2022)

- Attention is being drawn towards big cities, most likely due to frequent attacks on Ukrainian electrical infrastructure.

Future work

- Yearly trending words have potential, but it still shows words that should probably be counted as stopwords. TF-IDF might be a good technique to use instead.
- Most common names would be interesting to compare to other places, and see how most common names differ, for instance on different social media.
- Most mentioned cities map can be improved by changing the way we detect the mentions of a city. For example, it can take into consideration capitalization, context, and grammar.



github.com/74R45/telegram-eda



**Thank you for your
attention!**

