

# Telegram data-analysis of personal data

Student: Stoliaruk Yuliia  
Teacher: Andrew Kurochkin

Course: Computational Social Science  
Date: 11.12.2023

# Plan of the presentation

1. Introduction
2. How to get data
3. Data overview
4. Exploratory Data Analysis
5. Final results
6. Further work
7. Git Repository

# Introduction

The main purpose of this project is to

- explore messenger data
- investigated Telegram behavioural patterns
- find as much insight as possible

# How to get data

To get personal data from Telegram I used two repositories, that were provided:

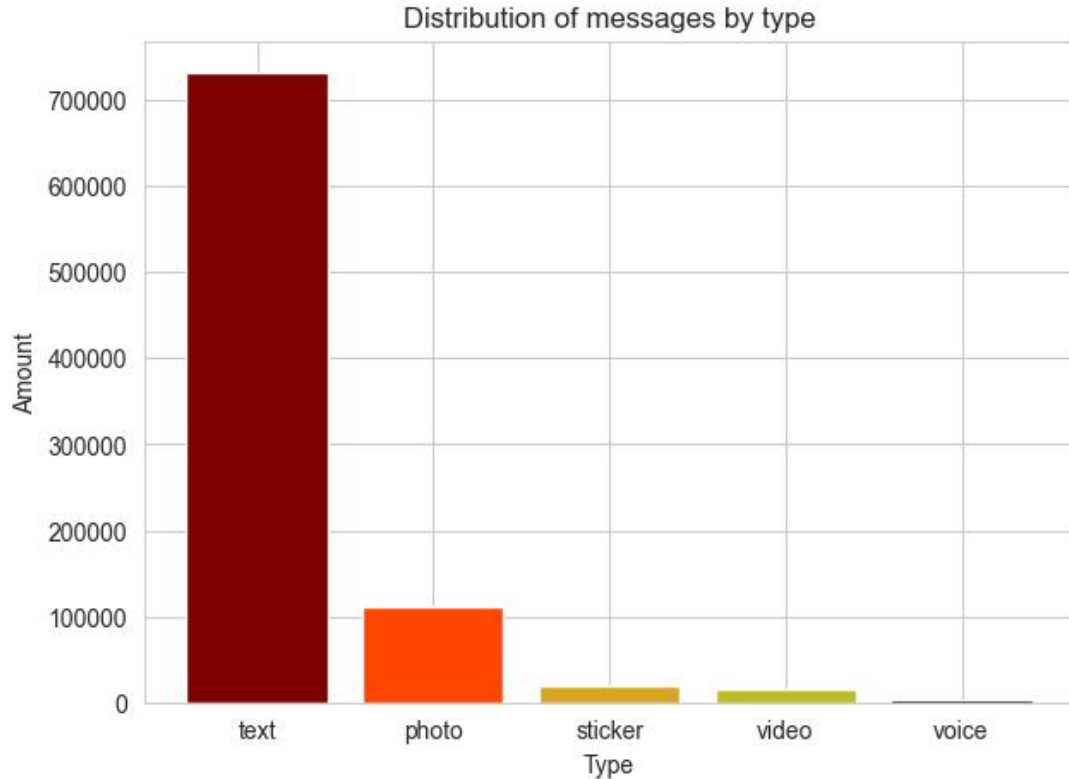
1. To download raw data from Telegram use this repository [SanGreel/telegram-data-collection \(github.com\)](https://github.com/SanGreel/telegram-data-collection)

Some chats could have too many messages therefore it's better to limit maximum amount of messages to 100 000 for one chat.

2. Then merge collected data into csv datasets using this repository [SanGreel/telegram-dialogs-analysis-v2 \(github.com\)](https://github.com/SanGreel/telegram-dialogs-analysis-v2)

The main complexity of this task was time spent to download raw data.

