Telegram data-analysis

Student: Stoliaruk Yuliia
Teacher: Andrew Kurochkin

Plan of the presentation

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

Introduction

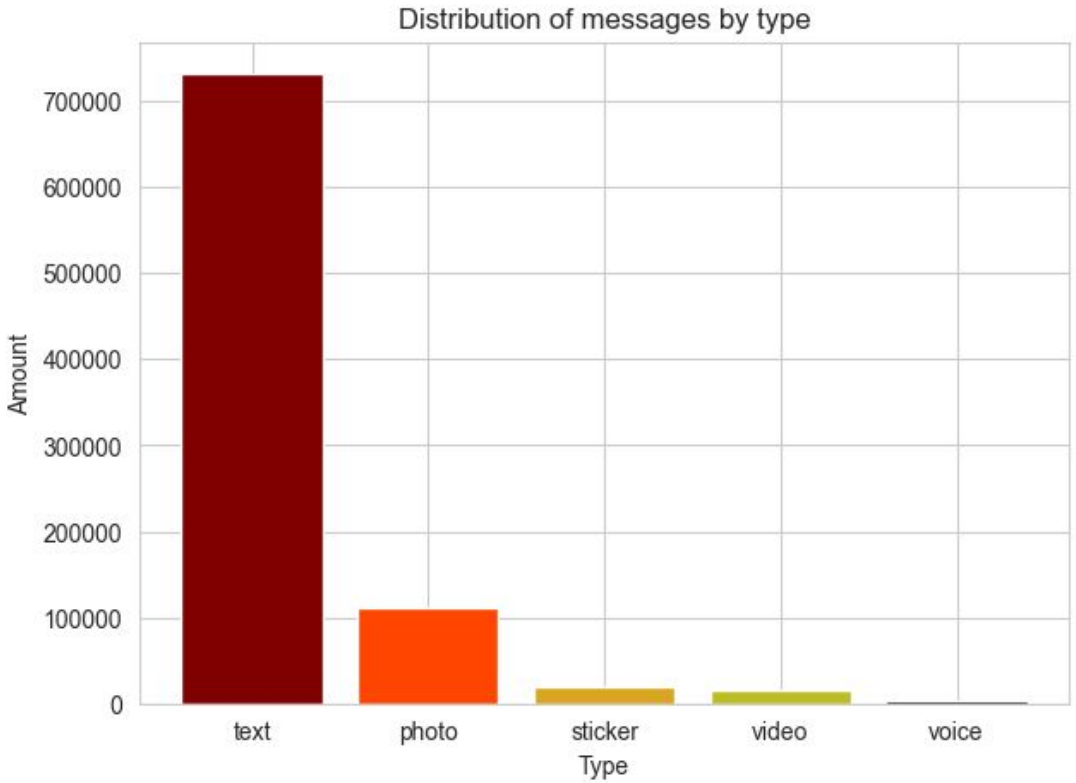
The main purpose of this project is to analyse activity of users and find hidden patterns in messenger Telegram.

How to get data

To get your data from Telegram:

1. To download data from Telegram use this repository SanGreel/telegram-data-collection (github.com)
2. Then merge collected data using this repository SanGreel/telegram-dialogs-analysis-v2 (github.com)

Data overview (Messages data)

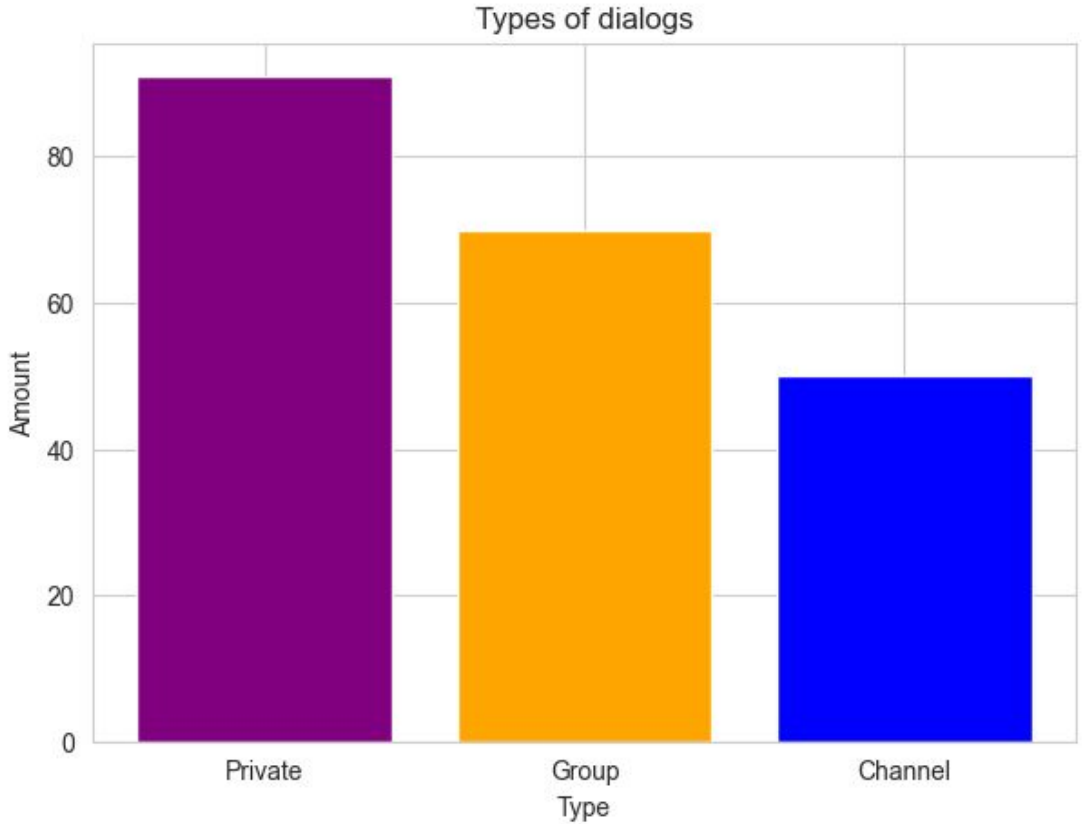


Size: 337.849 MB

Messages number: 880 940

Data overview (Dialogs data)

