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:

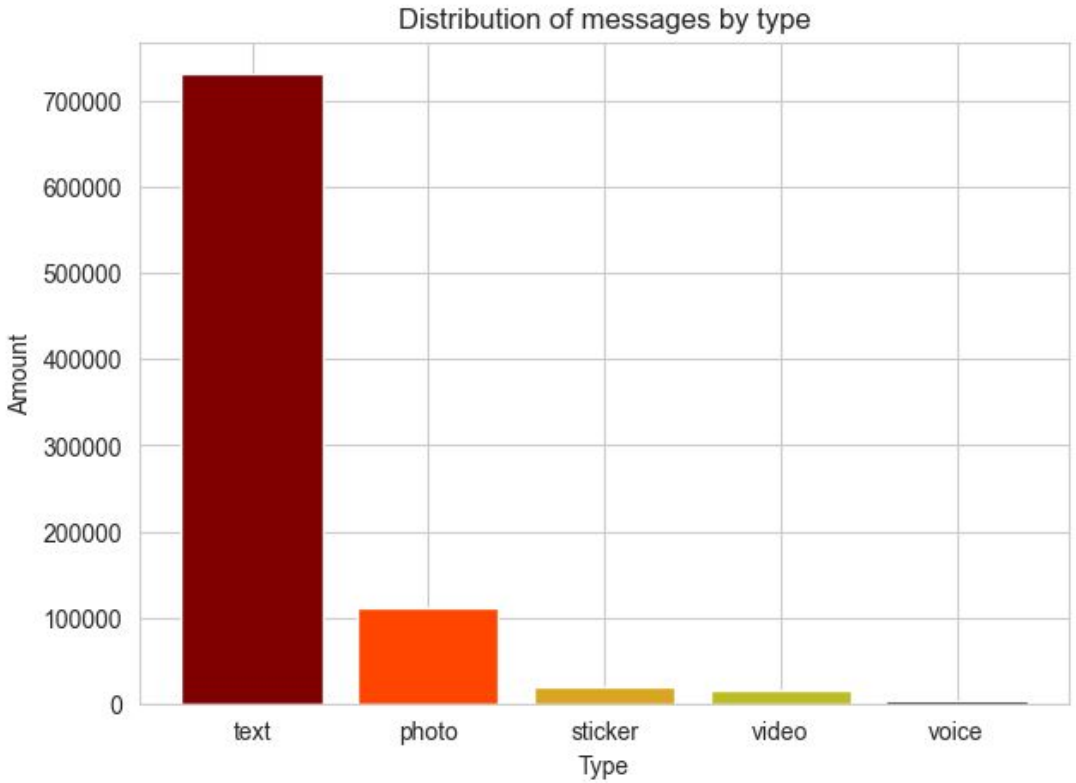
1. To download raw data from Telegram use this repository SanGreel/telegram-data-collection (github.com)

   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)

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

Data overview (Messages data)

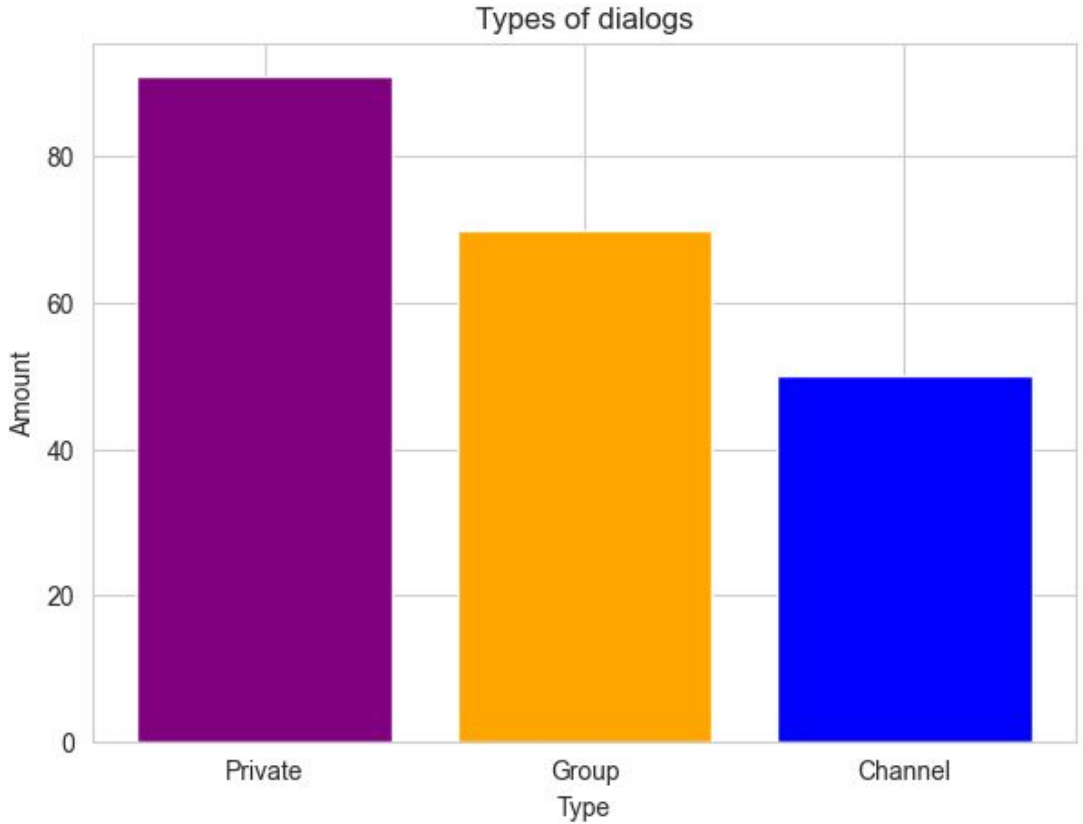


Size: 206.634 MB

Messages number: 880 940

Data overview (Dialogs data)