# Data overview (Messages data)

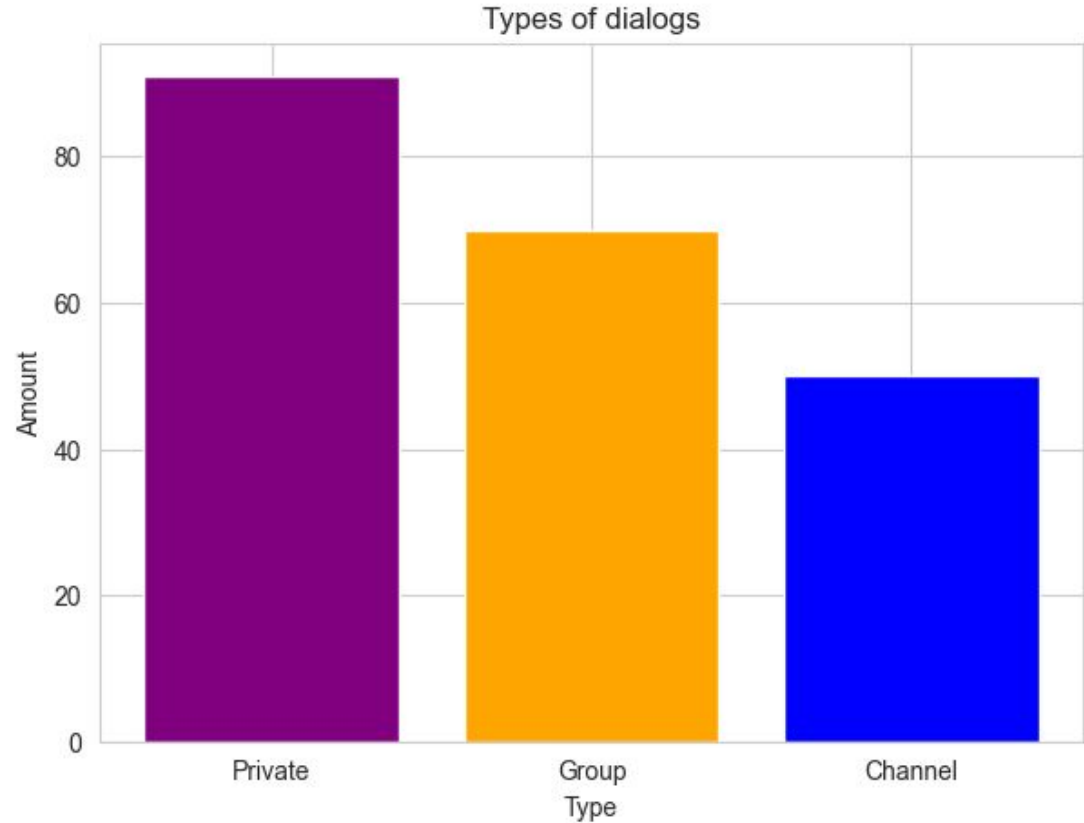


Size: 206.634 MB

Messages number: 880 940

# Data overview (Dialogs data)

Size: 1.85 MB

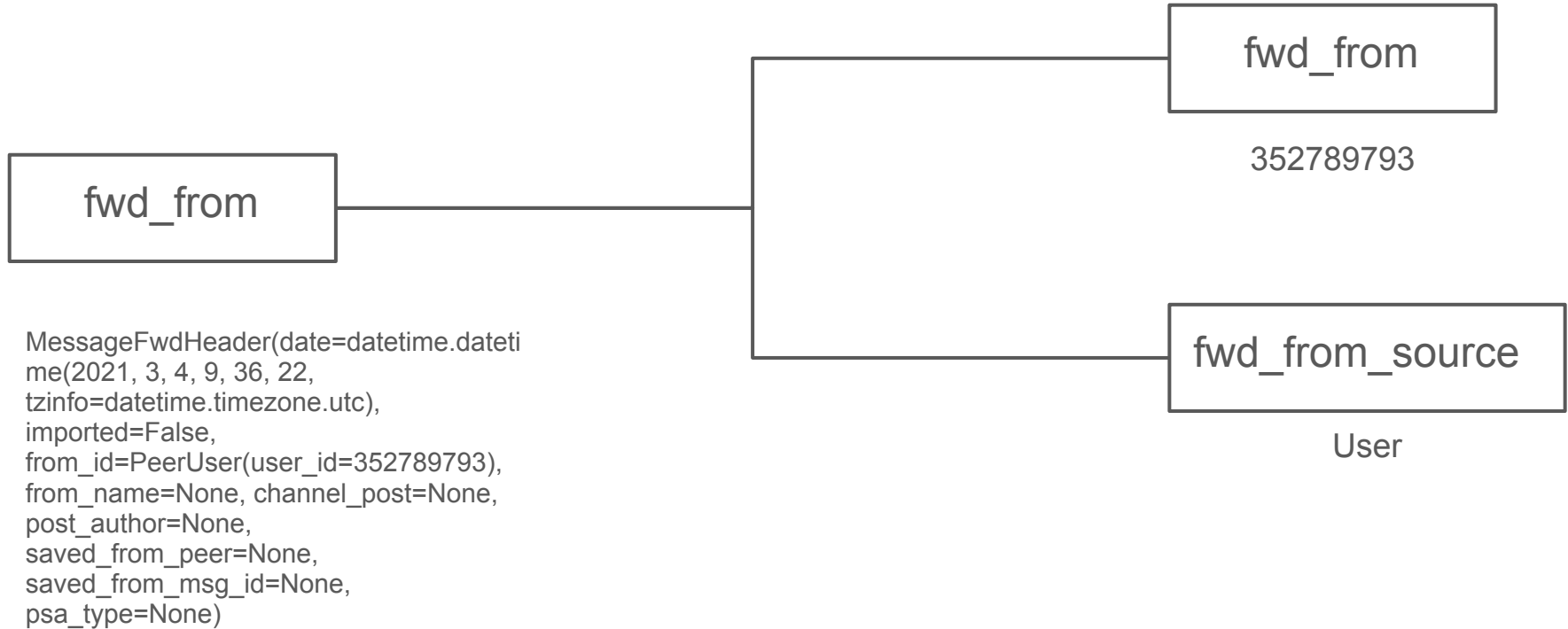


# Exploratory Data Analysis (Data cleaning)

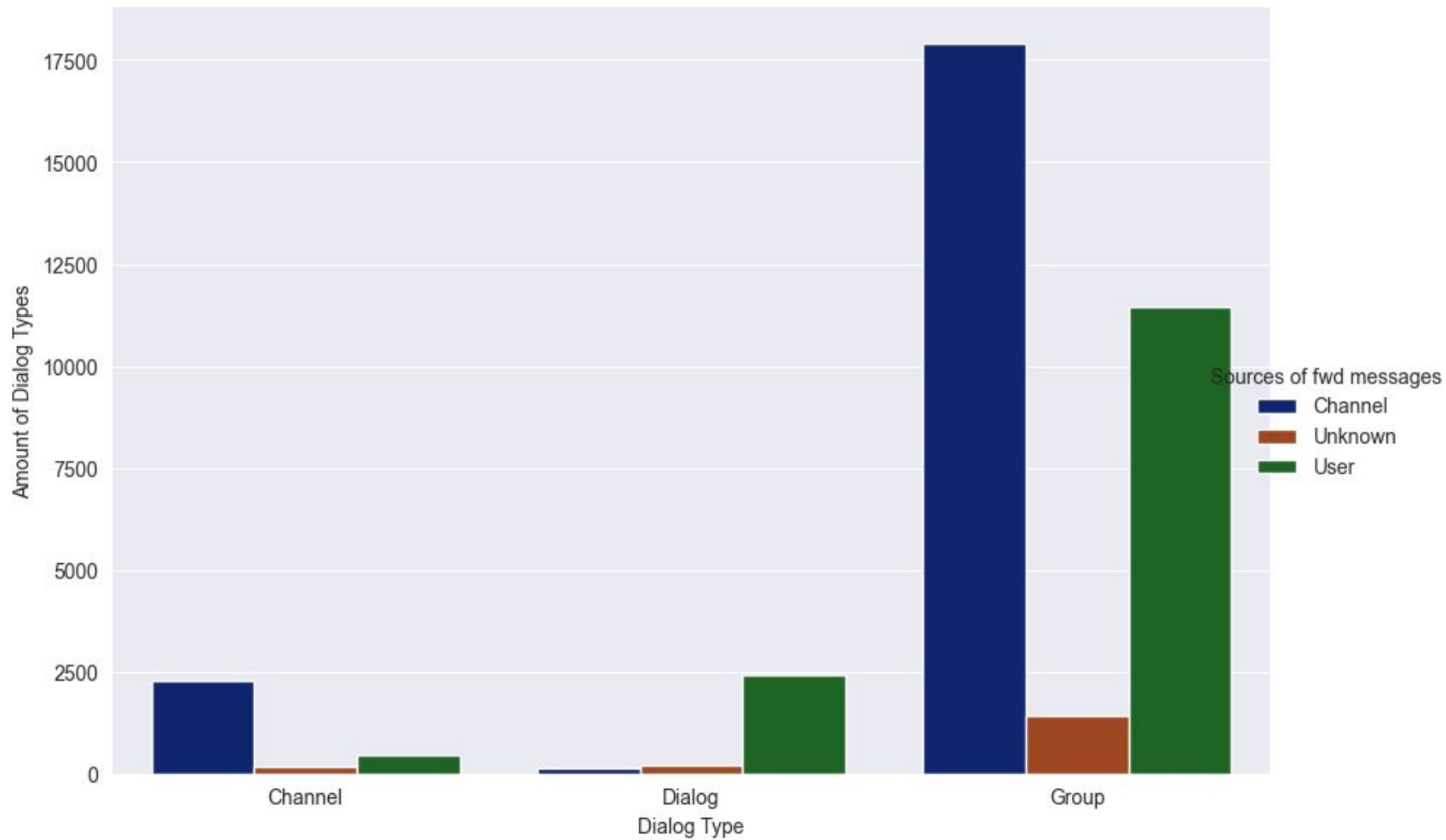
The first step was to clear the data.

1. Cleaning 'fwd\_from' column
2. Cleaning 'from\_id' column
3. Cleaning 'message' column

# Cleaning 'fwd\_from' column







## Cleaning 'from\_id' column

To use column 'from\_id' in further analysis i cleaned it from unnecessary data.

`'PeerUser(user_id=475253228)'`



`'475253228'`

# Cleaning 'message' column

To delete unnecessary data from text of messages I used this steps:

- Removed small words, punctuation, digits, links
- Removed stop-words

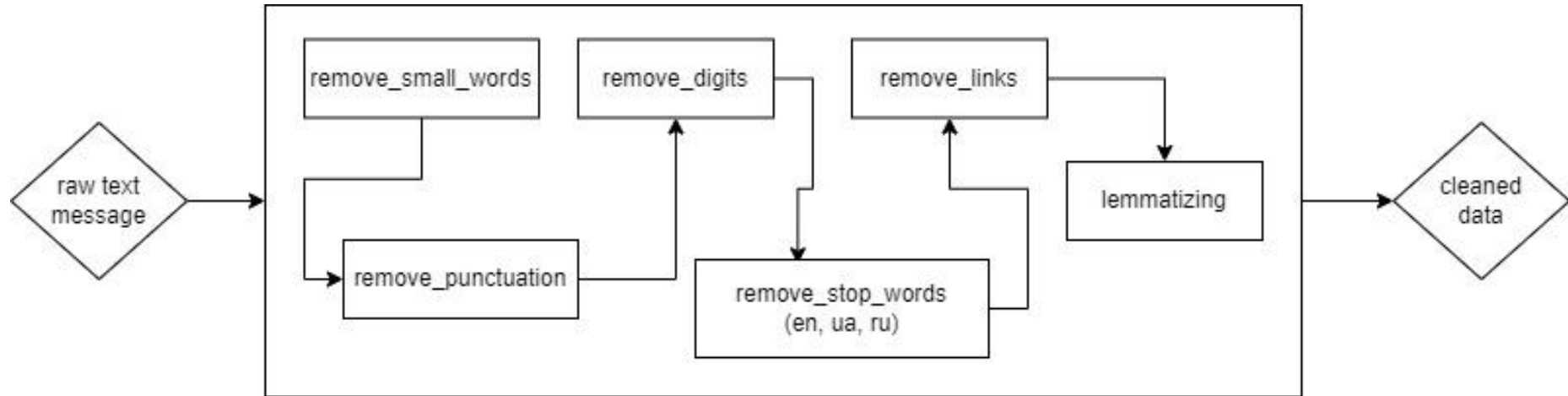
To remove stop-words I used dictionary of english and russian words from “nltk” library.

For ukrainian words I used dictionary from [this](#) repo made by Serhii Kupriienko.

- Words lemmatizing

To remove inflectional endings only and to return the base of a word, which is known as the lemma I used Wordnet Lemmatizer.

# Cleaning 'message' column



# Final results

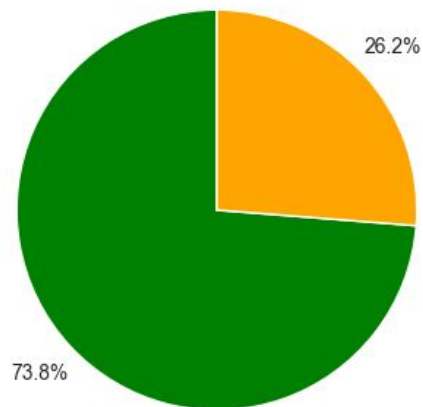
- General analysis
- Private dialog analysis
- Groups analysis
- Channels analysis

# General analysis of my messages

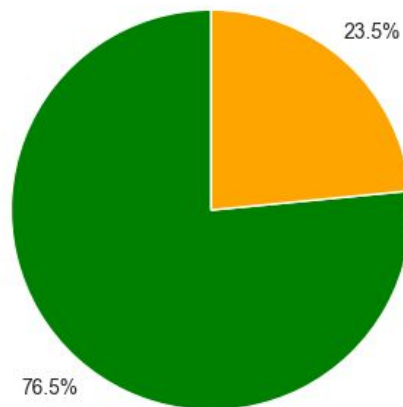
# General analysis of my messages

Language distribution

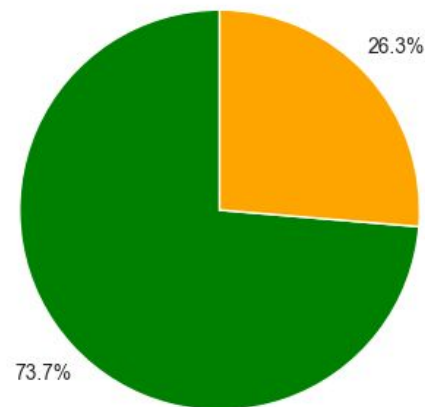
All Messages



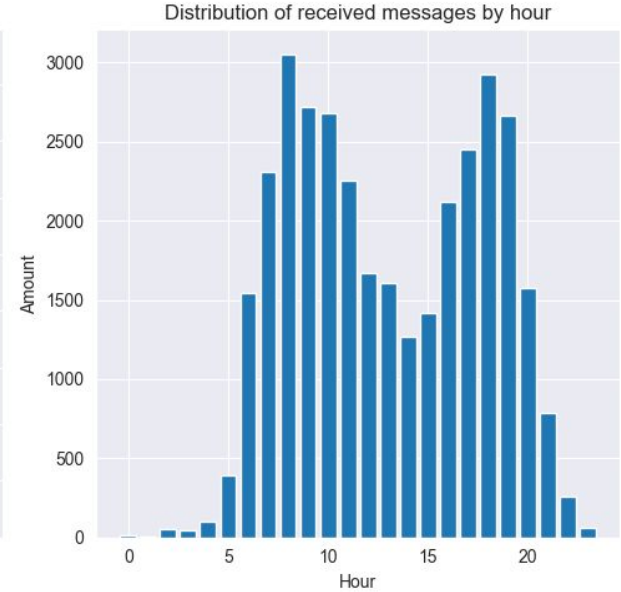
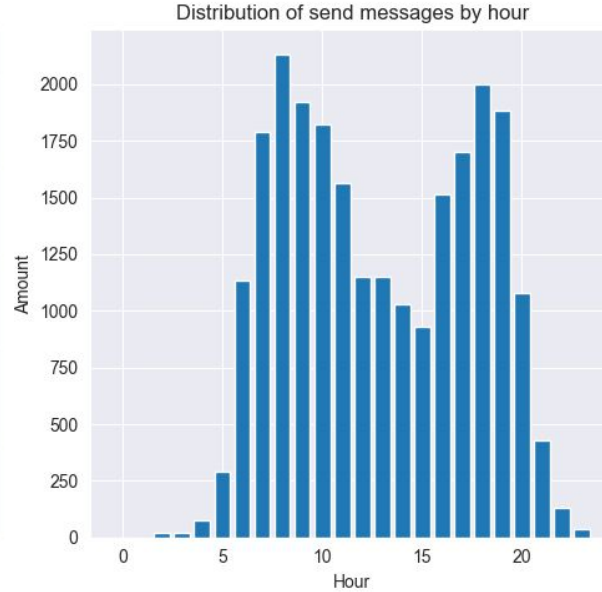
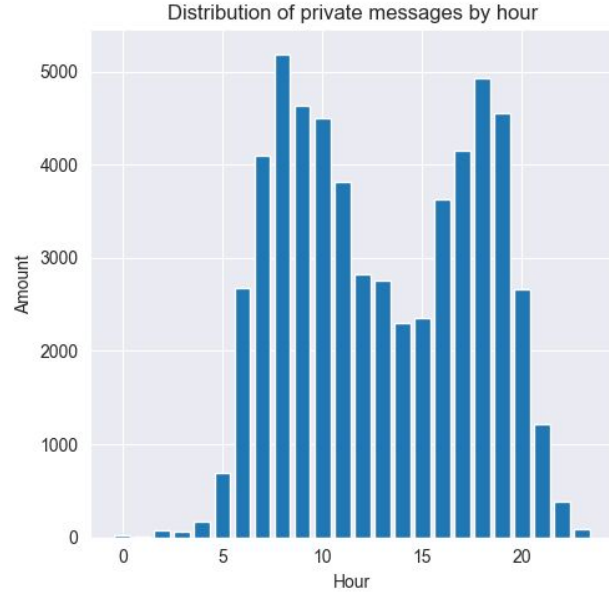
Sent Messages