Size: 1.85 MB

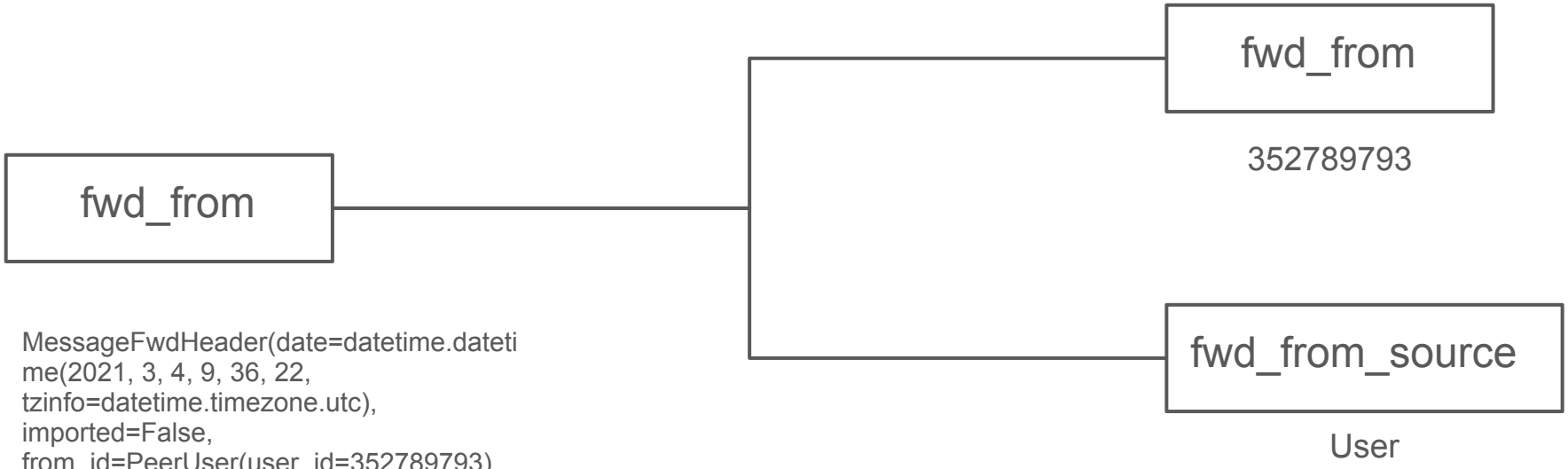


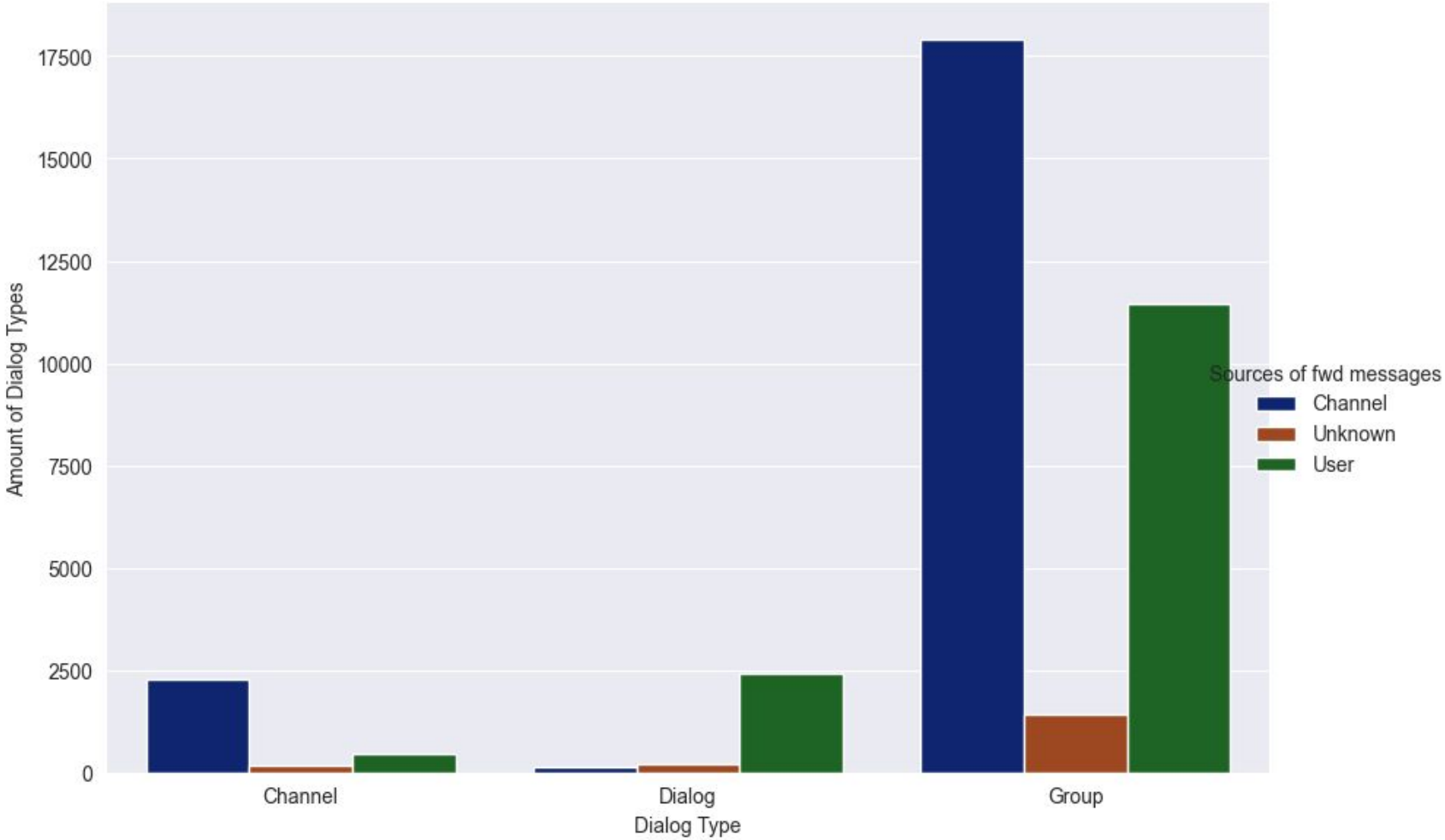
Exploratory Data Analysis (Data cleaning)

The first step was to clear the data.

1. Cleaning ‘fwd_from’ column
2. Cleaning ‘from_id’ column
3. Cleaning ‘message’ column

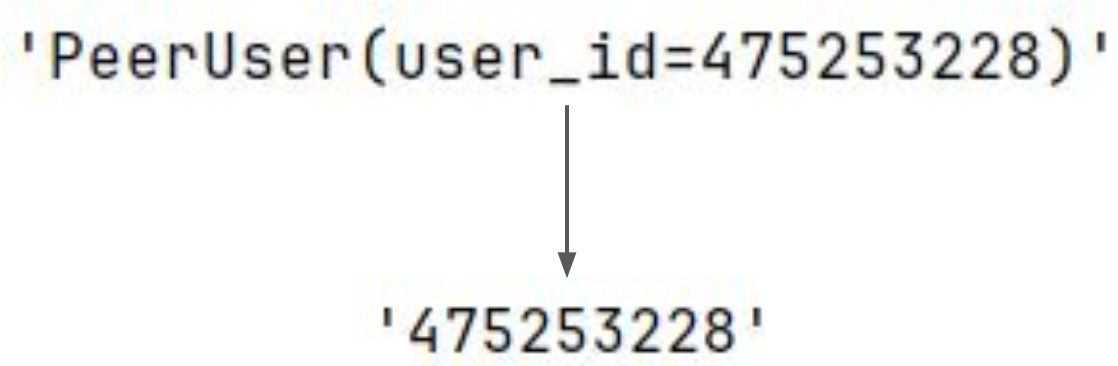
Cleaning ‘fwd_from’ column



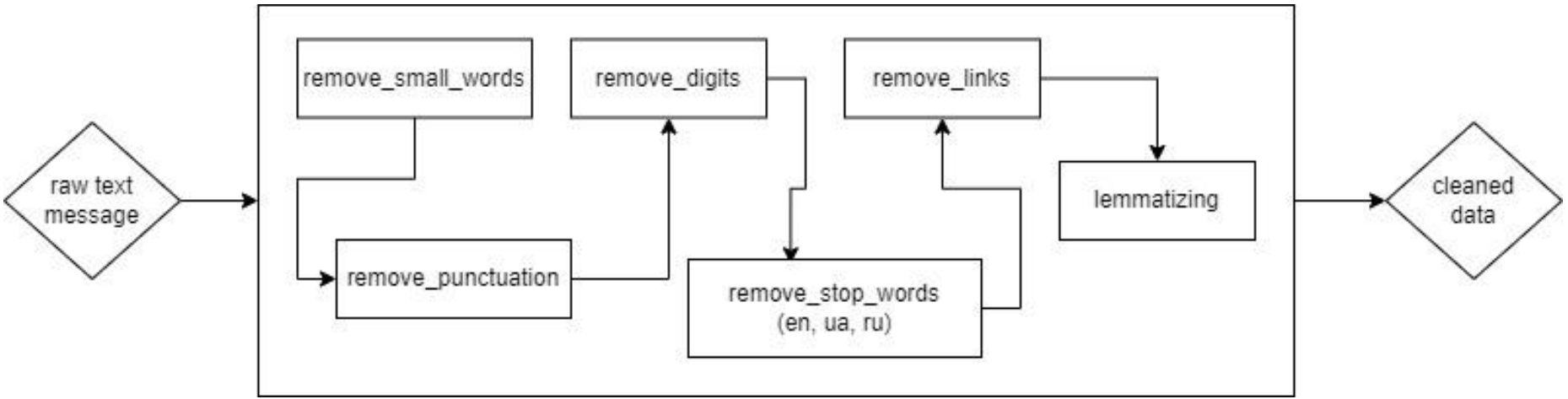

17500
15000
12500
10000
7500
5000
2500
0
Amount of Dialog Types
Channel
Dialog
Group
Dialog Type
Sources of fwd messages
Channel
Unknown
User