Size: 1.85 MB

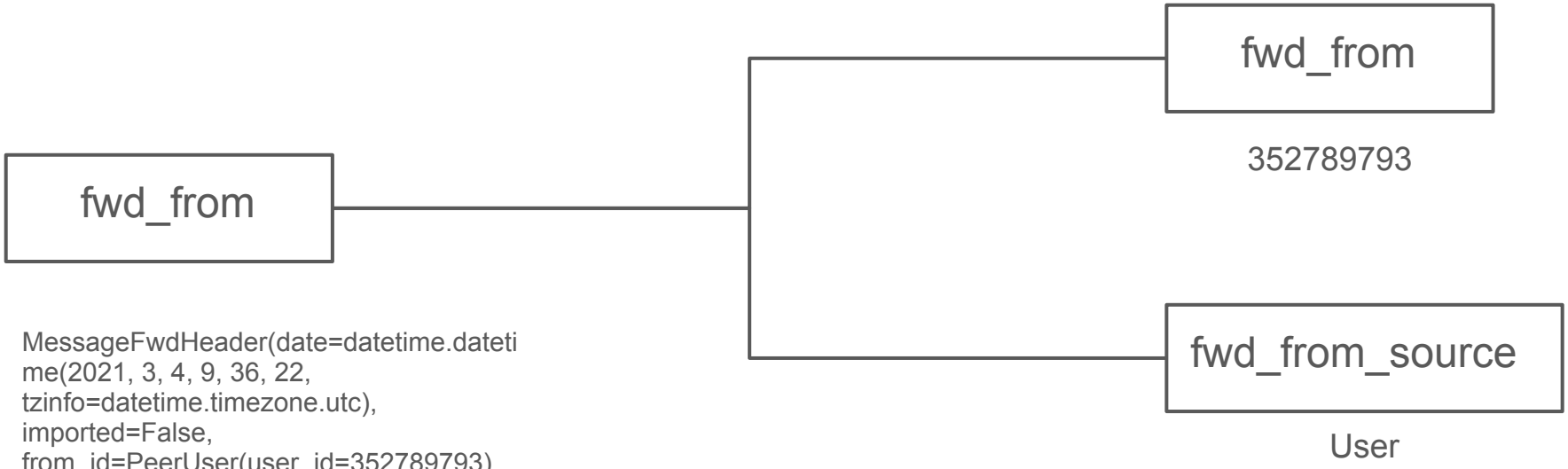


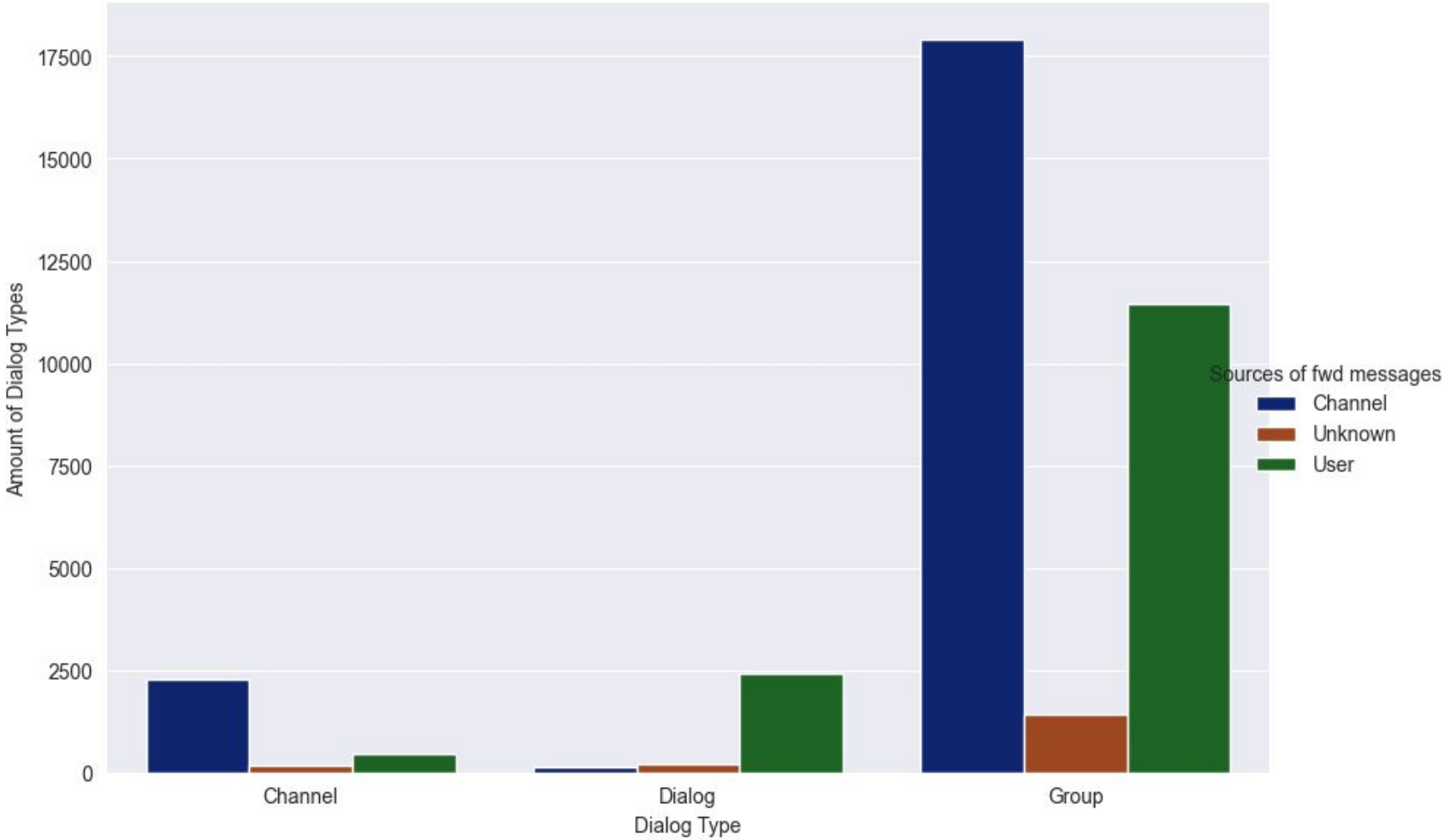
Exploratory Data Analysis (Data cleaning)

The first step was to clear the data.

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

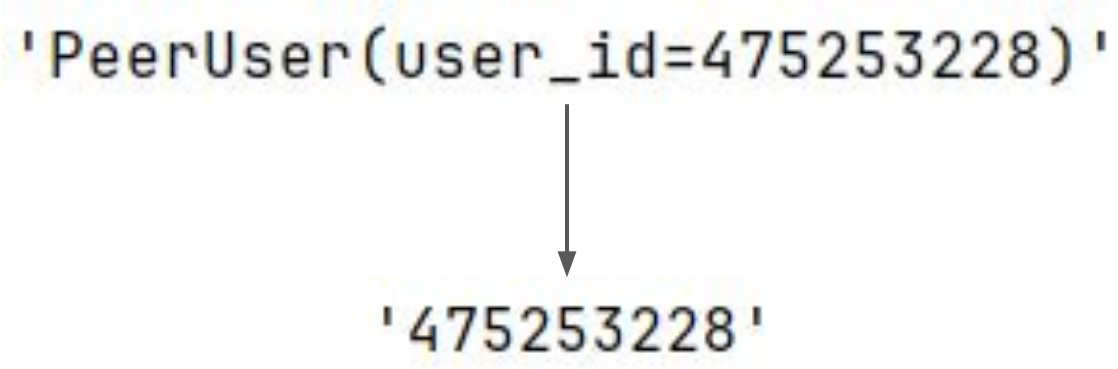
Cleaning ‘fwd_from’ column




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

To use column ‘from_id’ in further analysis i cleaned it from unnecessary data.



Cleaning ‘message’ column

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

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

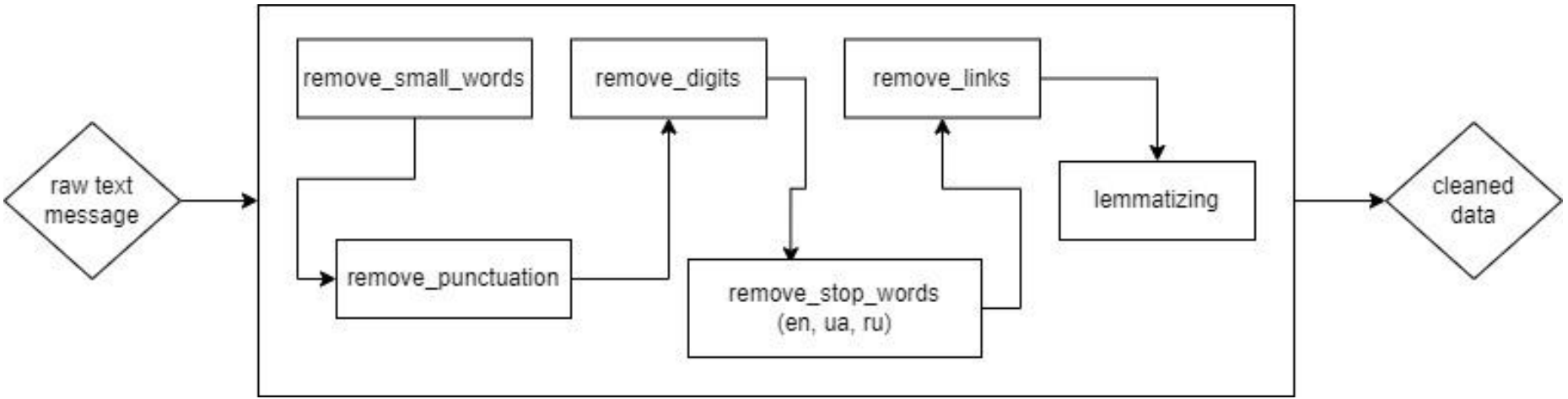
To remove stop-words I used dictionary of english and russian words from “nltk” library.