Received Messages



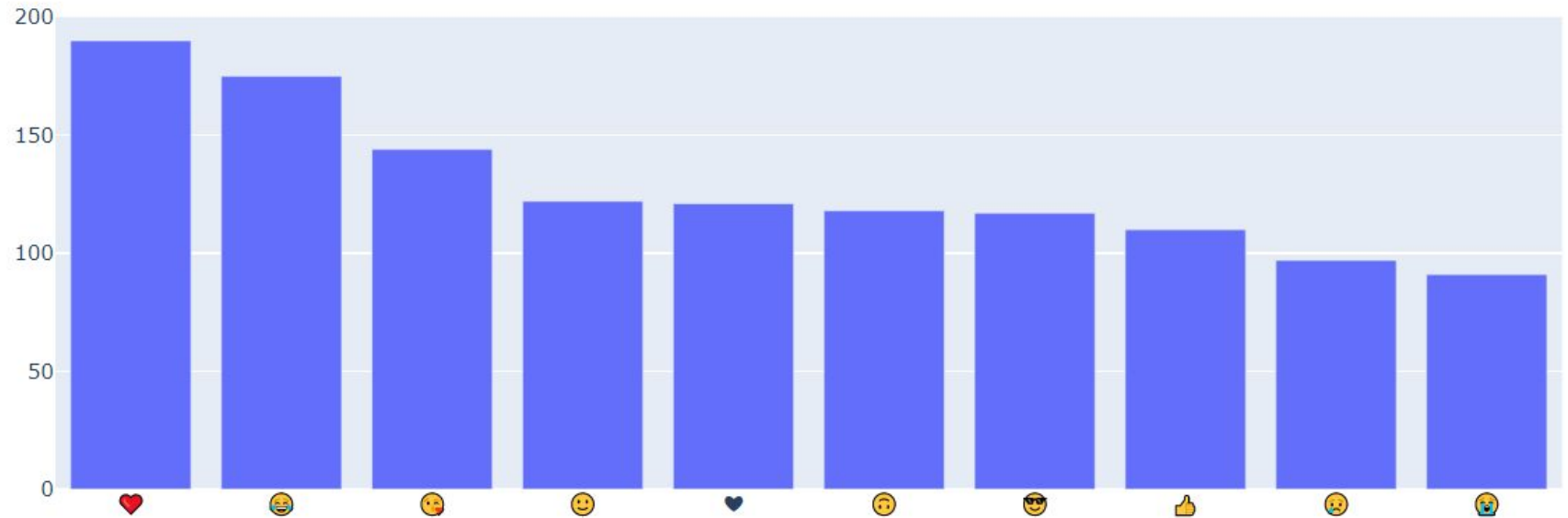
# General analysis of my messages





# General analysis of my messages

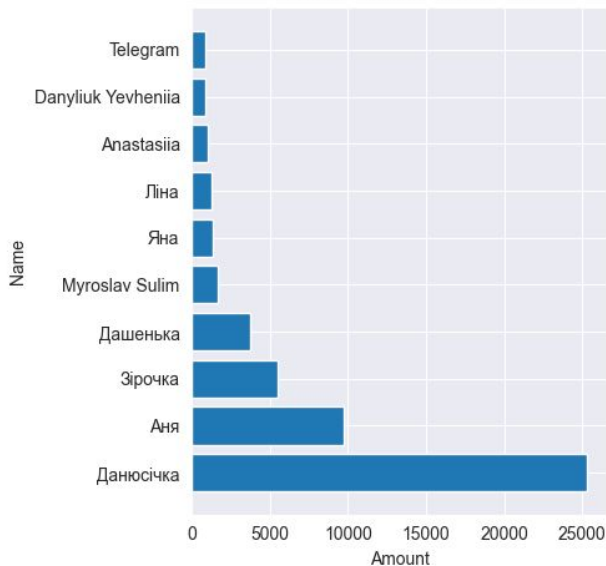
My top-10 emojis



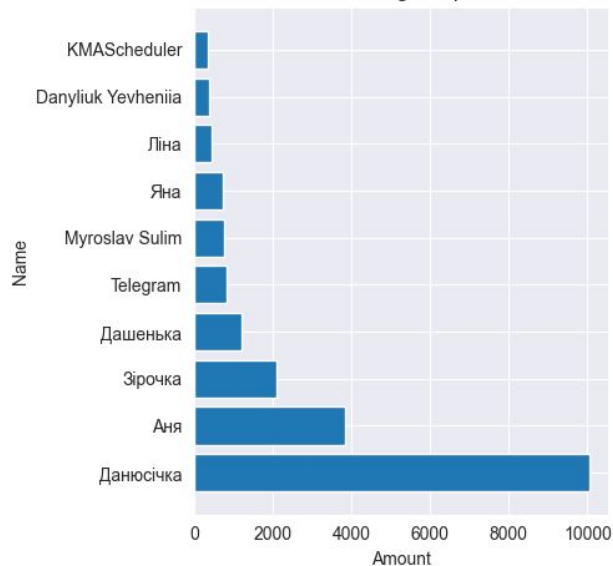
# Private dialogs analysis

# Private dialogs analysis

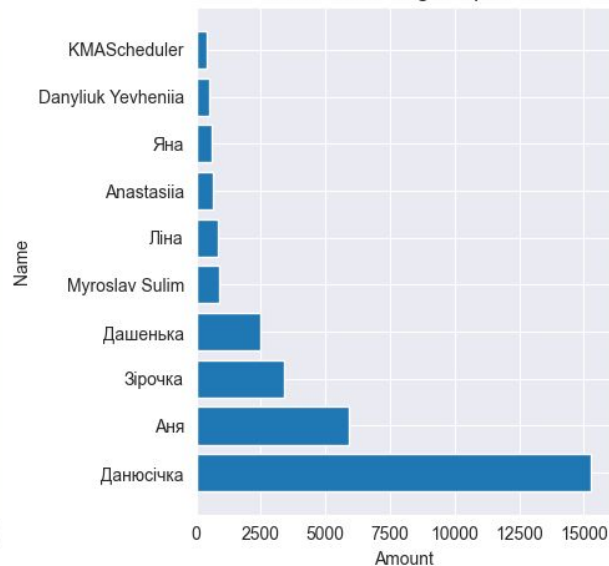
Users with most interections



Send messages top users

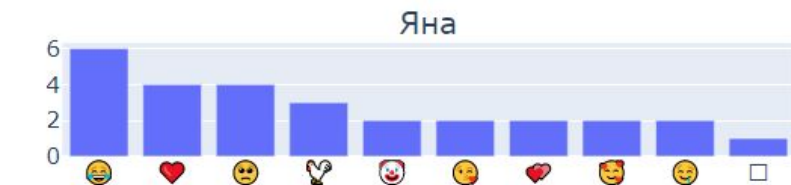
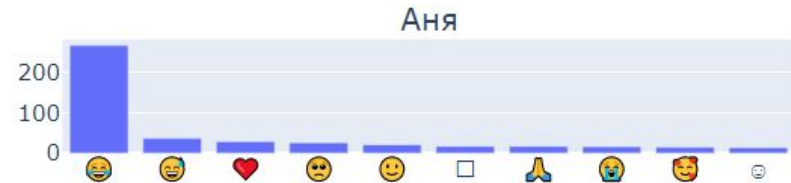
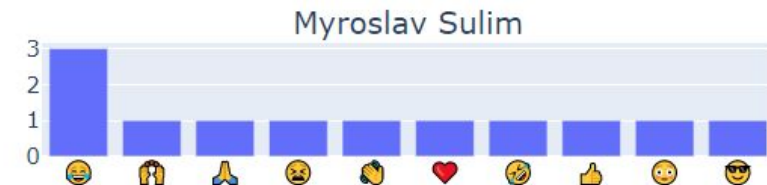
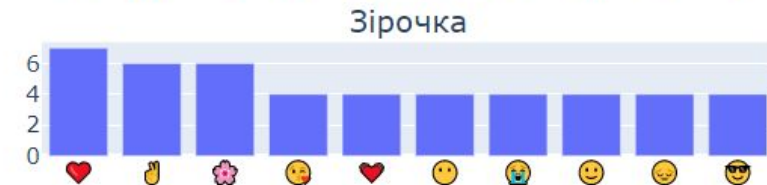
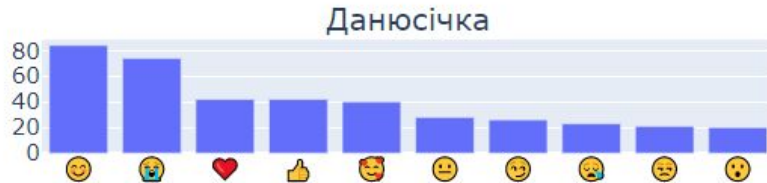


Received messages top users



# Top emojis of users with whom I interact the most

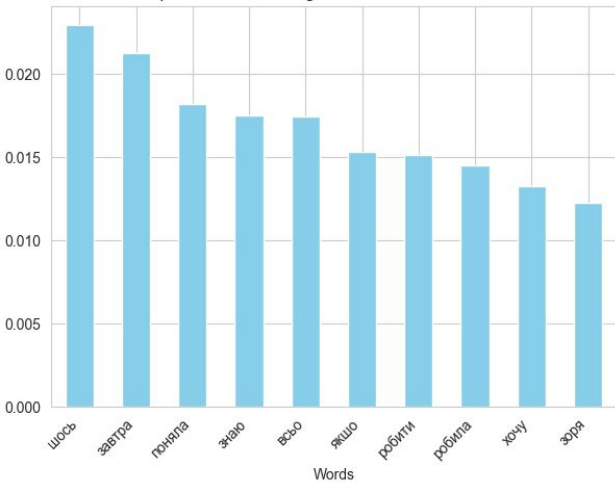
## Top-10 Emojis in Received Messages



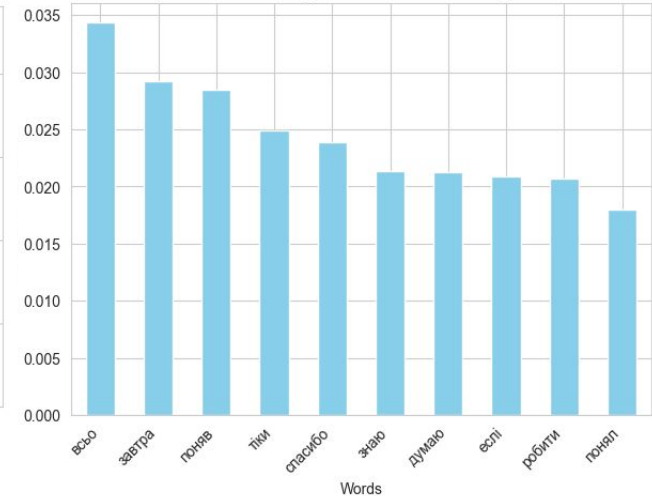
# TF-IDF (Term Frequency Inverse Document Frequency) of dialogs

To calculate TF-IDF of words, I splitted dialogs into documents by dates and calculated TF-IDF of words using library 'sklearn'.