Cleaning ‘from_id’ column

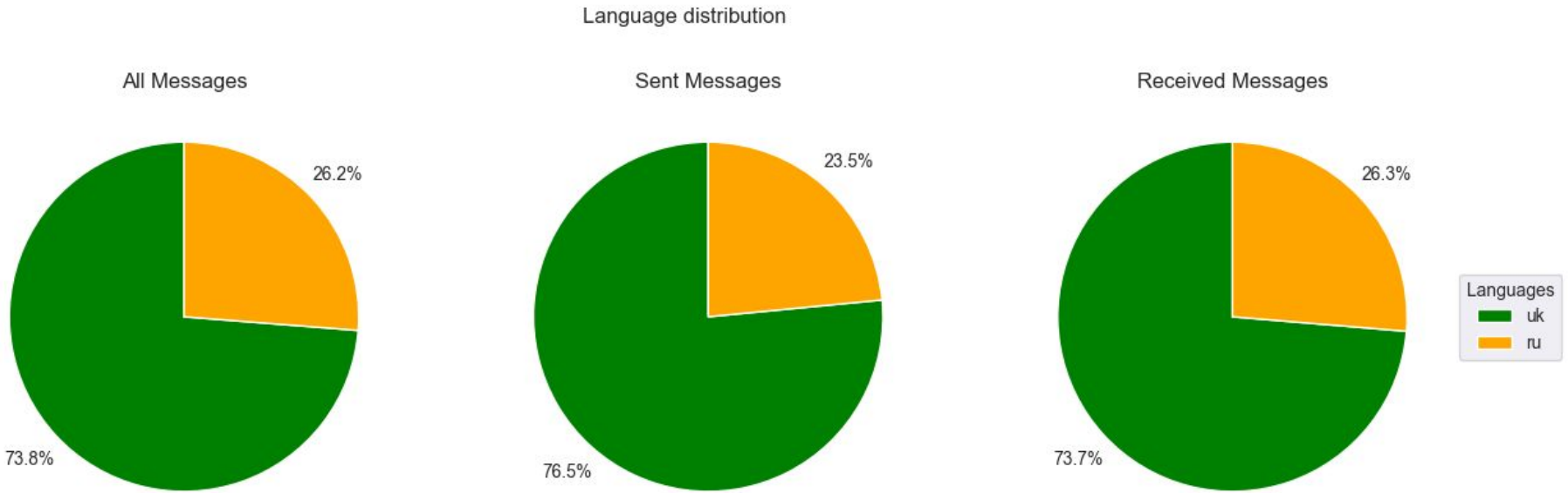
I cleaned this column for easier analysis.