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

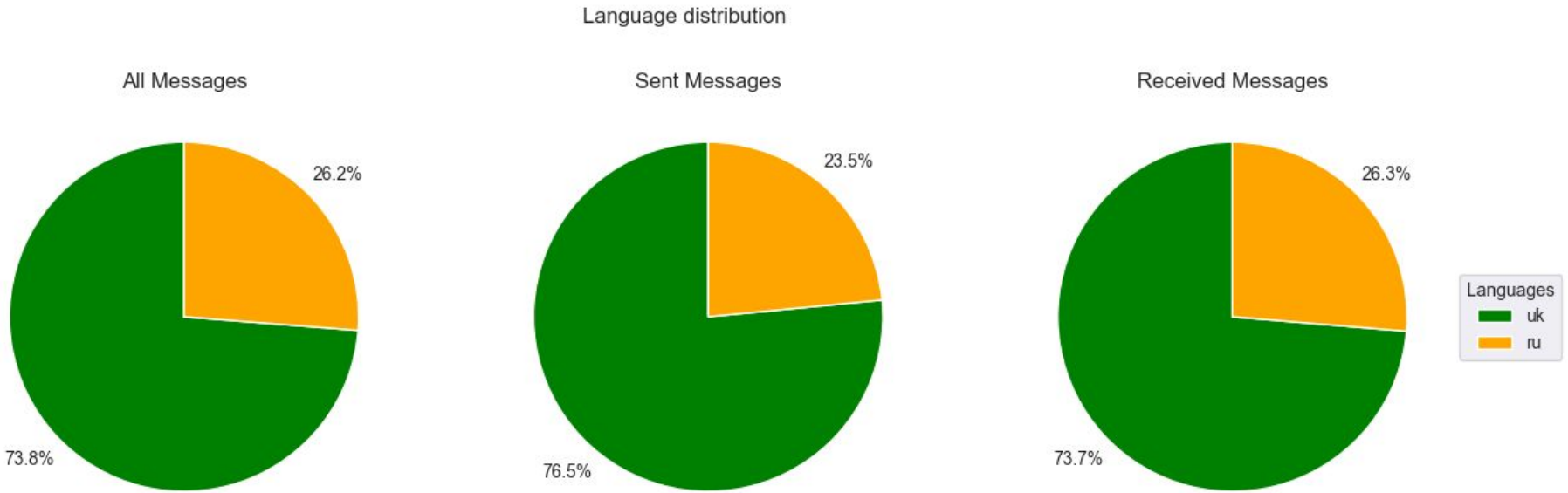


Final results

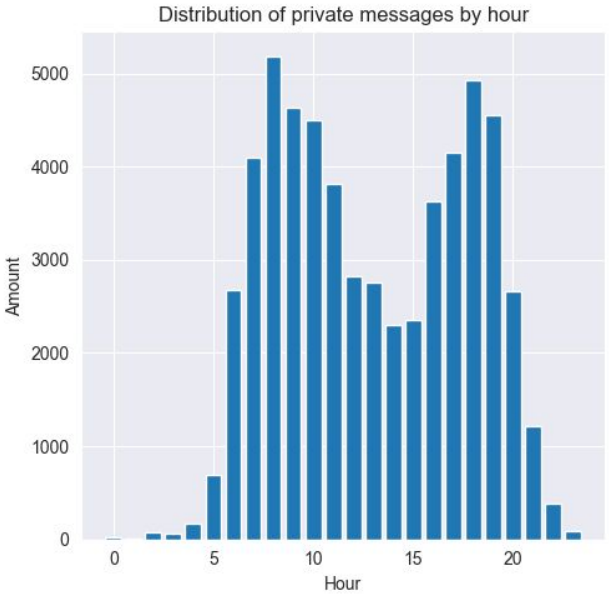
- General analysis
- Private dialog analysis
- Groups analysis
- Chanels analysis