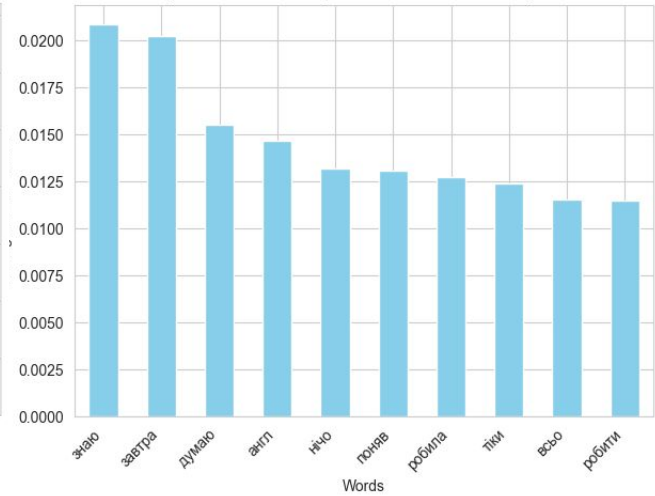
Top 10 Words with Highest TF-IDF Scores for Аня



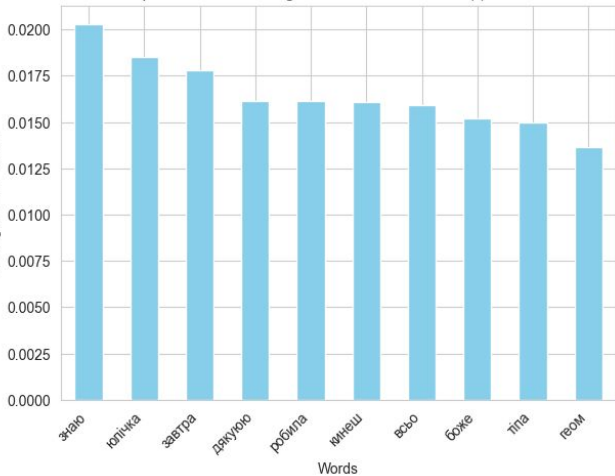
Top 10 Words with Highest TF-IDF Scores for Данюсічка



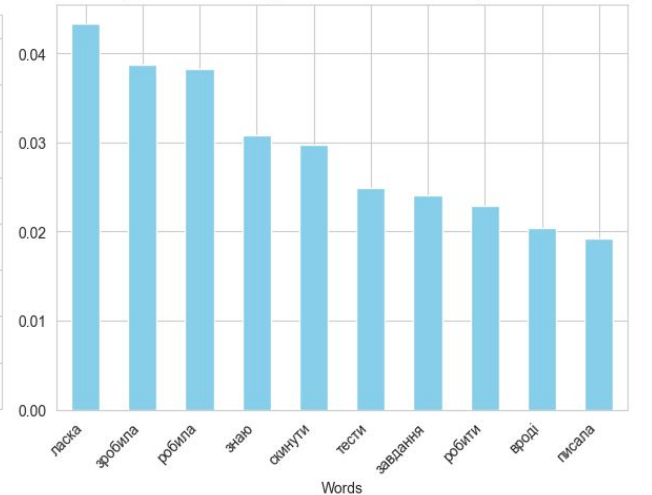
Top 10 Words with Highest TF-IDF Scores for Зірочка



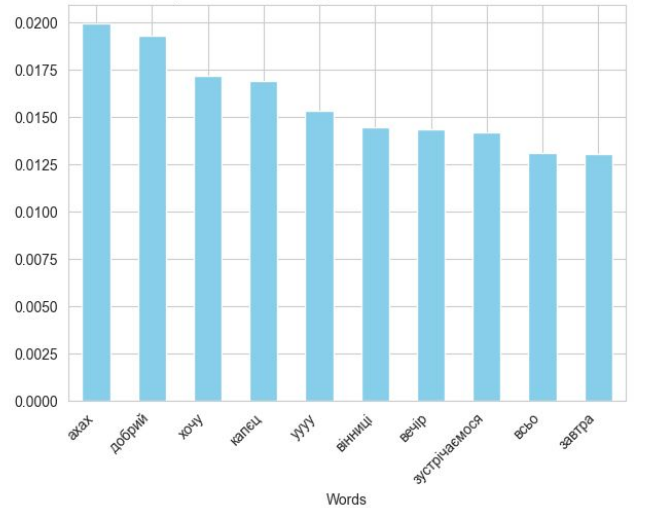
Top 10 Words with Highest TF-IDF Scores for Дашенька



Top 10 Words with Highest TF-IDF Scores for Myroslav Sulim



Top 10 Words with Highest TF-IDF Scores for Яна



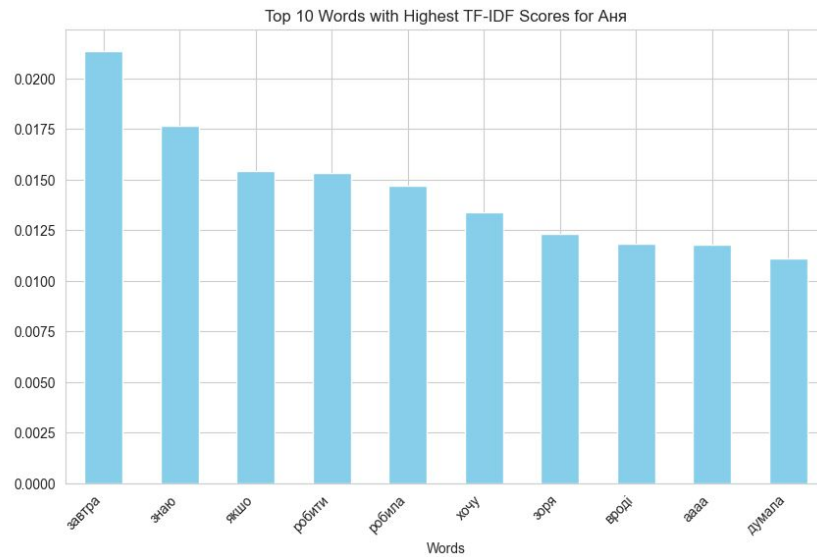
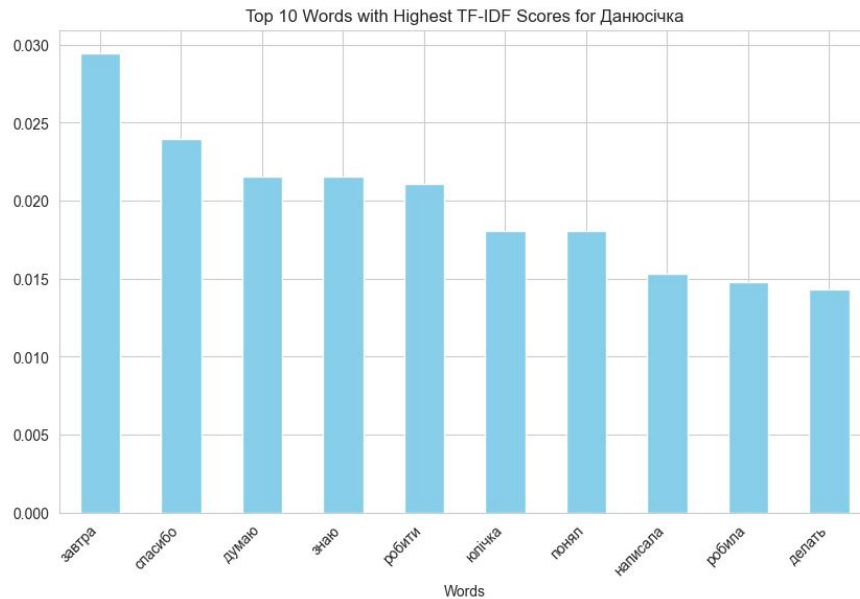
## TF-IDF (Data cleaning problem)

Histograms of words with top tf-idf show us that there are still a lot of words that don't give meaningful information about dialogs.

This happened because in most of my private dialogs consists of informal vocabulary.

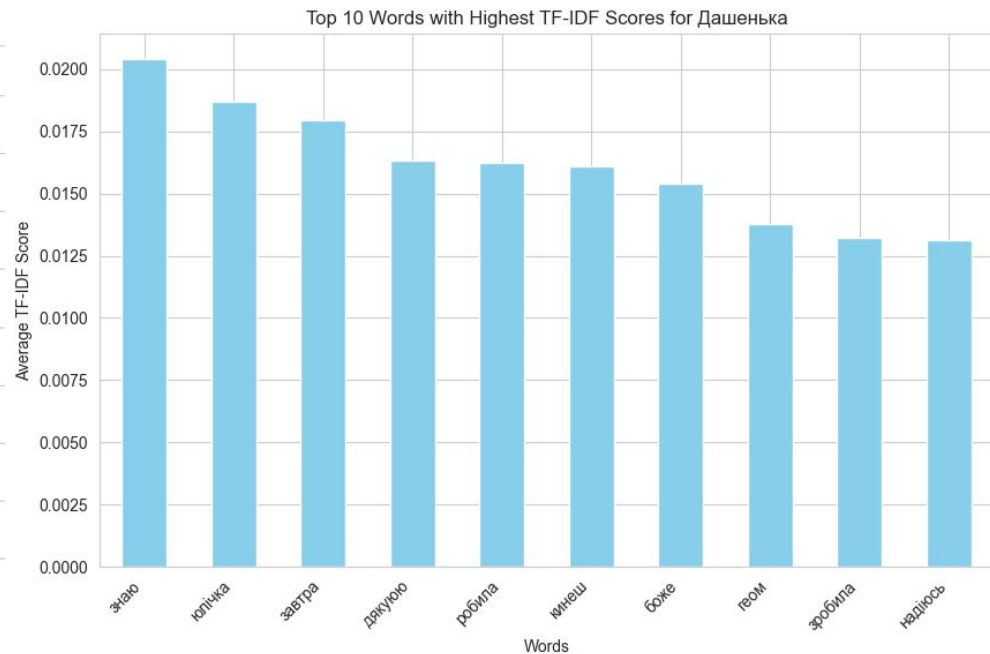
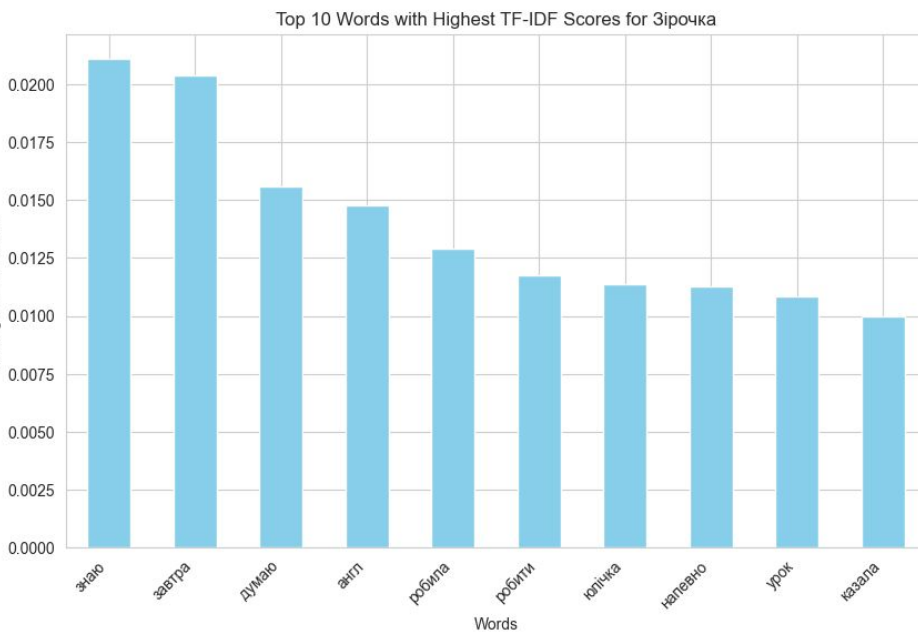
To fix this problem I created new additional dictionary of stop-words.