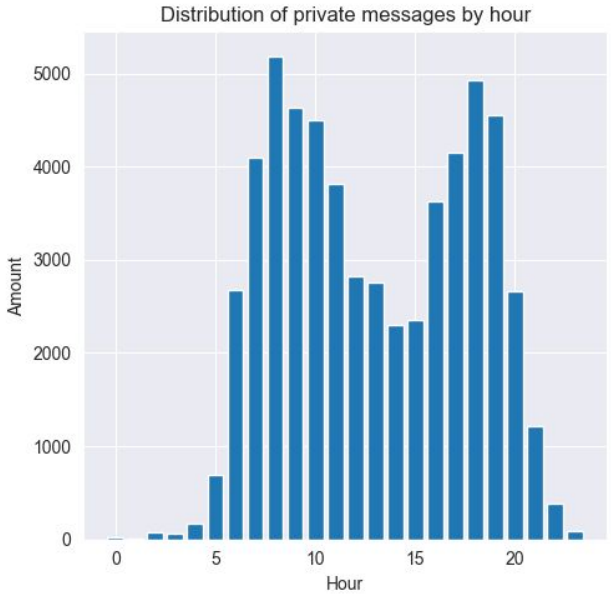
Cleaning ‘message’ column

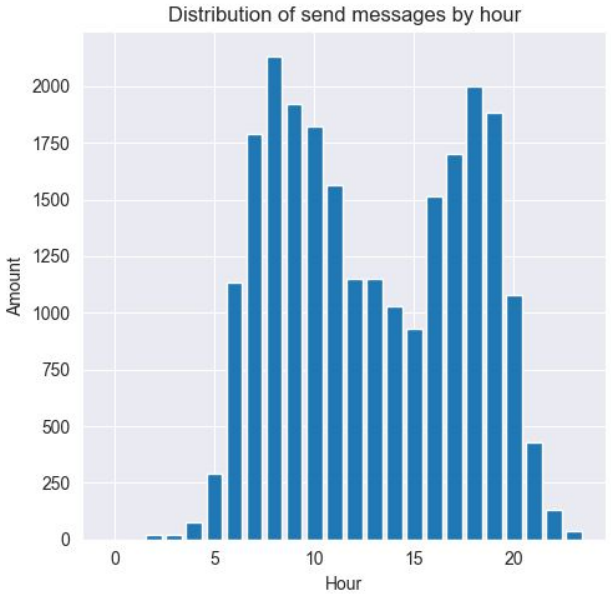


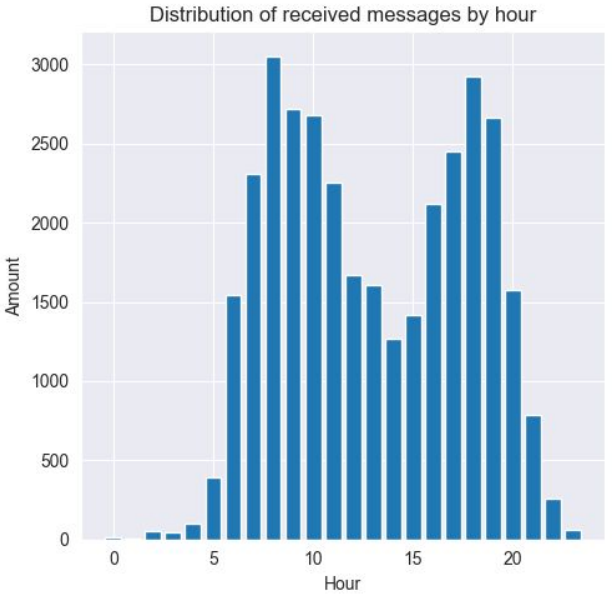
General analysis of my messages



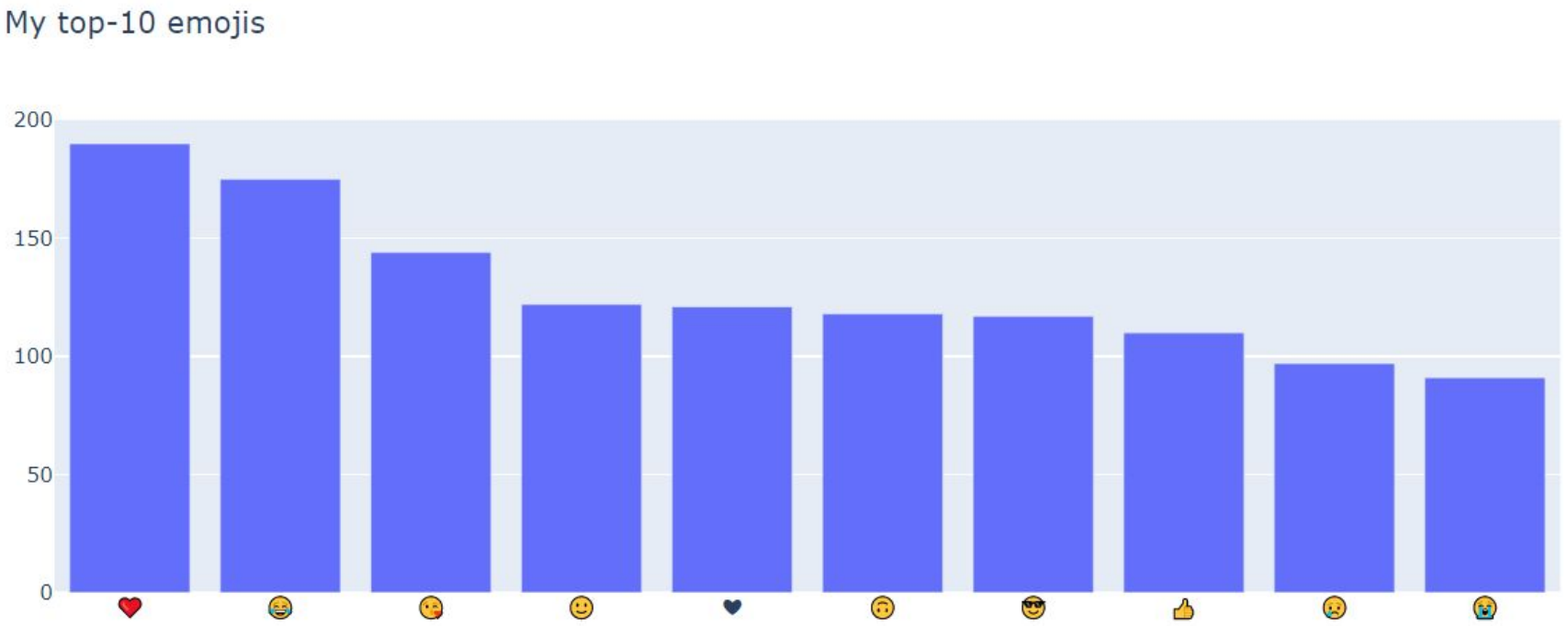
General analysis of my messages



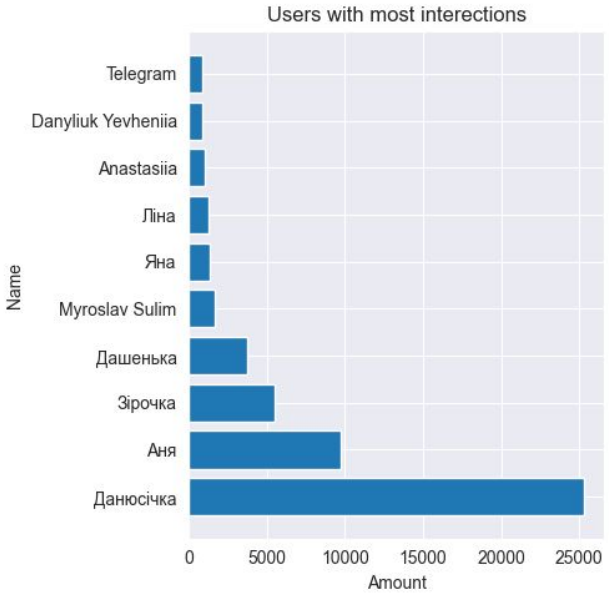


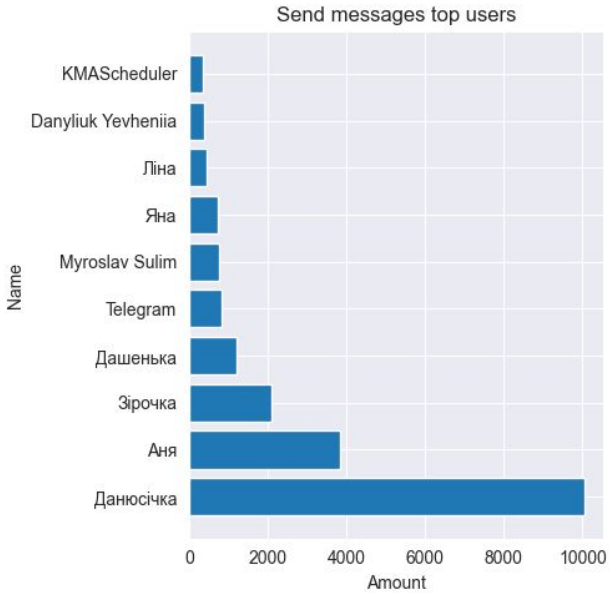


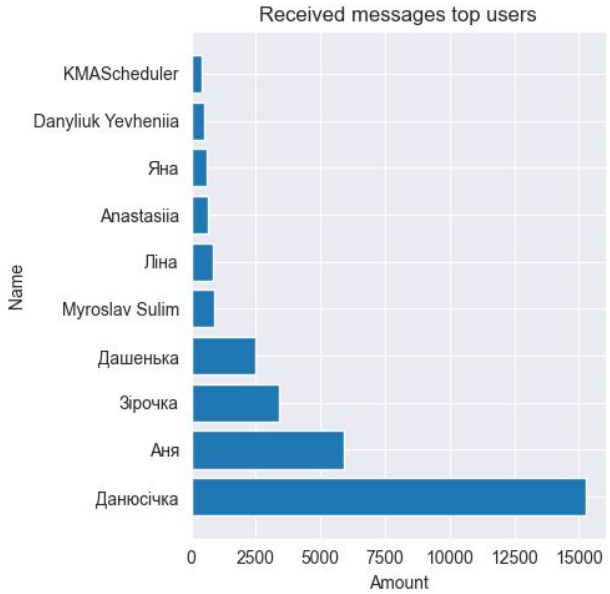
General analysis of my messages



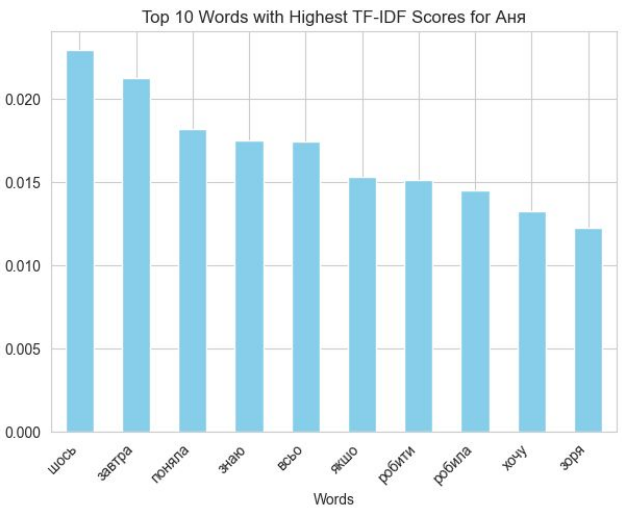
Private dialogs analysis

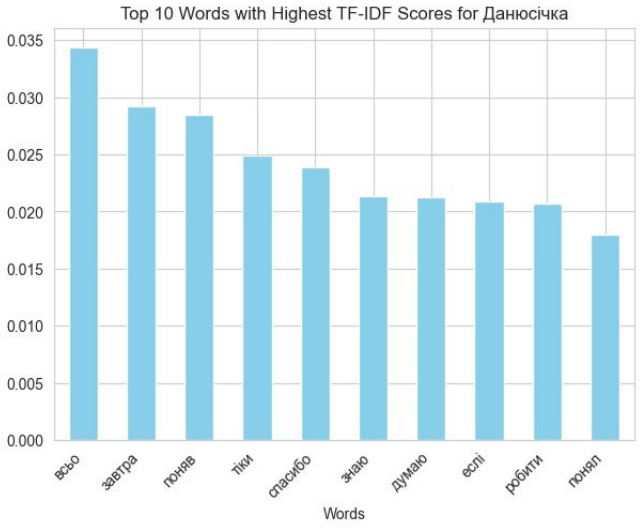


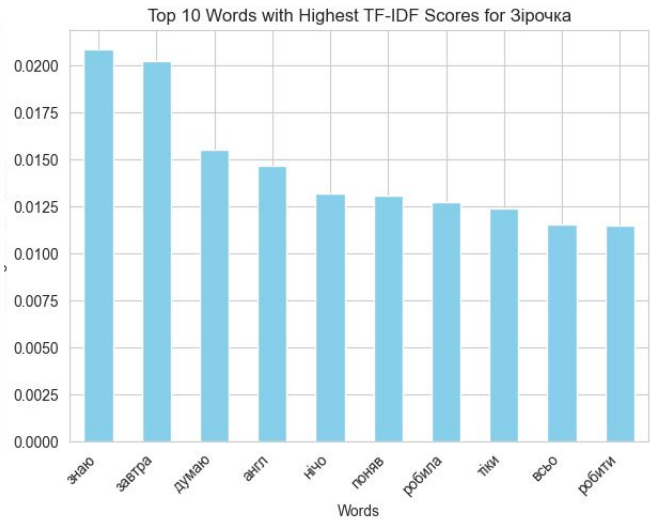


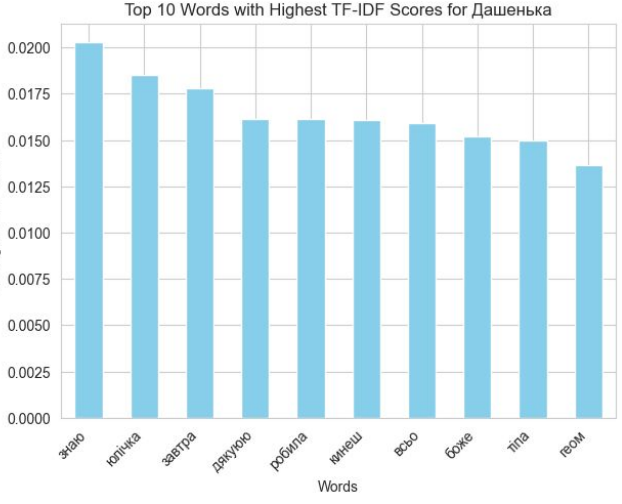


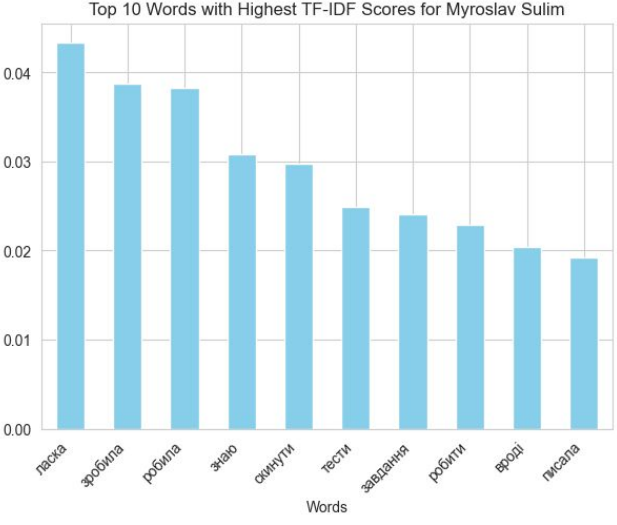
Top emojis of users with whom I interact the most

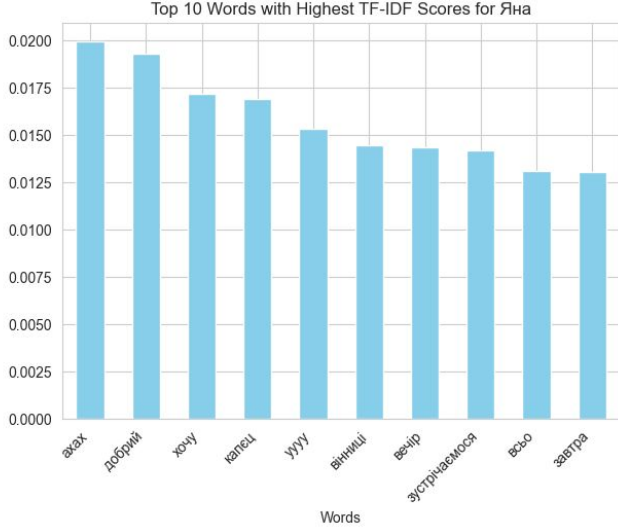

Top 10 Words with Highest TF-IDF Scores for Аня
0.020
0.015
0.010
0.005
0.000
шось
завтра
поняла
знаю
всьо
якшо
робити
робила
хочу
зоря
Words


Top 10 Words with Highest TF-IDF Scores for Данюсічка
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
всьо
завтра
поняв
тіки
спасибо
знаю
думаю
еслі
робити
понял
Words