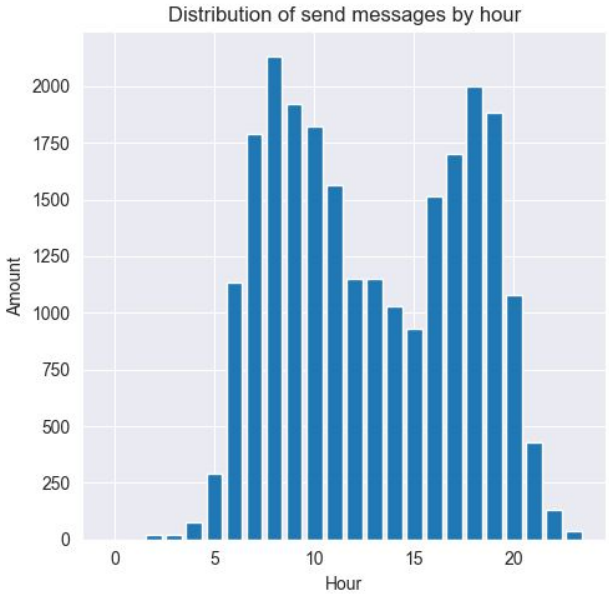
General analysis of my messages

General analysis of my messages

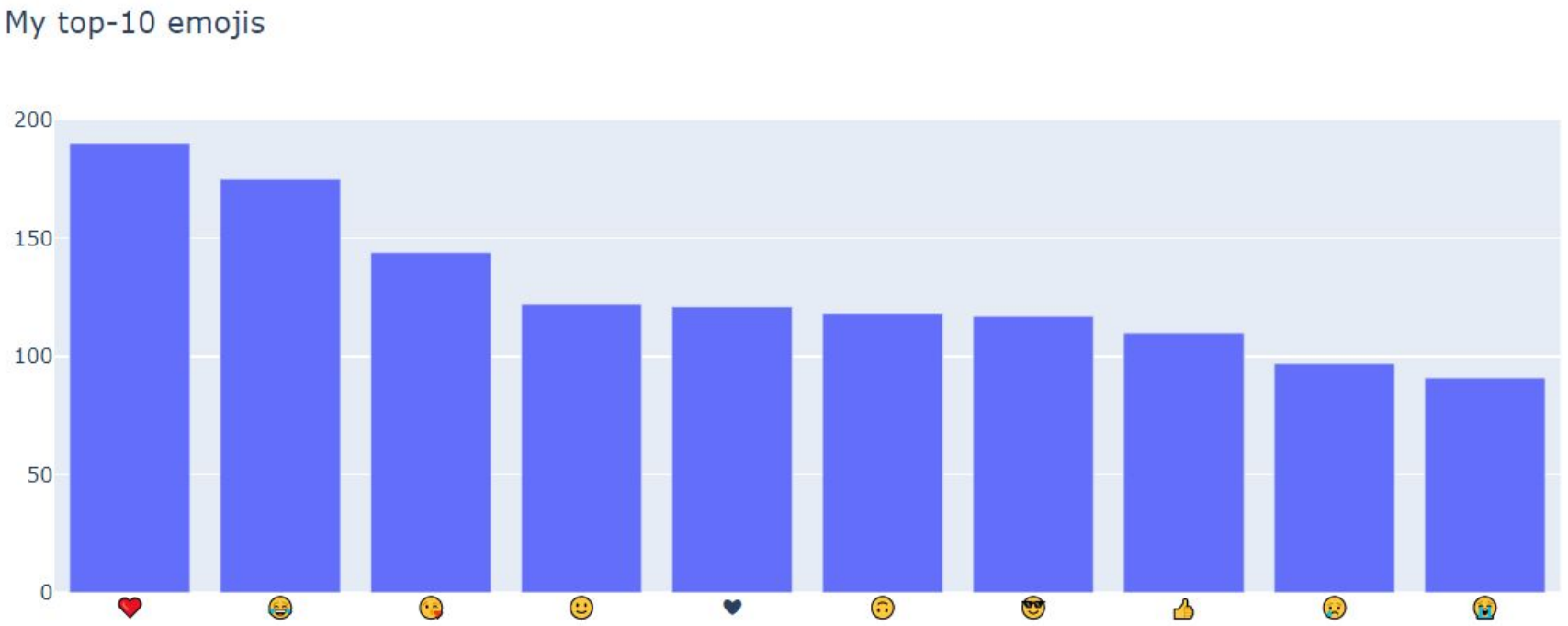


General analysis of my messages



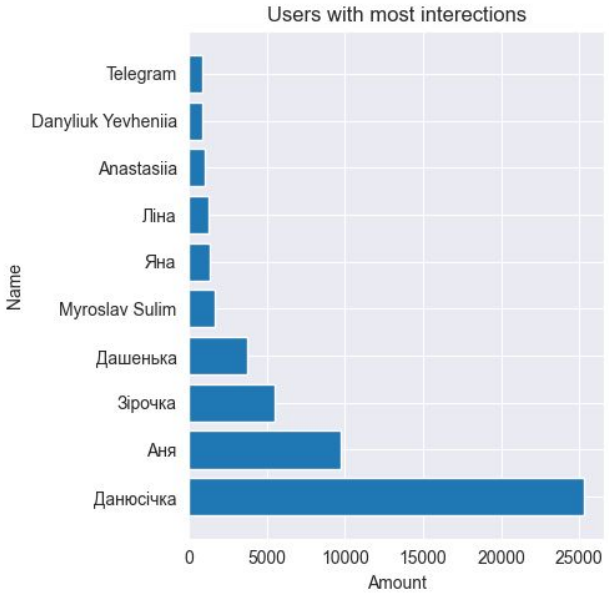


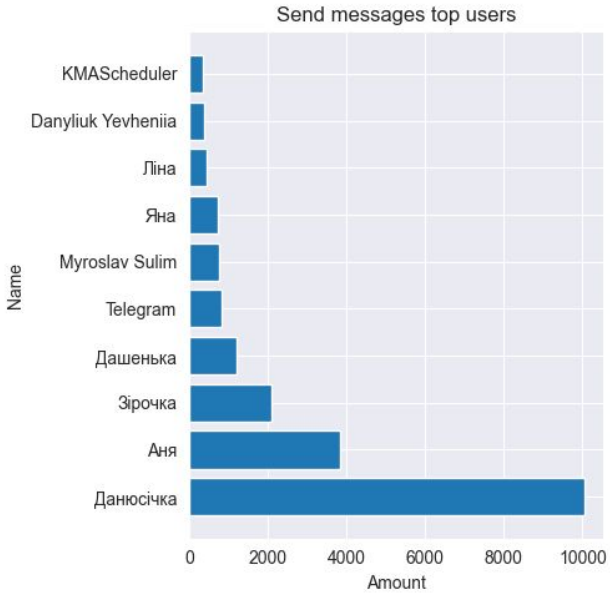
General analysis of my messages

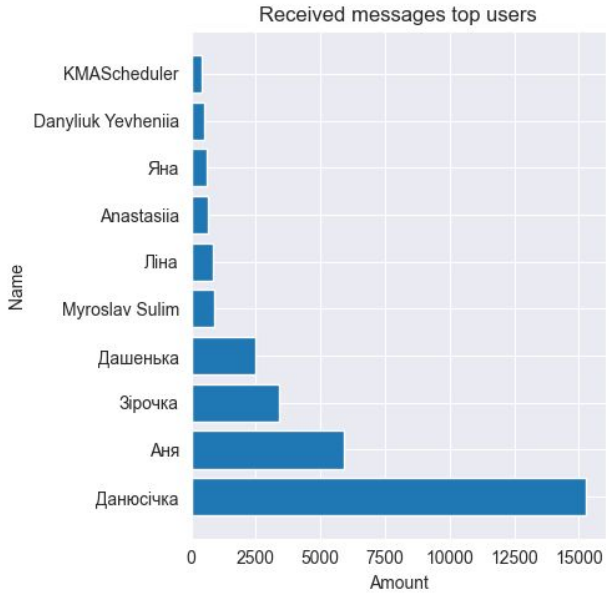


Private dialogs analysis

Private dialogs analysis



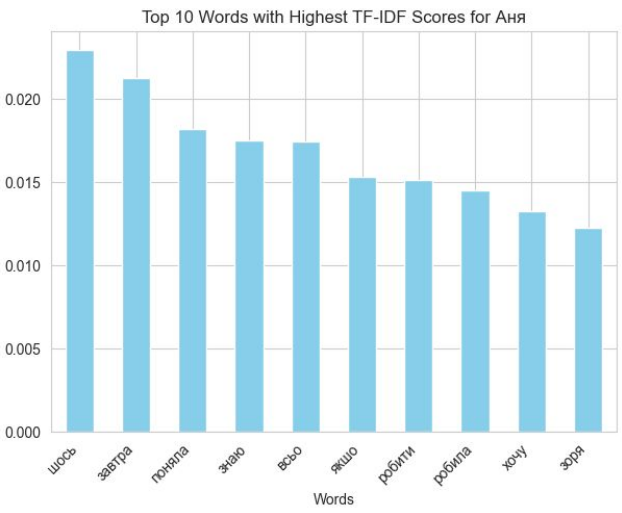


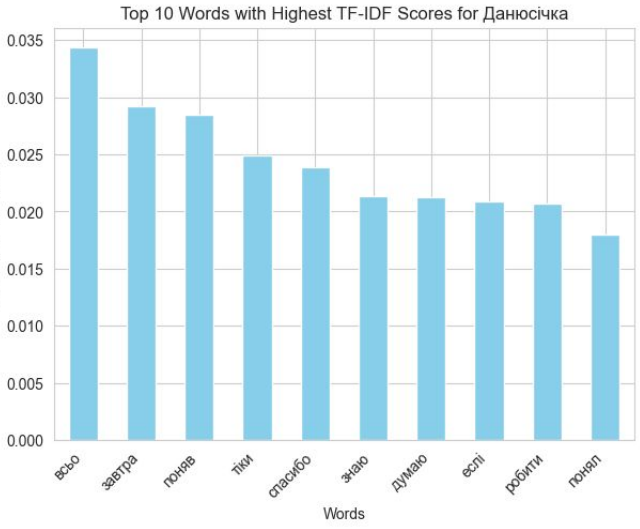


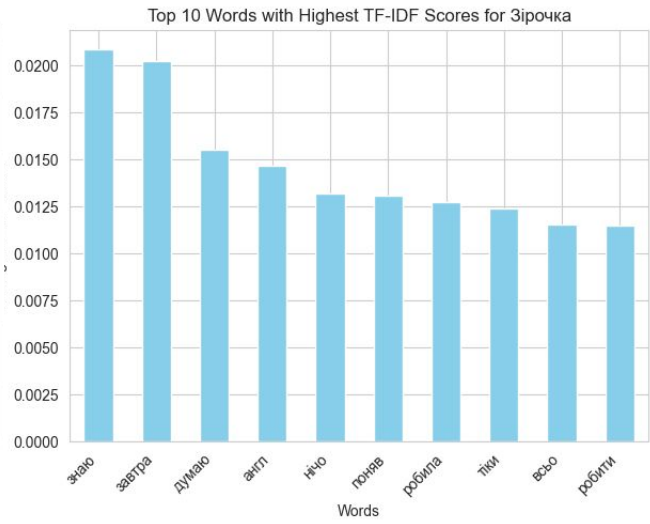
Top emojis of users with whom I interact the most

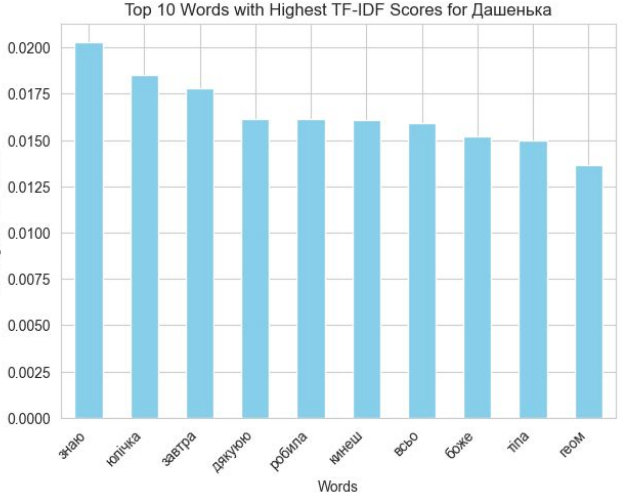
TF-IDF (Term Frequency Inverse Document Frequency) of dialogs

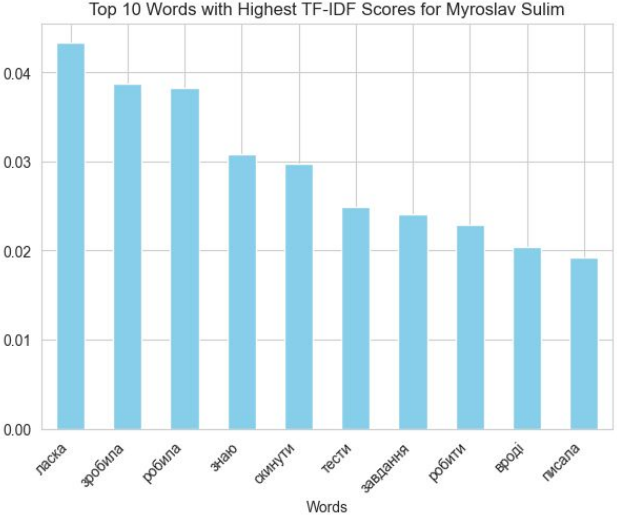
To calculate TF-IDF of words, I splitted dialogs into documents by dates and calculated TF-IDF of words using library ‘sklearn’.