# Updated tf-idf

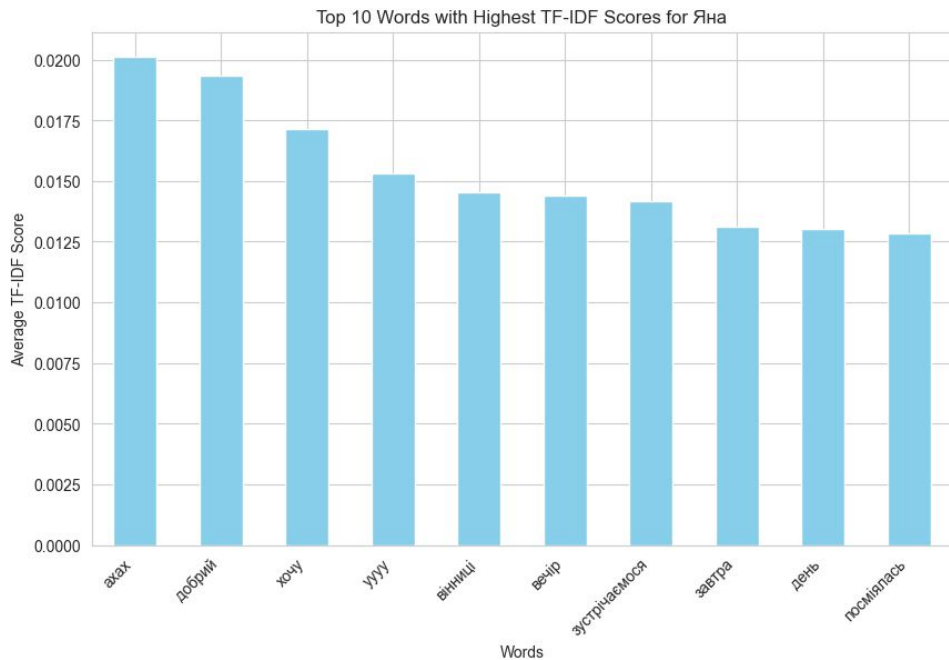
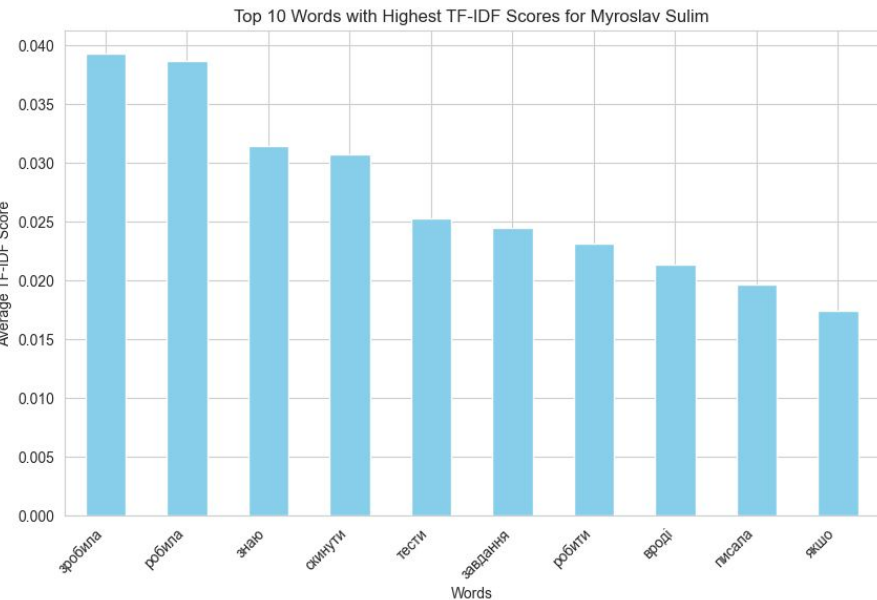




# Updated tf-idf

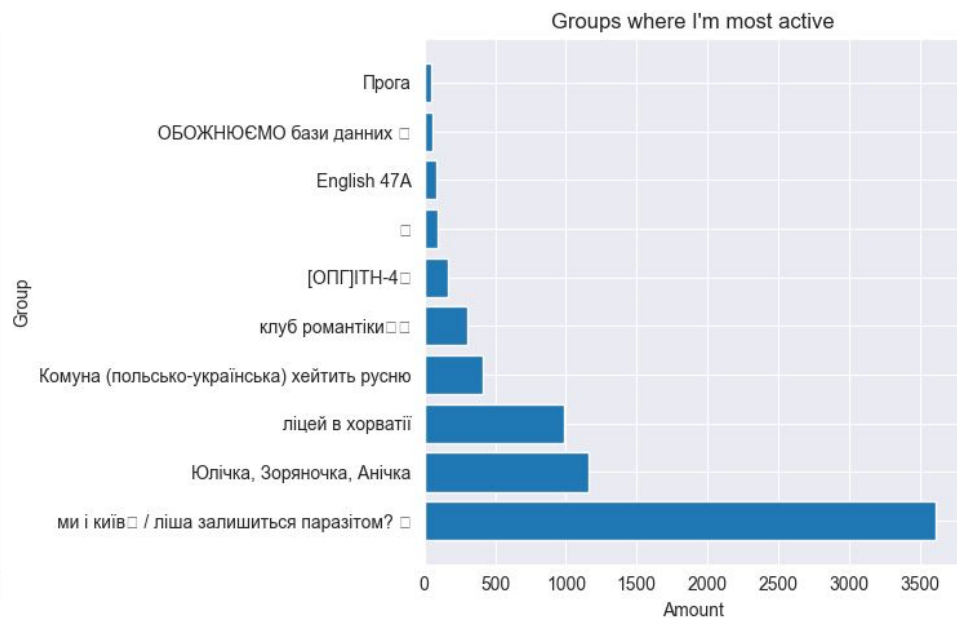
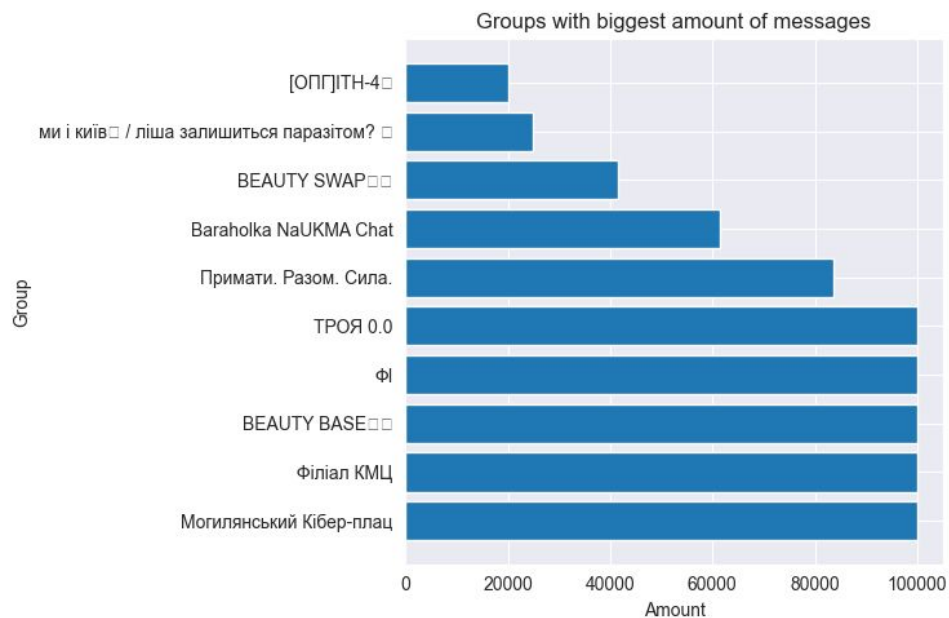


# Updated tf-idf



# Groups analysis

# Groups analysis



# Analysis of Могилянський Кібер-плац

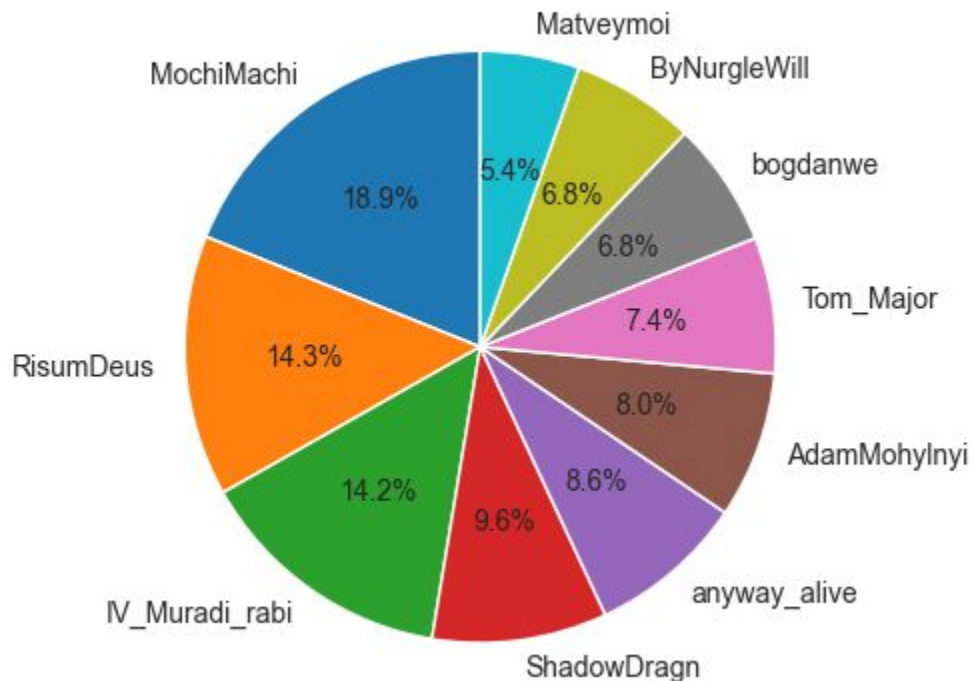
General:

Amount of messages: 100 000

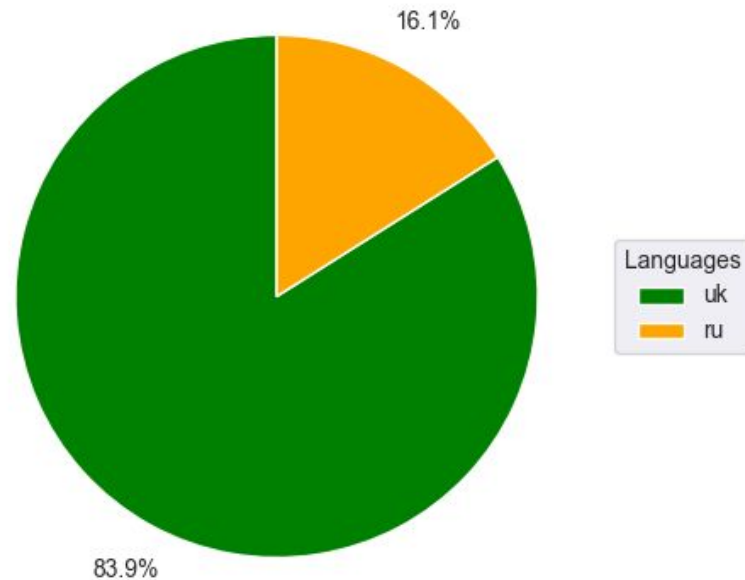
Amount of users: 1613

# Analysis of Могилянський Кібер-плац

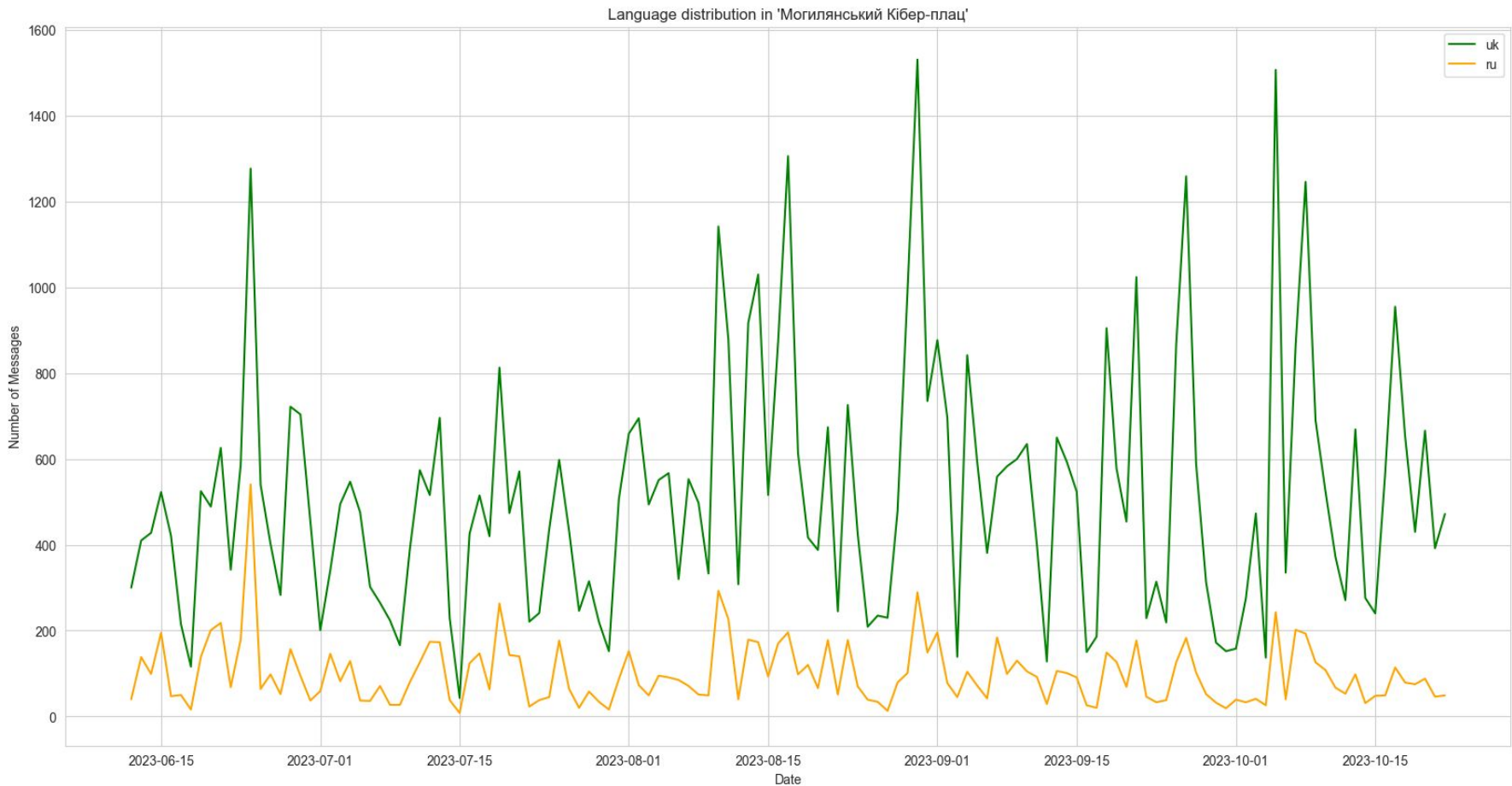
Most active users in 'Могилянський Кібер-плац'



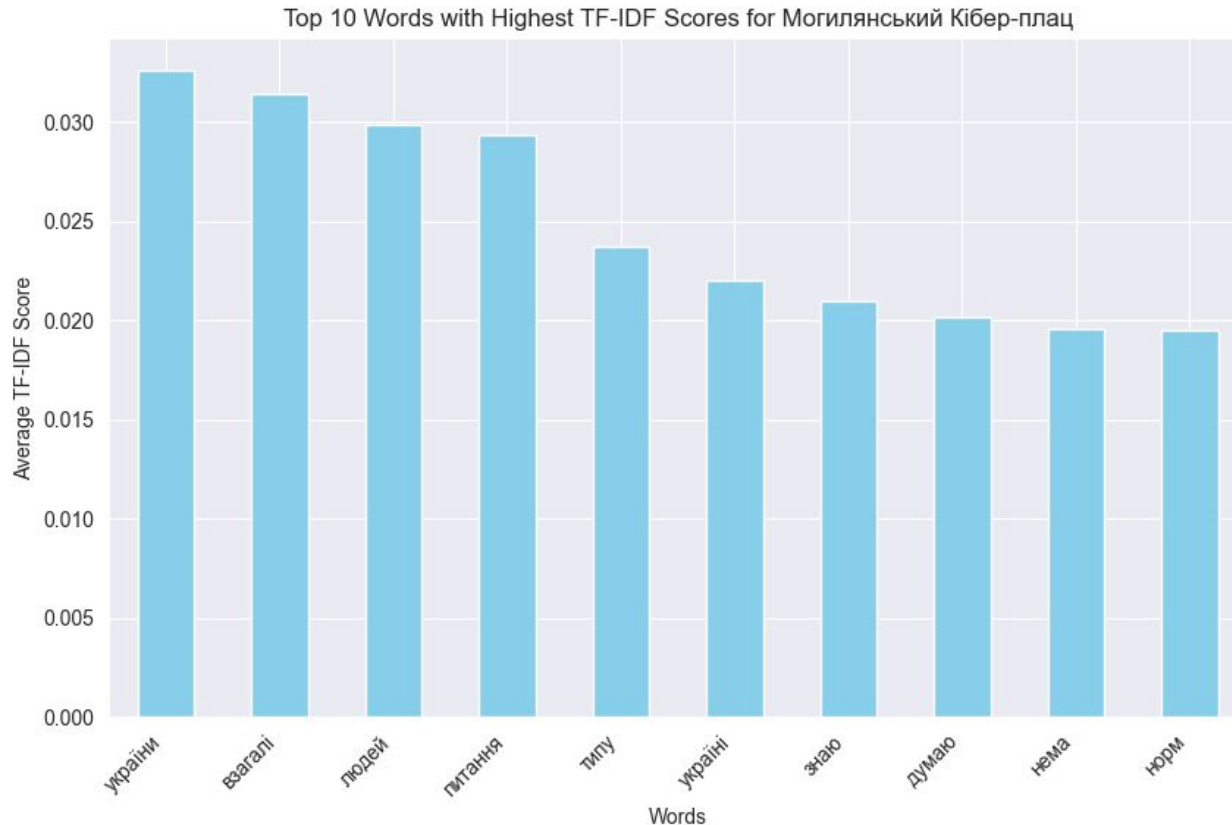
Language distribution in 'Могилянський Кібер-плац'



# Analysis of Могилянський Кібер-плац



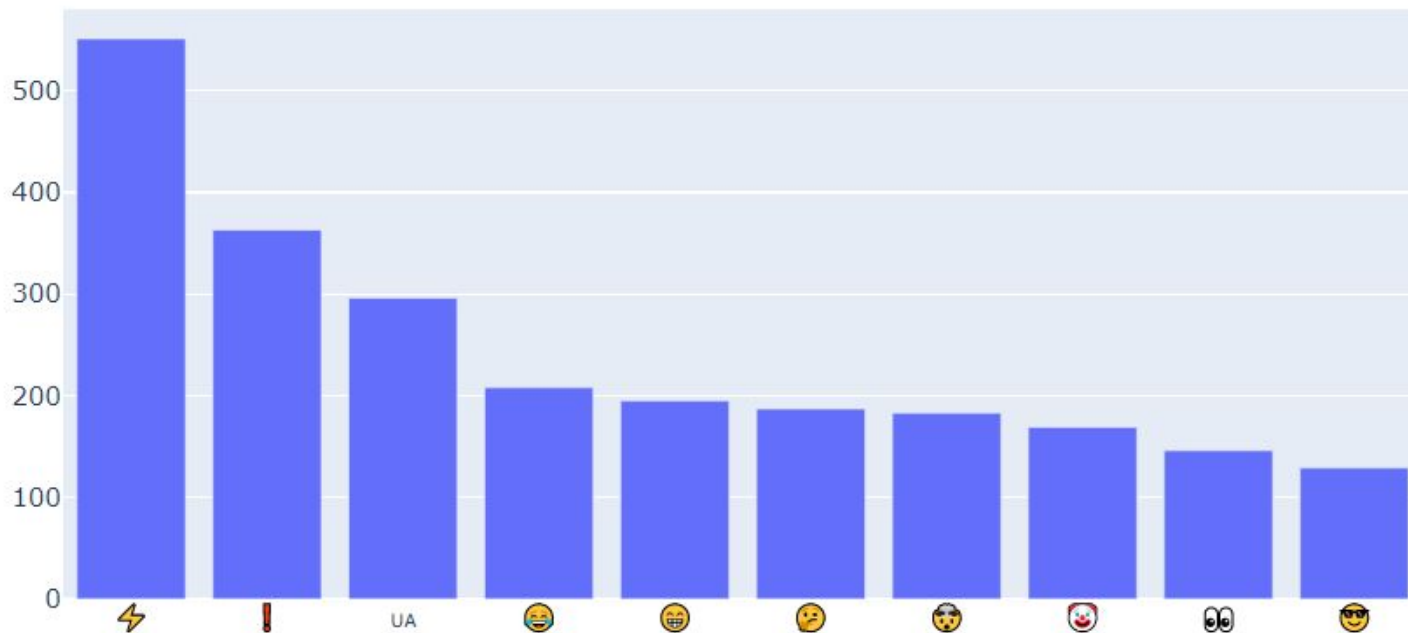
# Analysis of Могилянський Кібер-плац



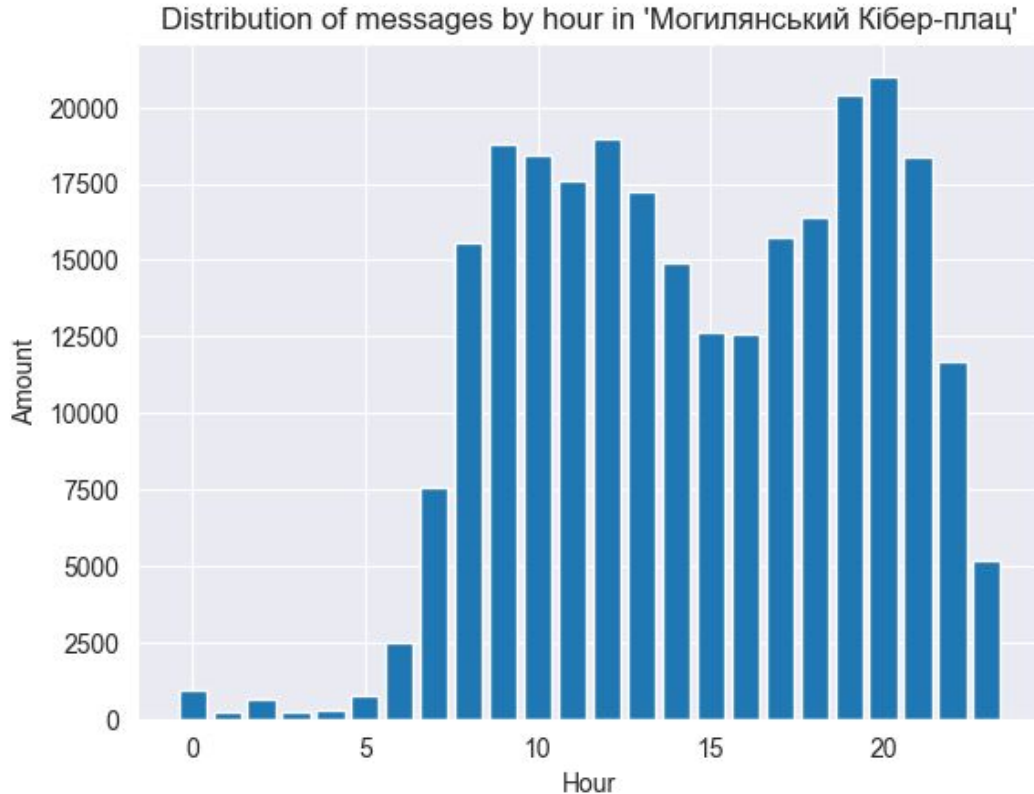


# Analysis of Могилянський Кібер-плац

Top-10 emojis of 'Могилянський Кібер-плац'