Top 10 Words with Highest TF-IDF Scores for Зірочка
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
знаю
завтра
думаю
англ
нічо
поняв
робила
тіки
всьо
робити
Words


Top 10 Words with Highest TF-IDF Scores for Дашенька
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
знаю
юлічка
завтра
дякую
робила
киньш
всьо
боже
тіпа
геом
Words


Top 10 Words with Highest TF-IDF Scores for Myroslav Sulim
0.04
0.03
0.02
0.01
0.00
ласка
зробила
робила
знаю
скинути
тести
завдання
робити
броді
писала
Words

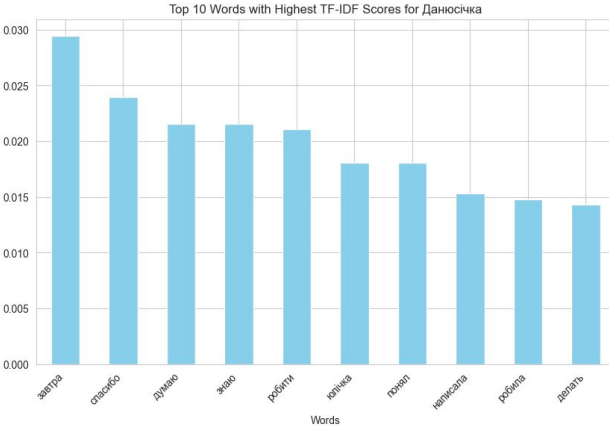

Top 10 Words with Highest TF-IDF Scores for Яна
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
ахах
добрий
хочу
капєц
ууу
вінніці
вечір
зустрічаємося
всьо
завтра
Words

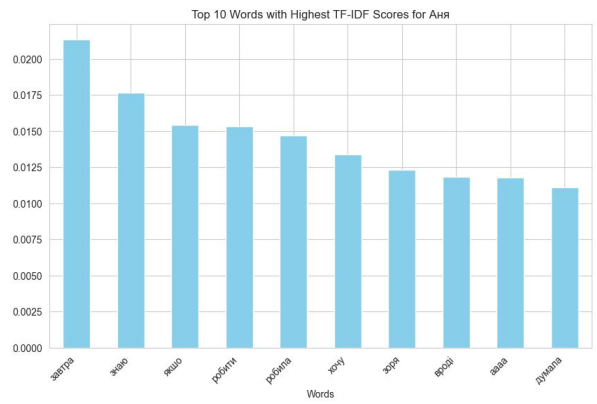
Data cleaning problem

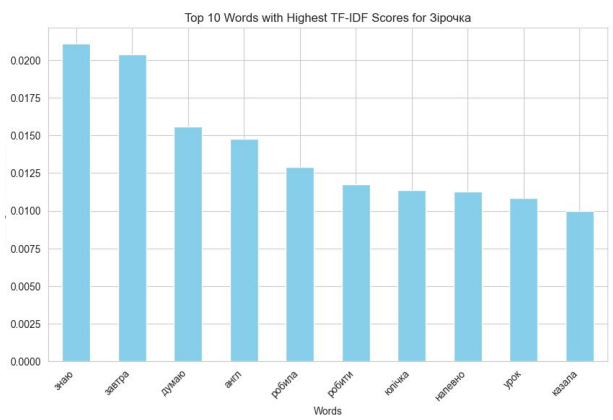
Histograms of words with top tf-idf show us that there are still a lot of words that don’t give meaningful information about dialogs.

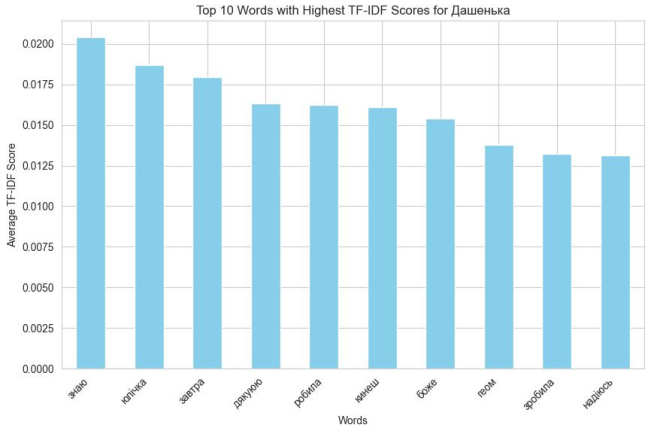
To fix this problem I created new additional dictionary of stop-words.