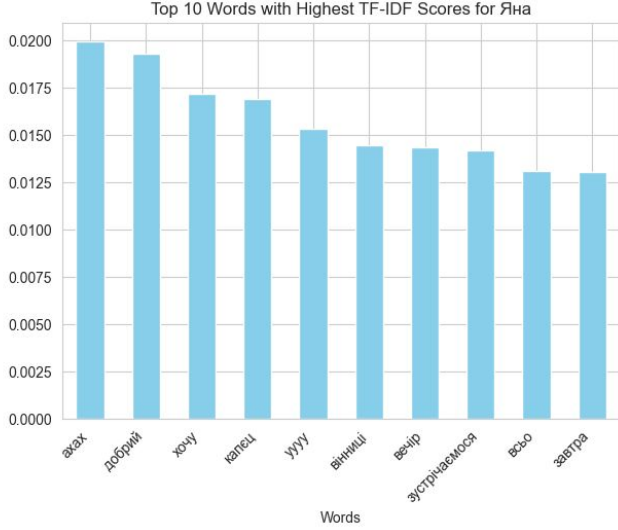

Top 10 Words with Highest TF-IDF Scores for Аня
0.020
0.015
0.010
0.005
0.000
шось
завтра
поняла
знаю
всьо
якшо
робити
робила
хочу
зоря
Words


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

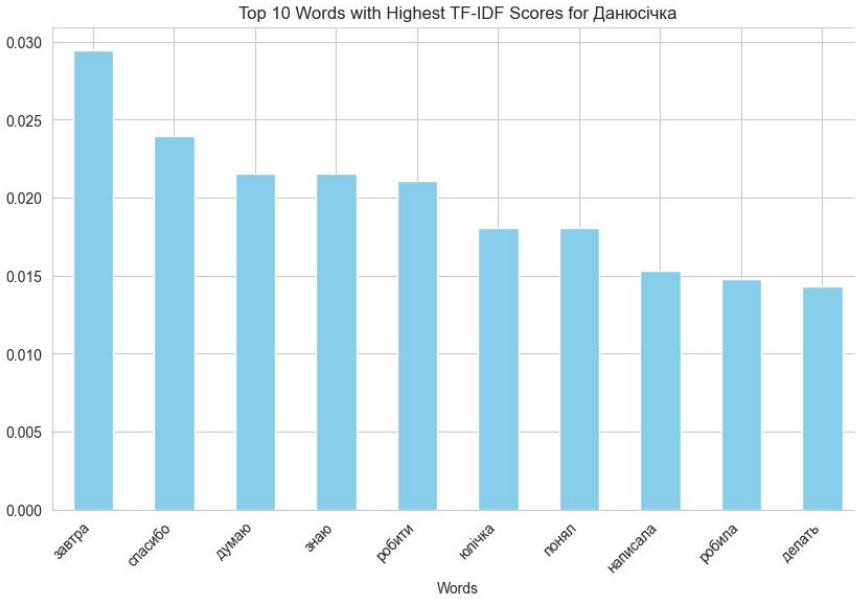
TF-IDF (Data cleaning problem)

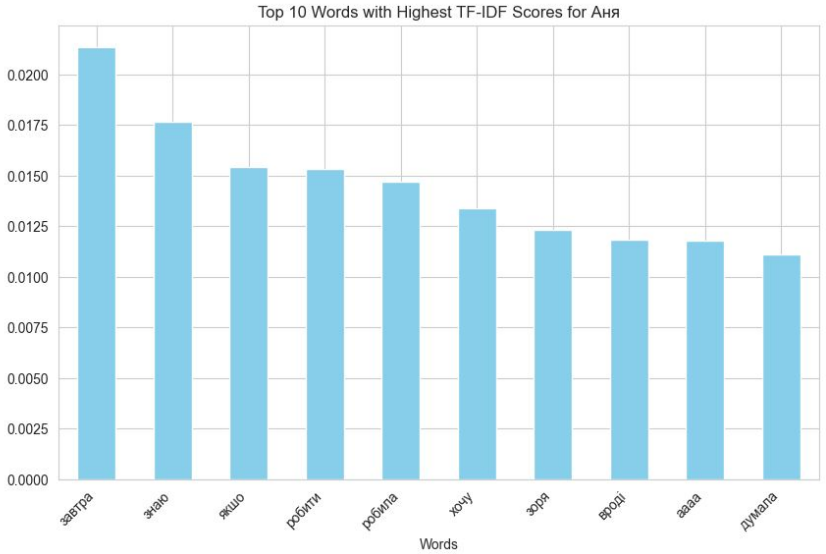
Histograms of words with top tf-idf show us that there are still a lot of words that don’t give meaningful information about dialogs.

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

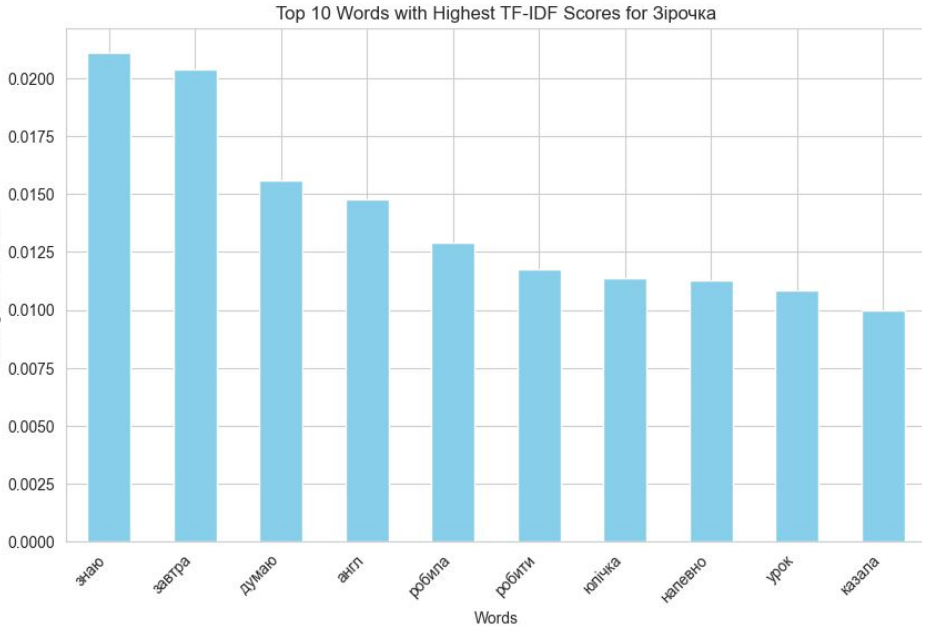
To fix this problem I created new additional dictionary of stop-words.