# Analysis of Могилянський Кібер-плац



# Analysis of BEAUTY BASE

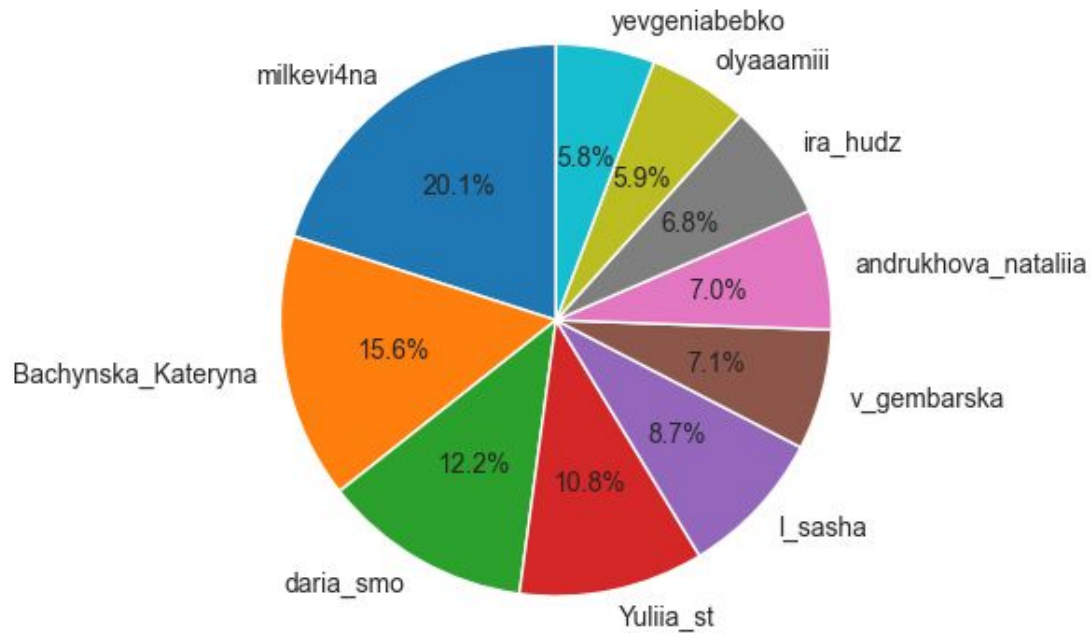
General:

Amount of messages: 100 000

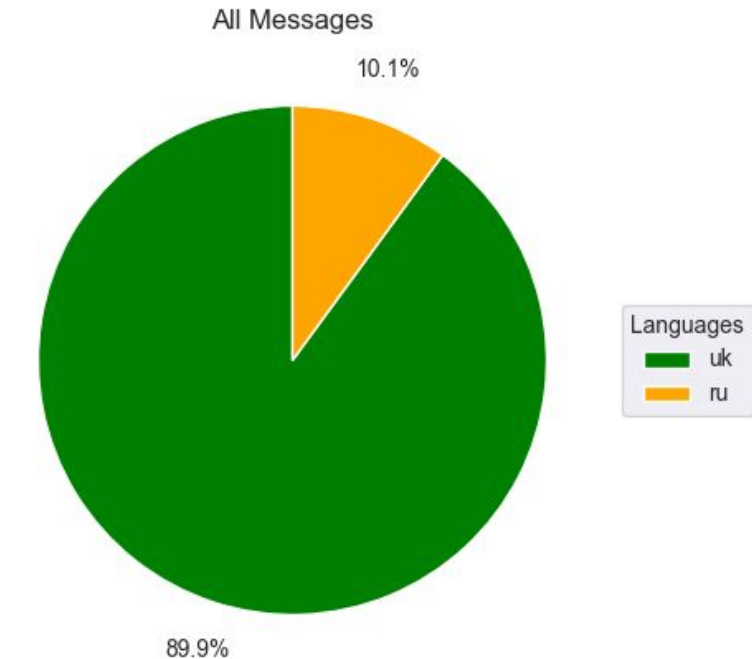
Amount of users: 1808

# Analysis of BEAUTY BASE

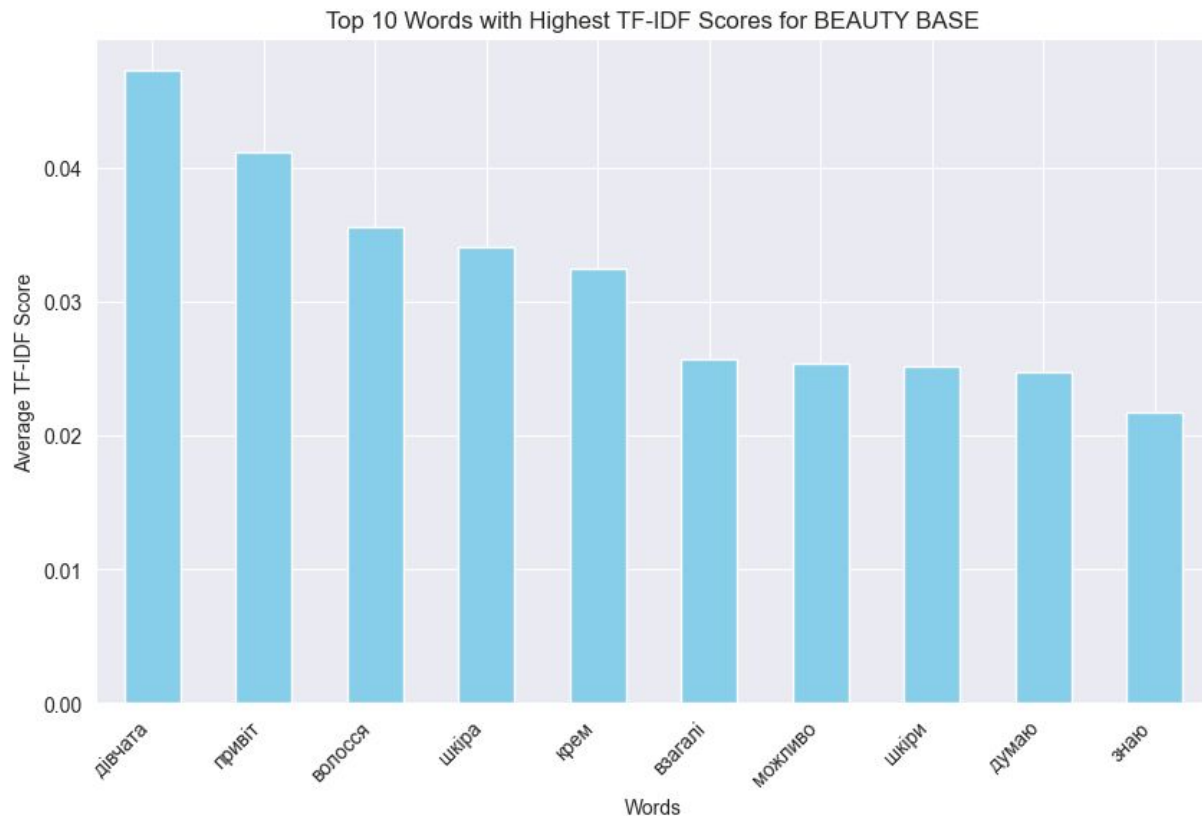
Most active users in 'BEAUTY BASE'



Language distribution in 'BEAUTY BASE'



# Analysis of BEAUTY BASE



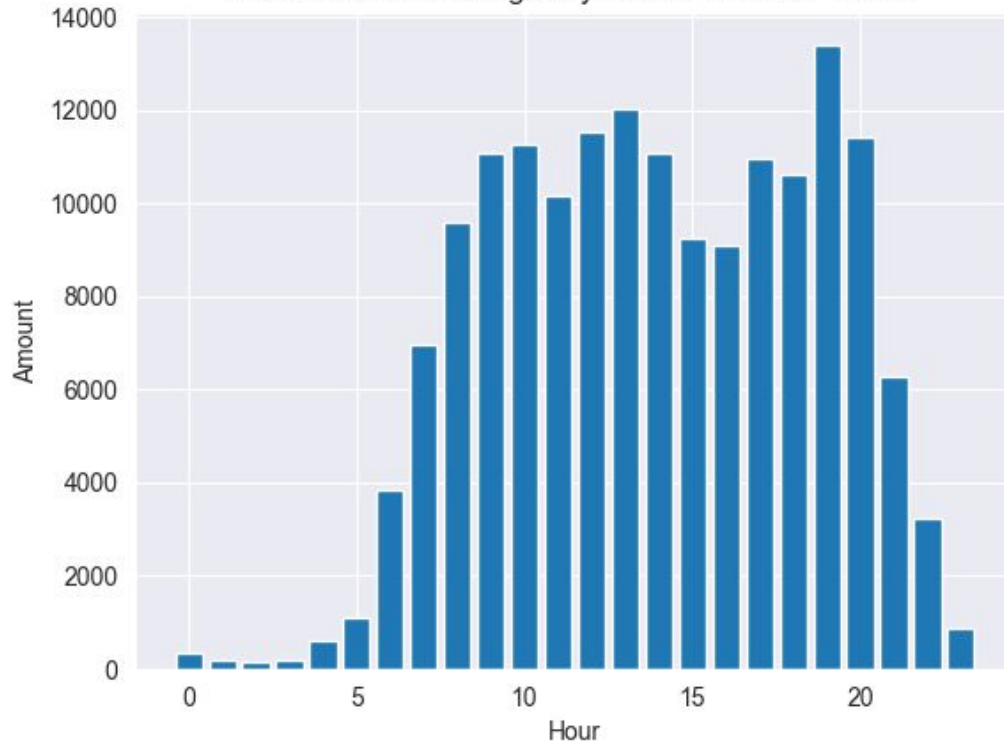
# Analysis of BEAUTY BASE 🧴 📝

Top-10 emojis of 'BEAUTY BASE'



# Analysis of BEAUTY BASE

Distribution of messages by hour in 'BEAUTY BASE'