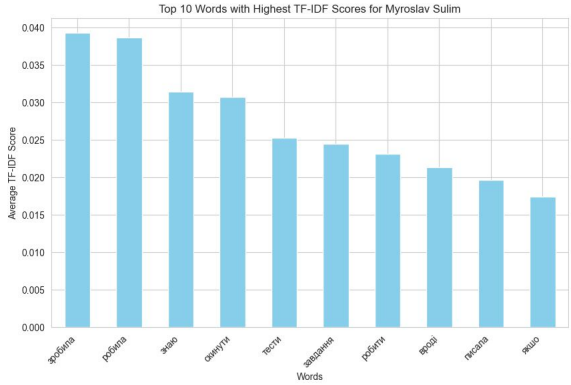
Updated tf-idf

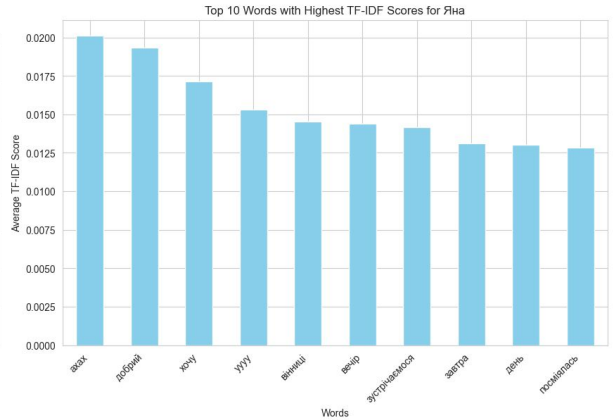




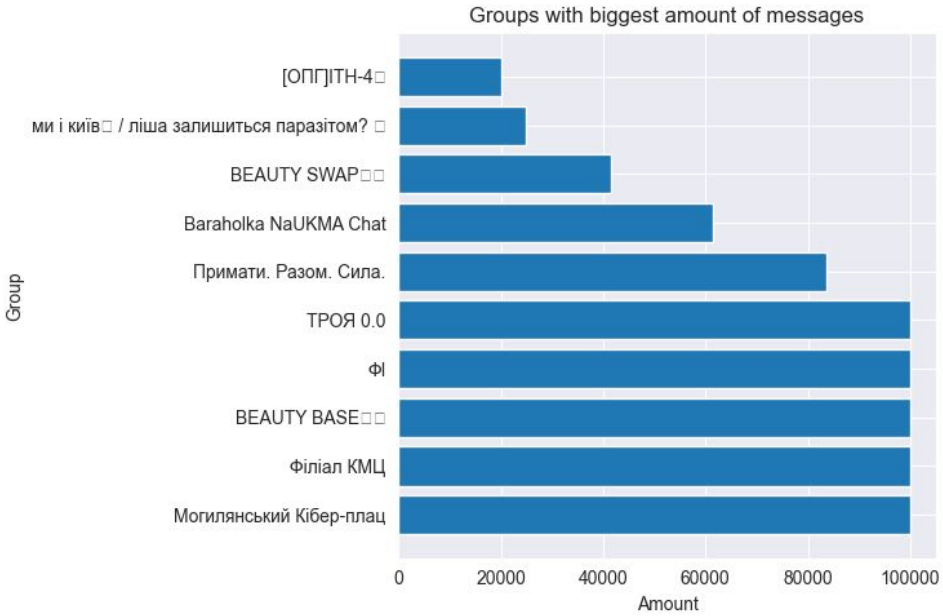


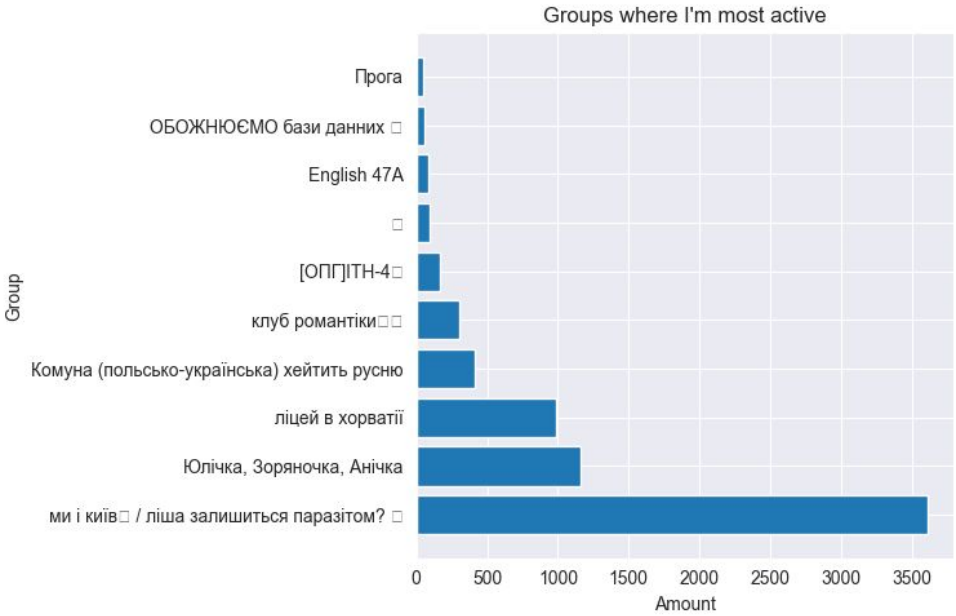






Groups analysis





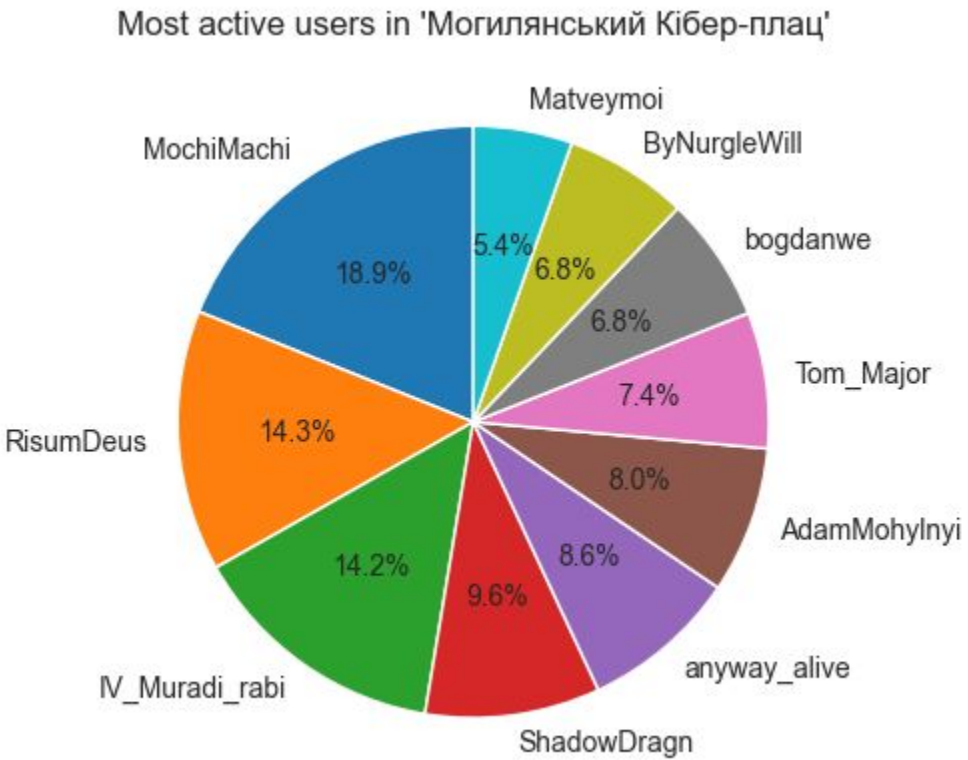
Analysis of Могилянський Кібер-плац

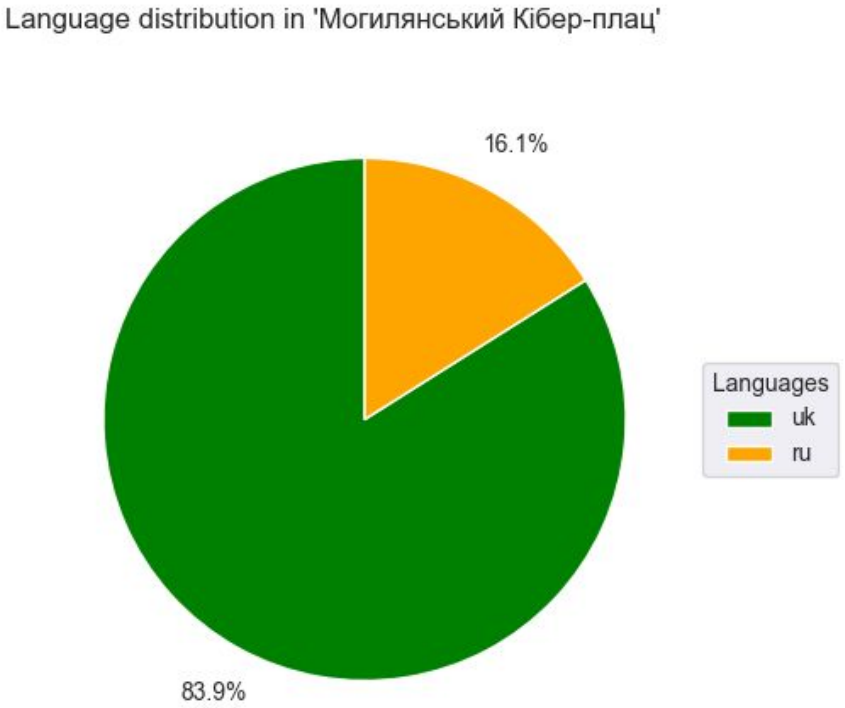
General:

Amount of messages: 100 000

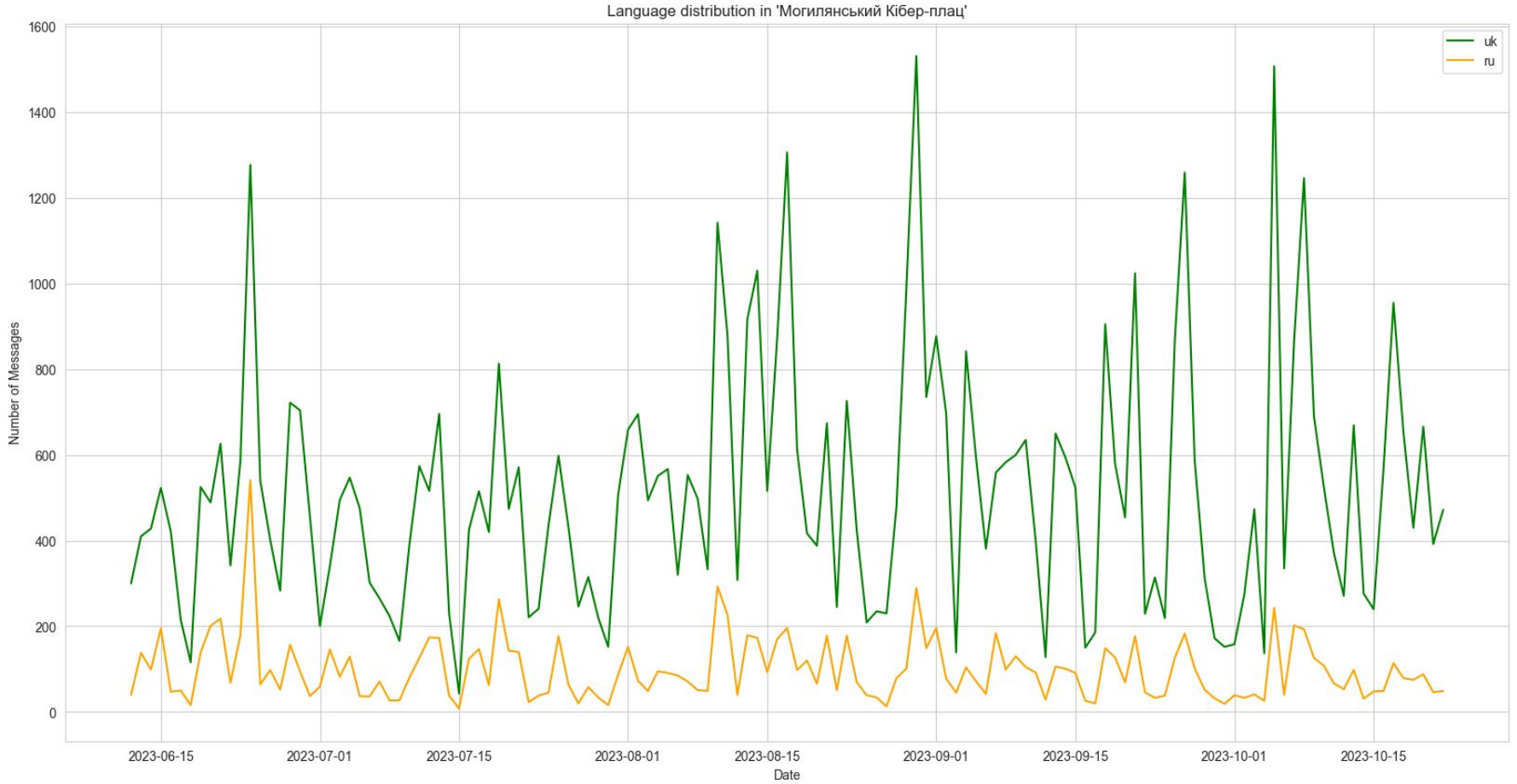
Amount of users: 1613

Analysis of Могилянський Кібер-плац

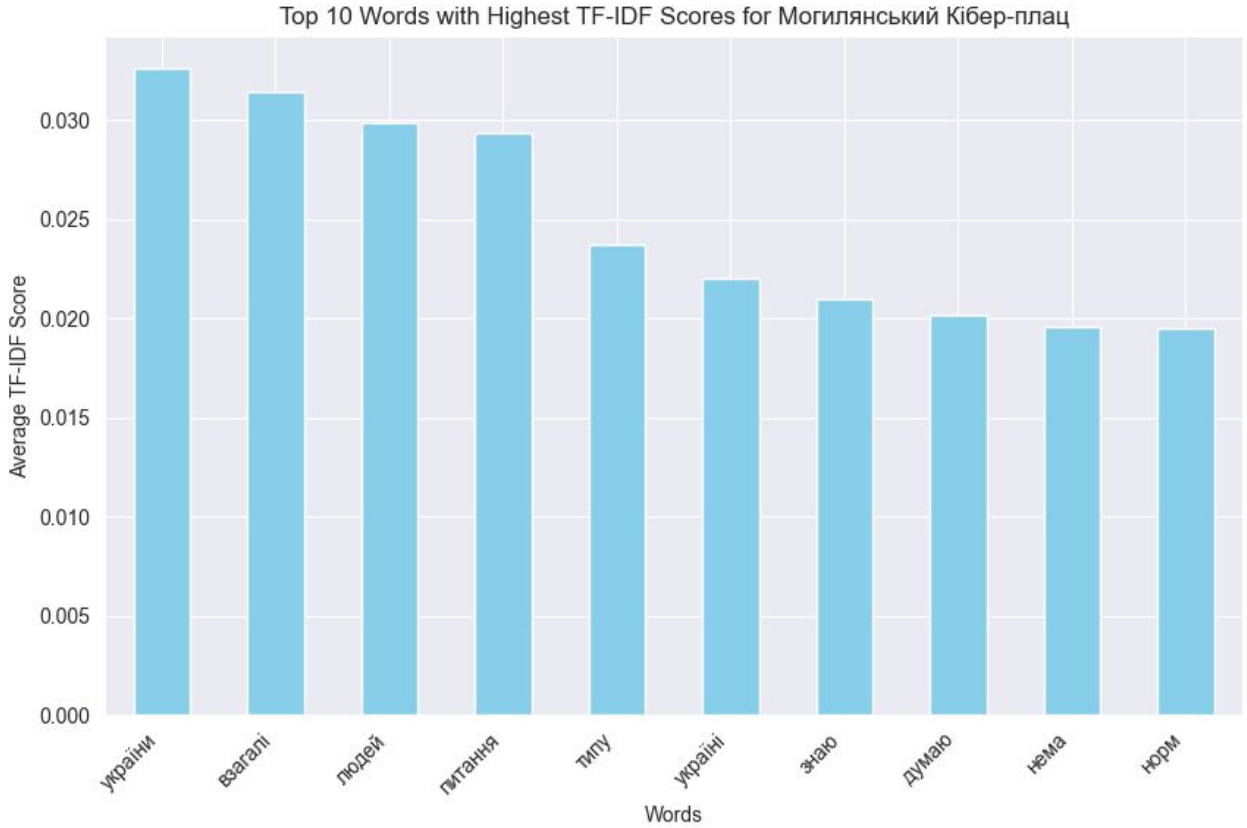




Analysis of Могилянський Кібер-плац



Analysis of Могилянський Кібер-плац



Analysis of Могилянський Кібер-плац

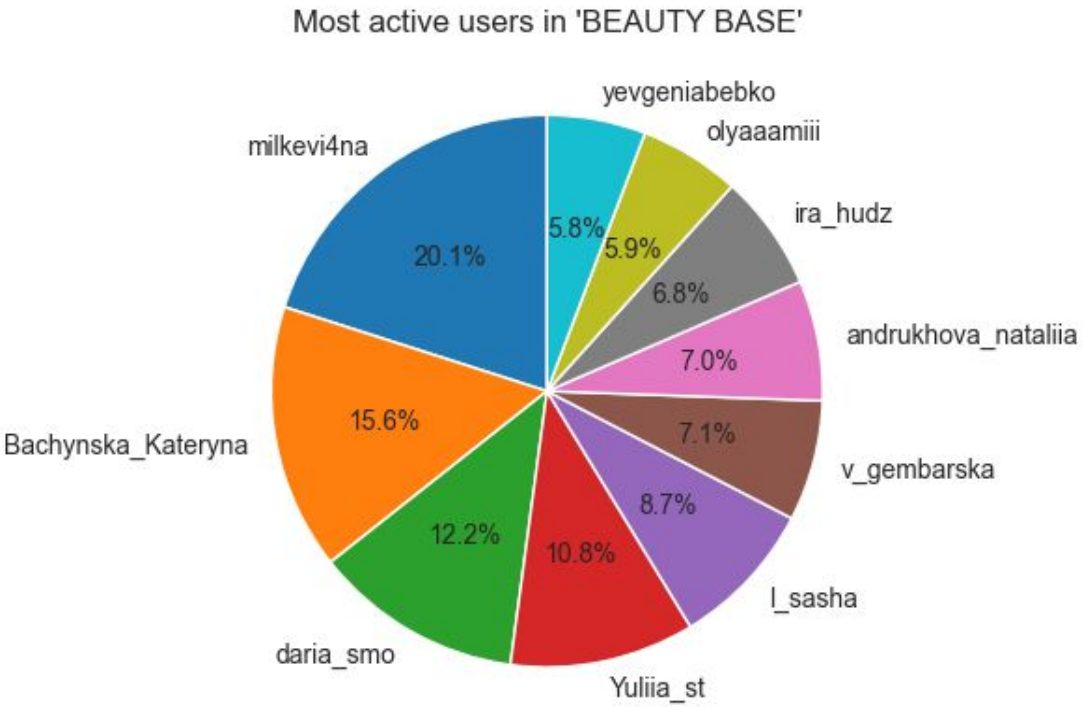
Analysis of BEAUTY BASE 🧴📝