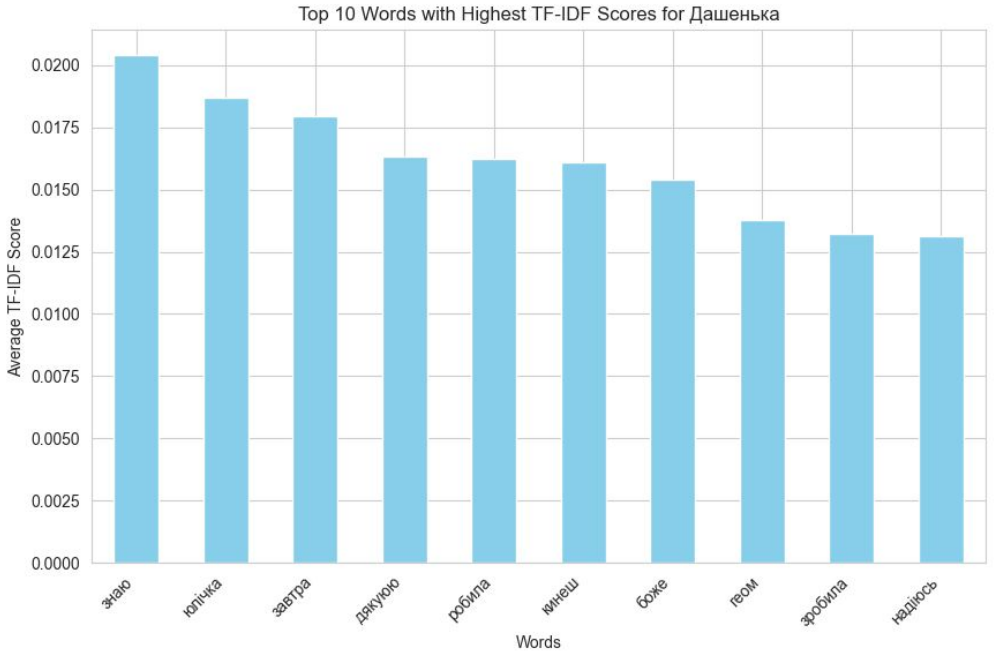
Updated tf-idf



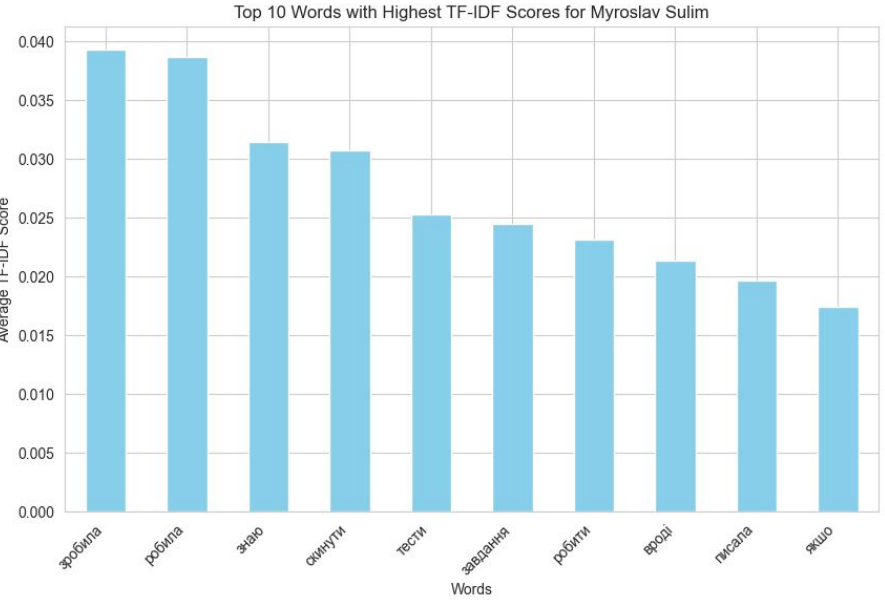


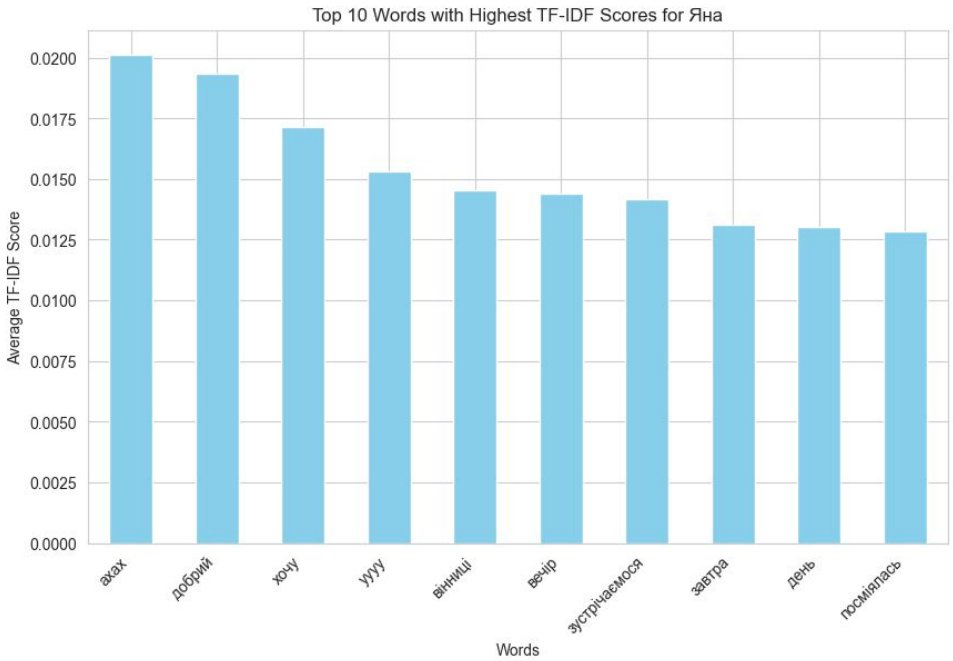
Updated tf-idf





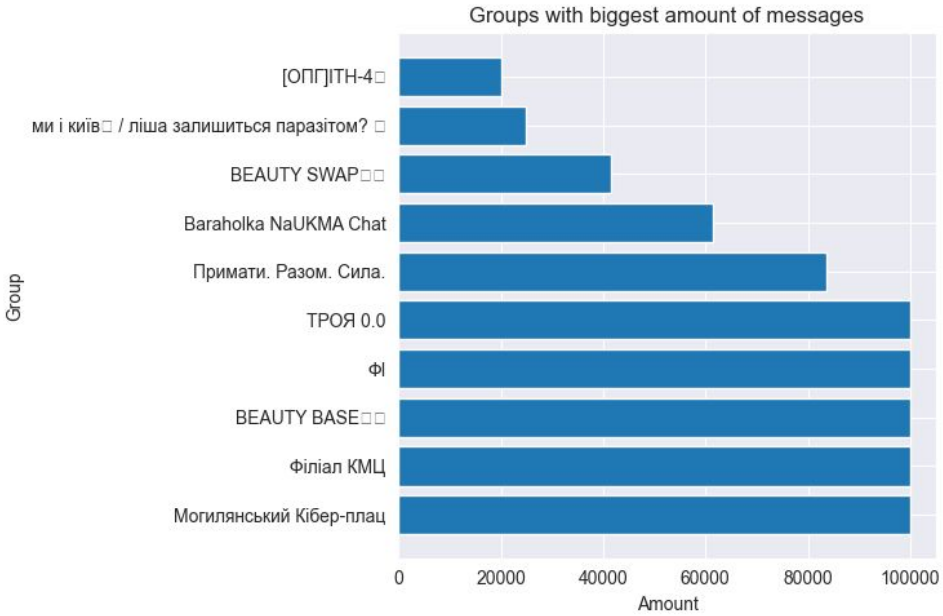
Updated tf-idf