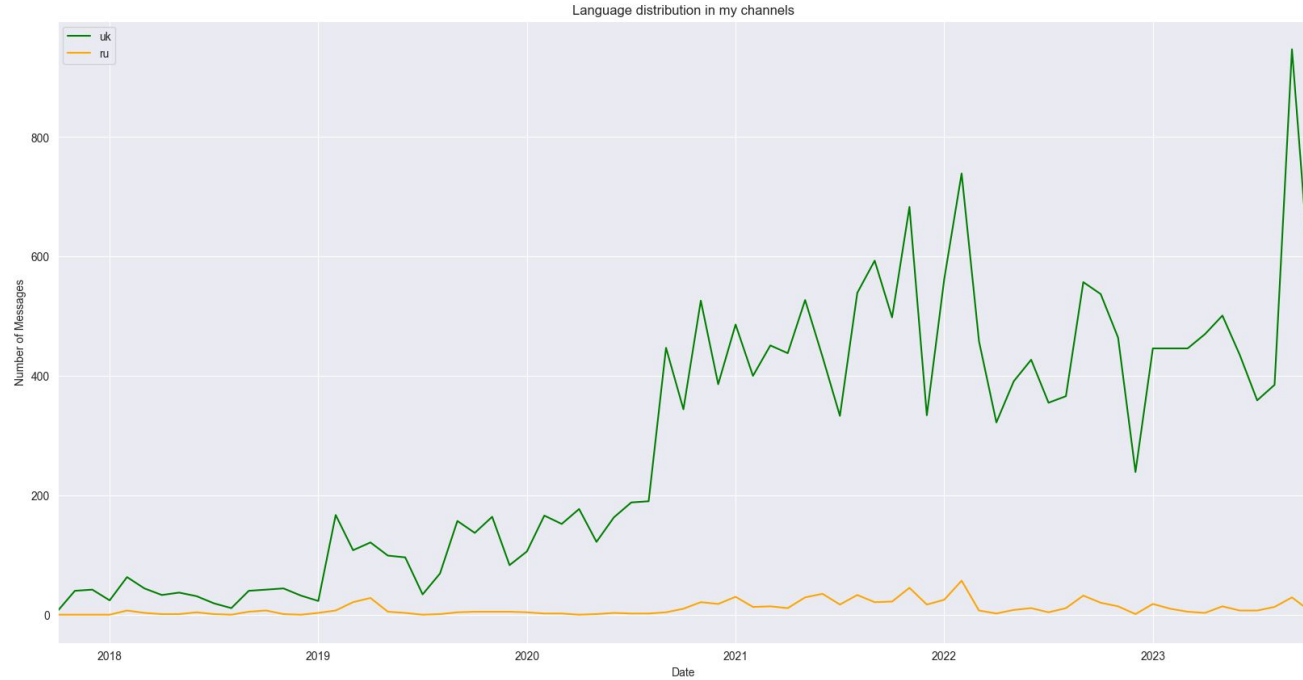
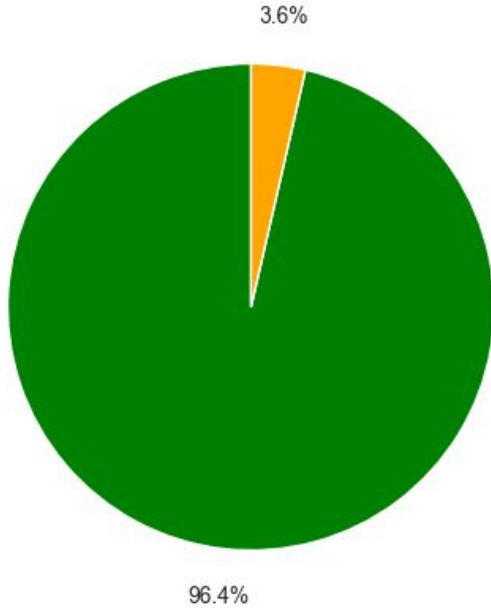


# Analysis of channels

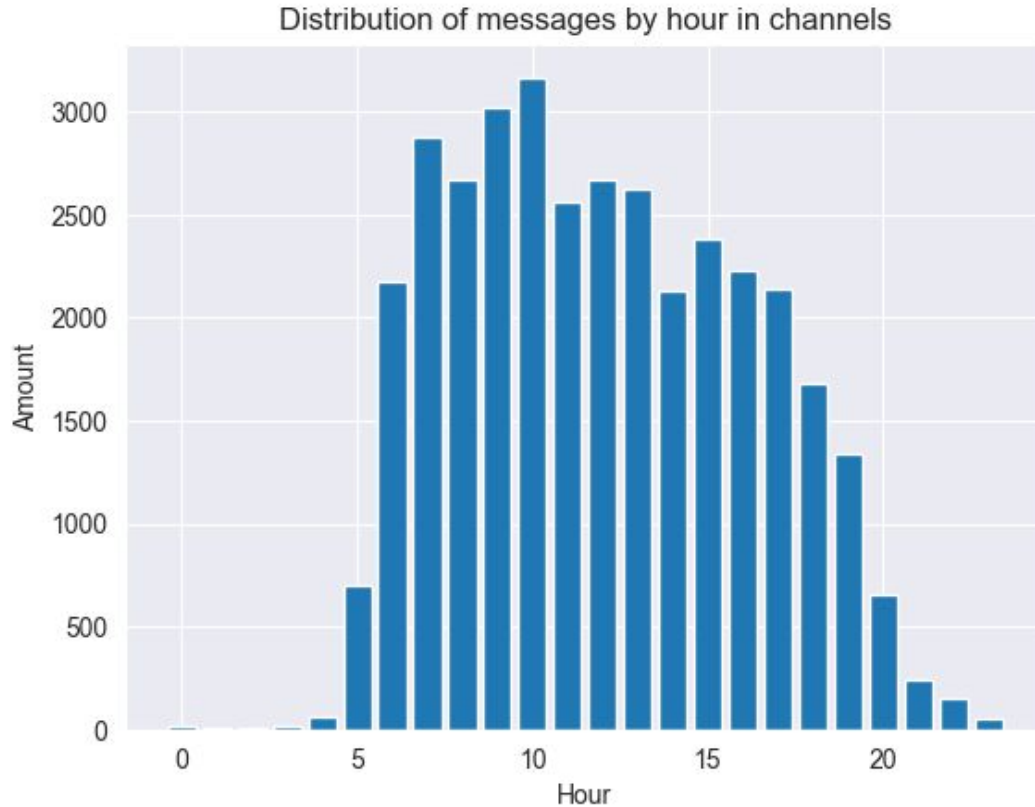


# Analysis of channels

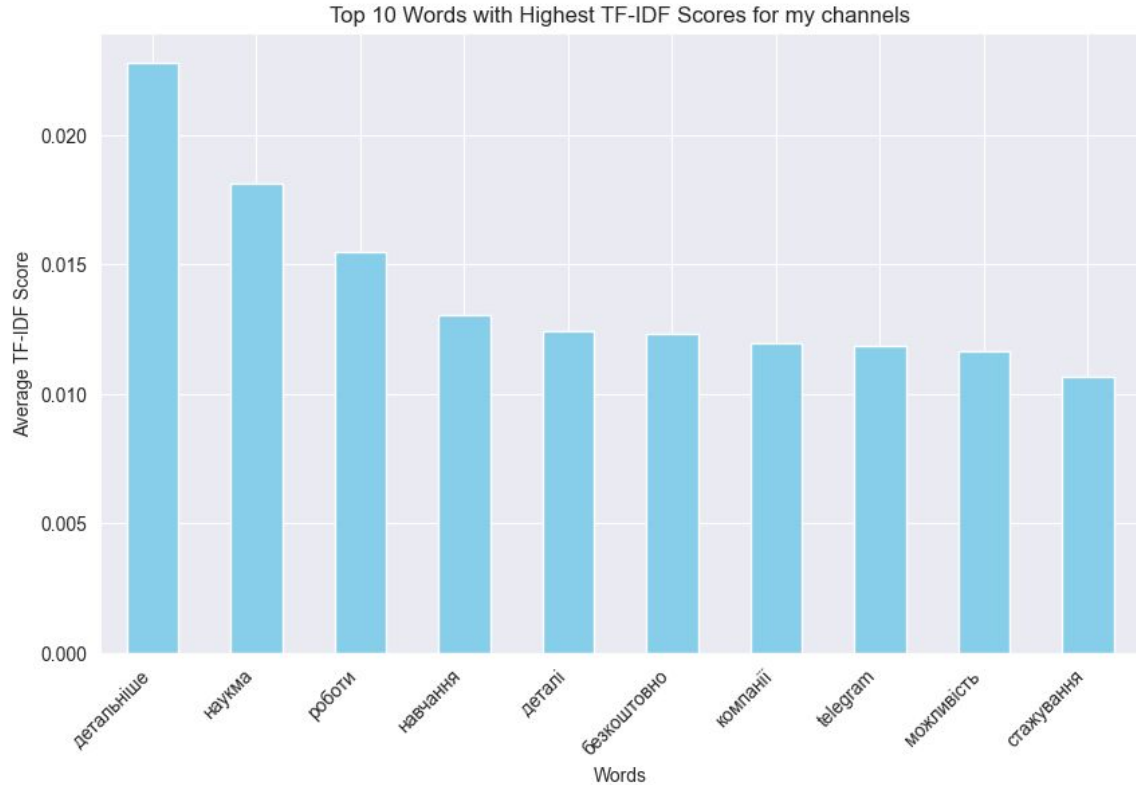
Language distribution in my channels



# Analysis of channels



# Analysis of channels



# Further work

Improving language detection of dialogs with informal vocabulary and mix of languages.

message	clean_message	language
Читай ше раз	читай	ru
Я беру CO i SiO2	беру sio	ru
Дякую дуже❤️		ru
Інна Штельмах приглашает вас на запланированну...	інна штельмах приглашает запланированную конф...	ru

# Git Repository

The screenshot shows a GitHub repository named 'telegram\_data\_analysis' by user 'yuliashtoliaruk'. The repository is public and has 1 branch (master) and 0 tags. The commit history shows three commits: 'Updated readme', 'Add main file of analysis and readme', and 'Add presentation', all from 5 days ago. The file list includes 'AnalysisPresentation.pdf', 'ExploratoryDataAnalysis.ipynb', and 'README.md'. The README content is visible, showing the title 'Telegram data analysis', a 'Files' section, and a 'How to use:' section with three steps. The right sidebar contains sections for 'About', 'Releases', 'Packages', and 'Languages', with a progress bar for 'Jupyter Notebook' at 100.0%.

telegram\_data\_analysis Public

Pin Unwatch 1 Fork 0 Star 0

master 1 branch 0 tags Go to file Add file <> Code

yuliashtoliaruk Updated readme 425fd48 5 days ago 3 commits

File	Commit	Time
AnalysisPresentation.pdf	Add presentation	5 days ago
ExploratoryDataAnalysis.ipynb	Add main file of analysis and readme	5 days ago
README.md	Updated readme	5 days ago

README.md

## Telegram data analysis

### Files

ExploratoryDataAnalysis.ipynb main file, where the data exploration is processing

AnalysisPresentation.pdf presentation of my data analysis

### How to use:

1. Download data using this [repository](#)
2. Merge your data using instructions from this [repository](#)
3. Run `ExploratoryDataAnalysis.ipynb`

### About

Data analysis of messages in Telegram

Readme Activity 0 stars 1 watching 0 forks

### Releases

No releases published [Create a new release](#)

### Packages

No packages published [Publish your first package](#)

### Languages

Jupyter Notebook 100.0%