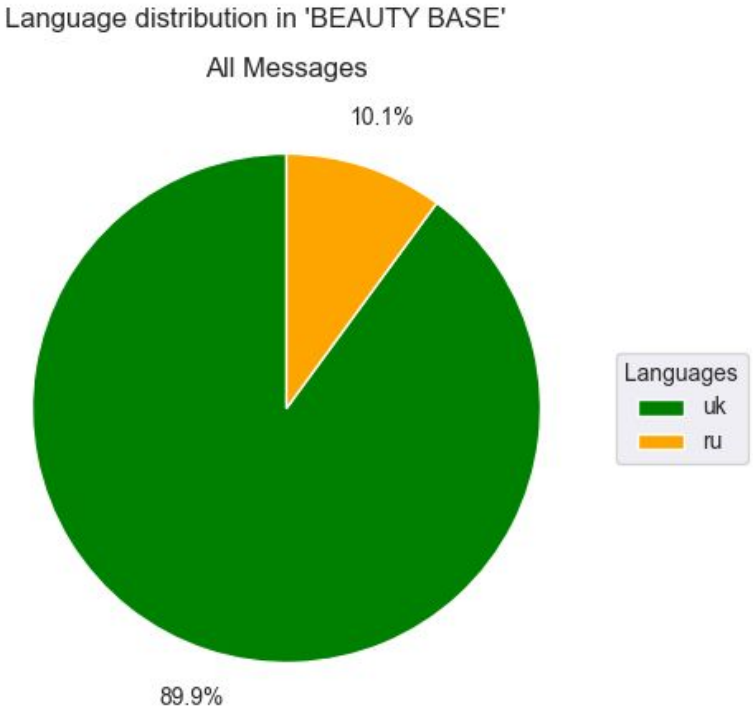
General:

Amount of messages: 100 000

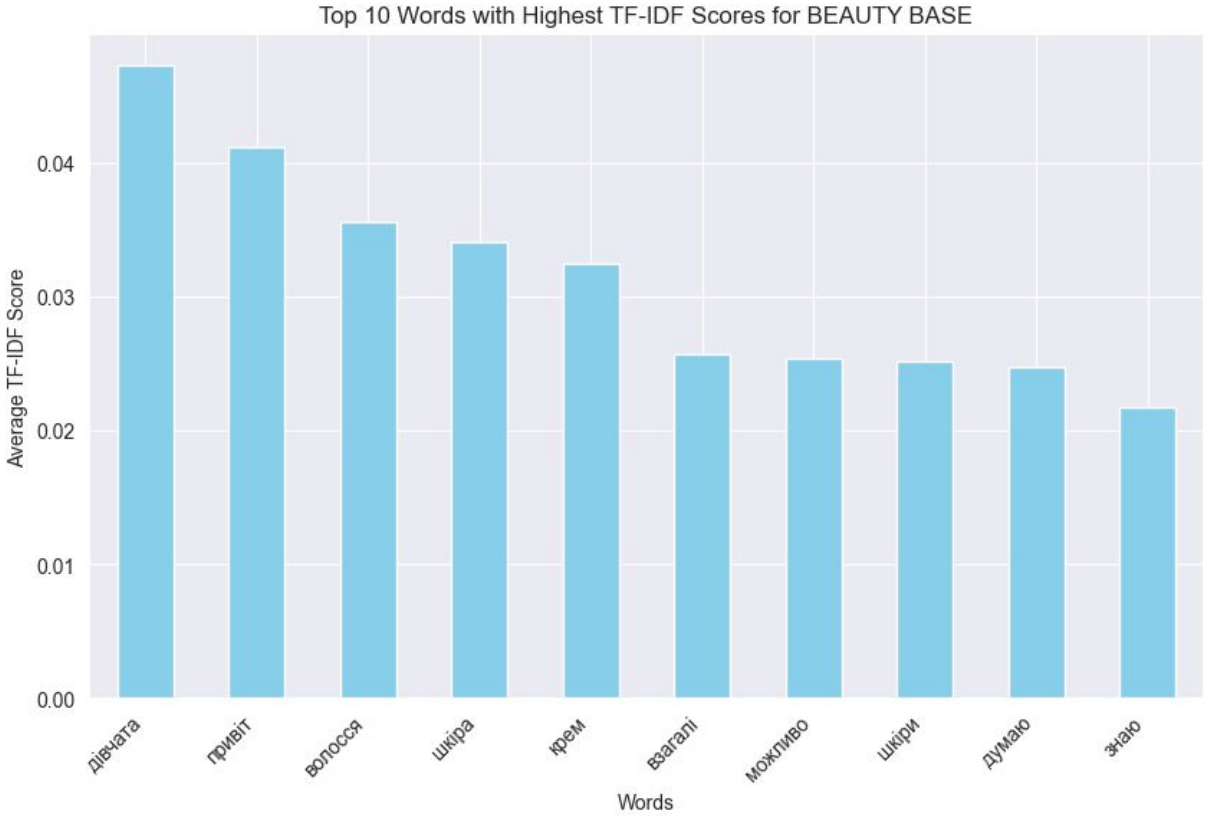
Amount of users: 1613

Analysis of BEAUTY BASE 🧴📝





Analysis of BEAUTY BASE 🧴📝



Analysis of BEAUTY BASE 🧴📝

Top-10 emojis of 'BEAUTY BASE'

Further work

Improving language detection of dialogs, where messages

Git Repository

GitLink