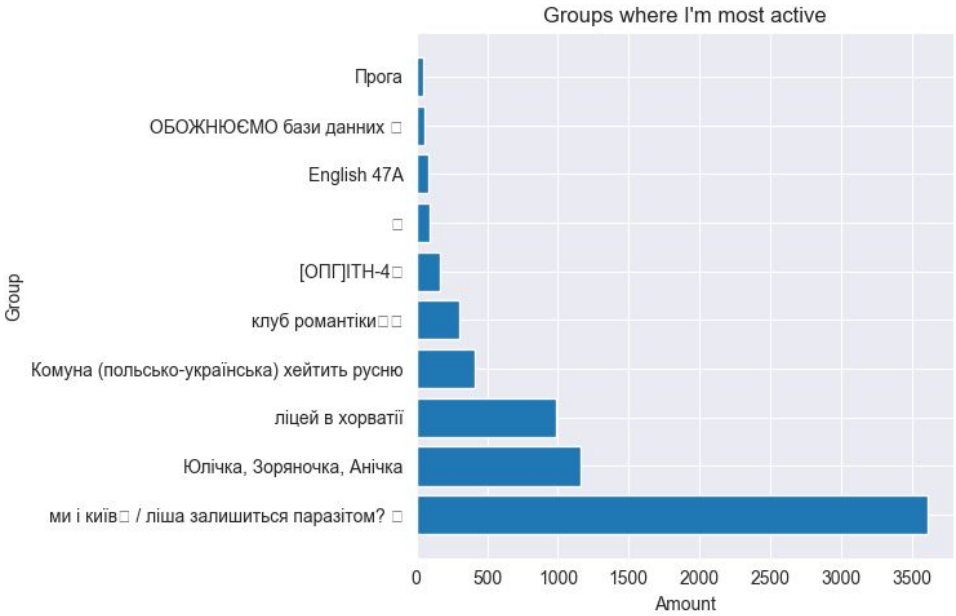




Groups analysis

Groups analysis





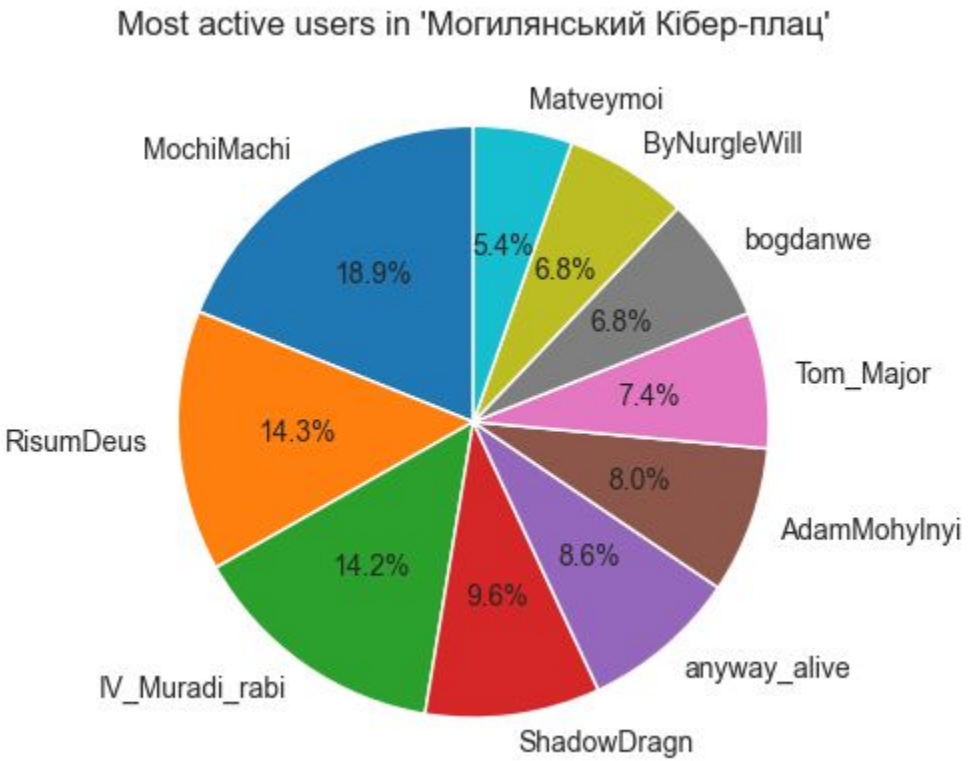
Analysis of Могилянський Кібер-плац

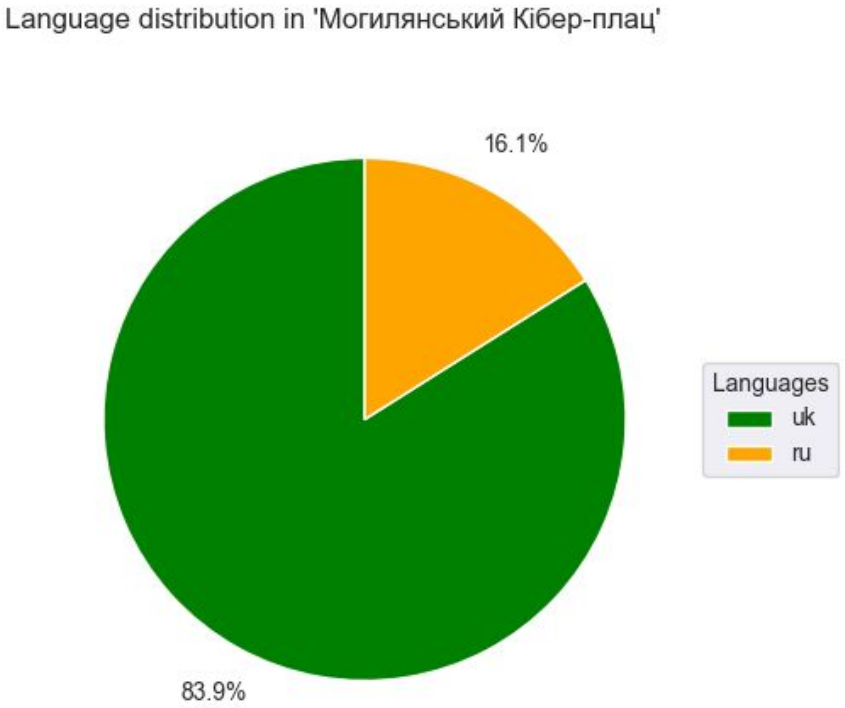
General:

Amount of messages: 100 000

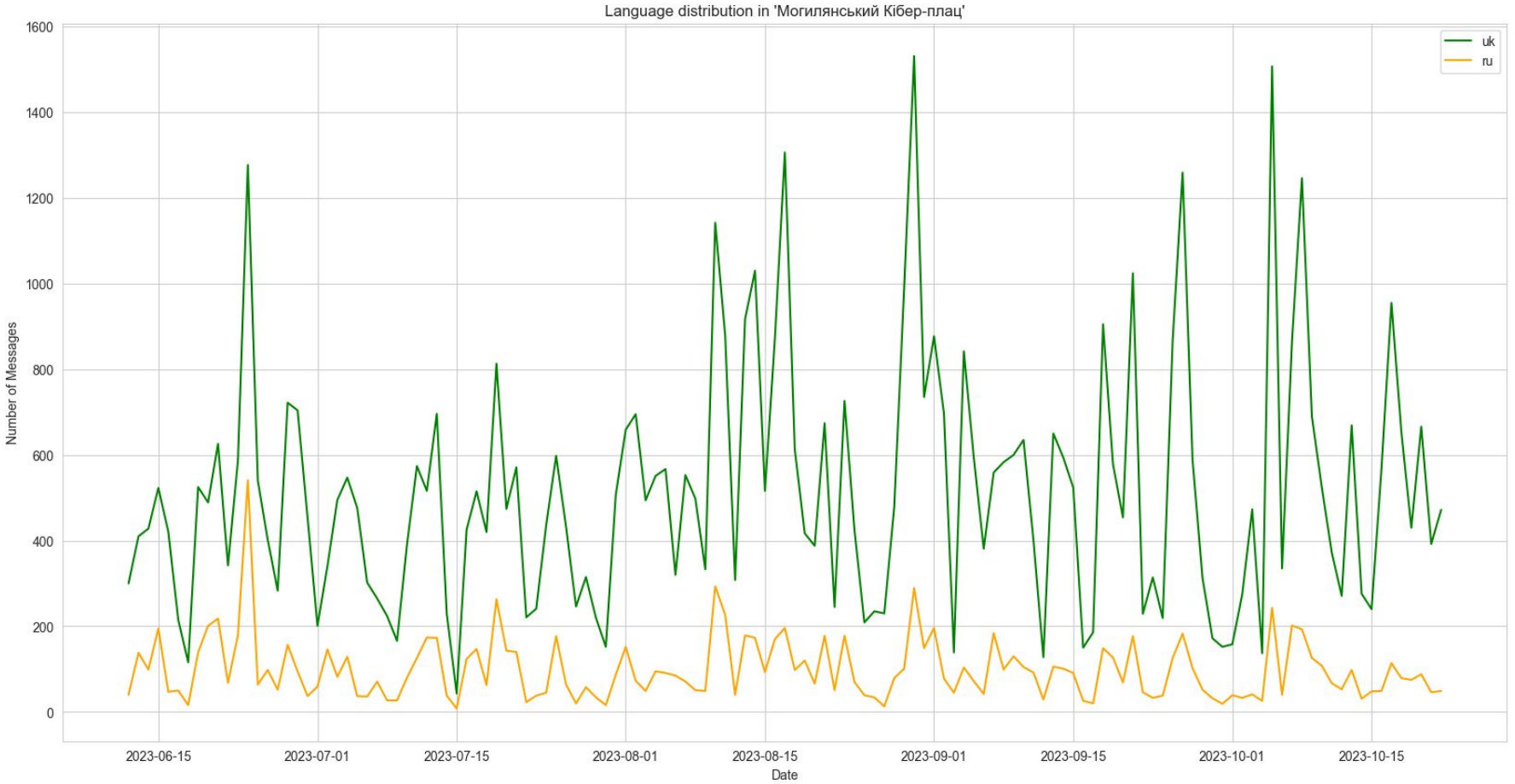
Amount of users: 1613

Analysis of Могилянський Кібер-плац

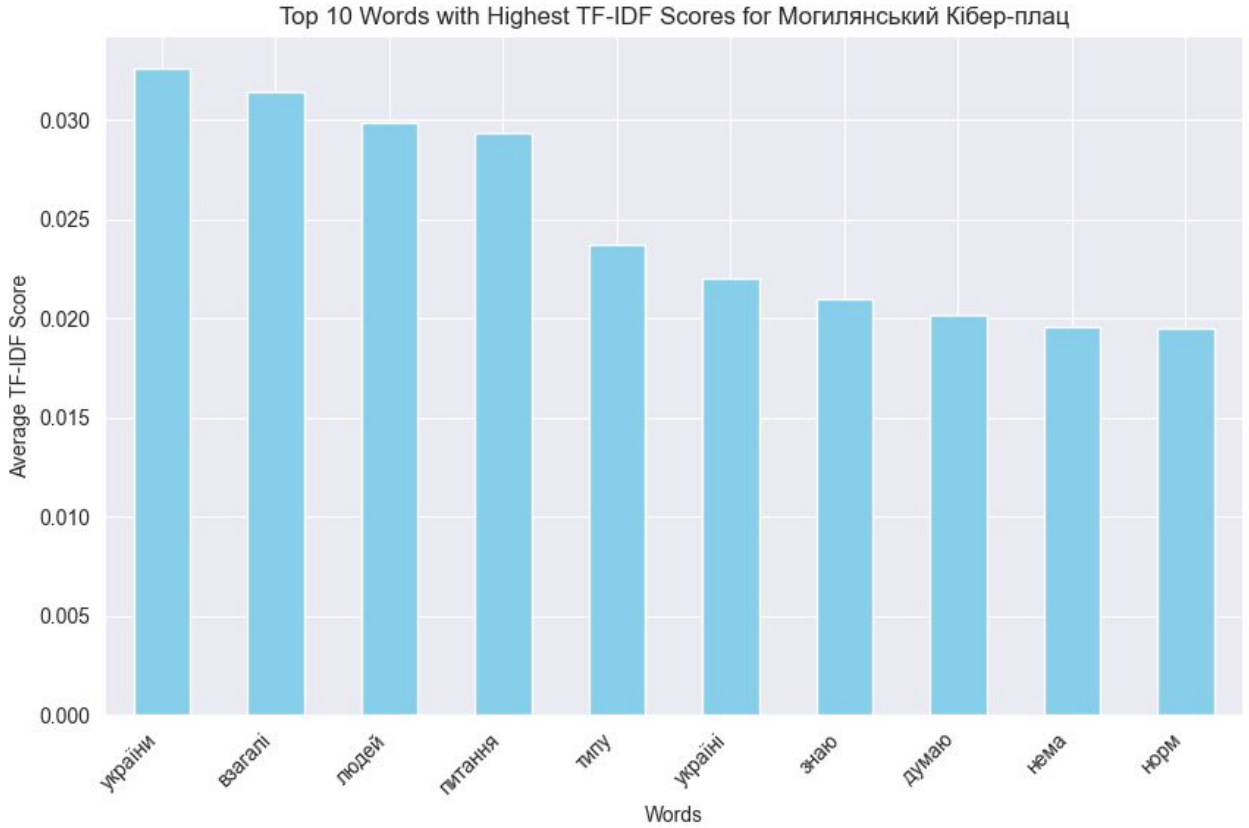




Analysis of Могилянський Кібер-плац

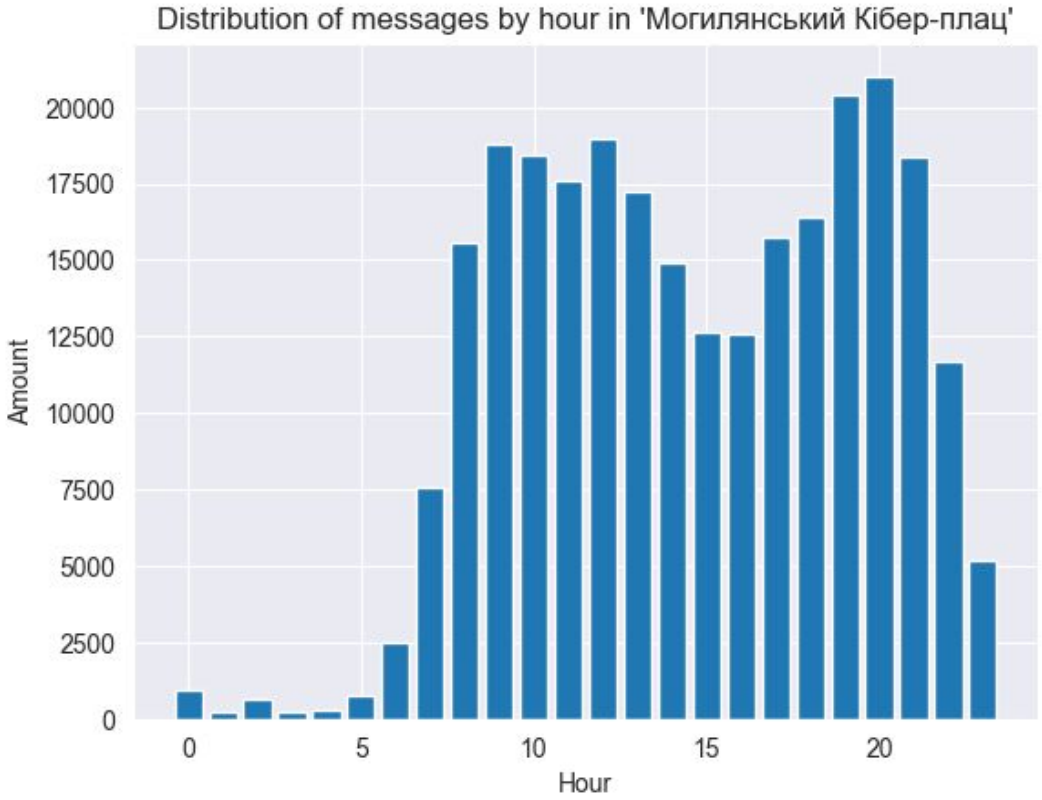


Analysis of Могилянський Кібер-плац



Analysis of Могилянський Кібер-плац

Analysis of Могилянський Кібер-плац



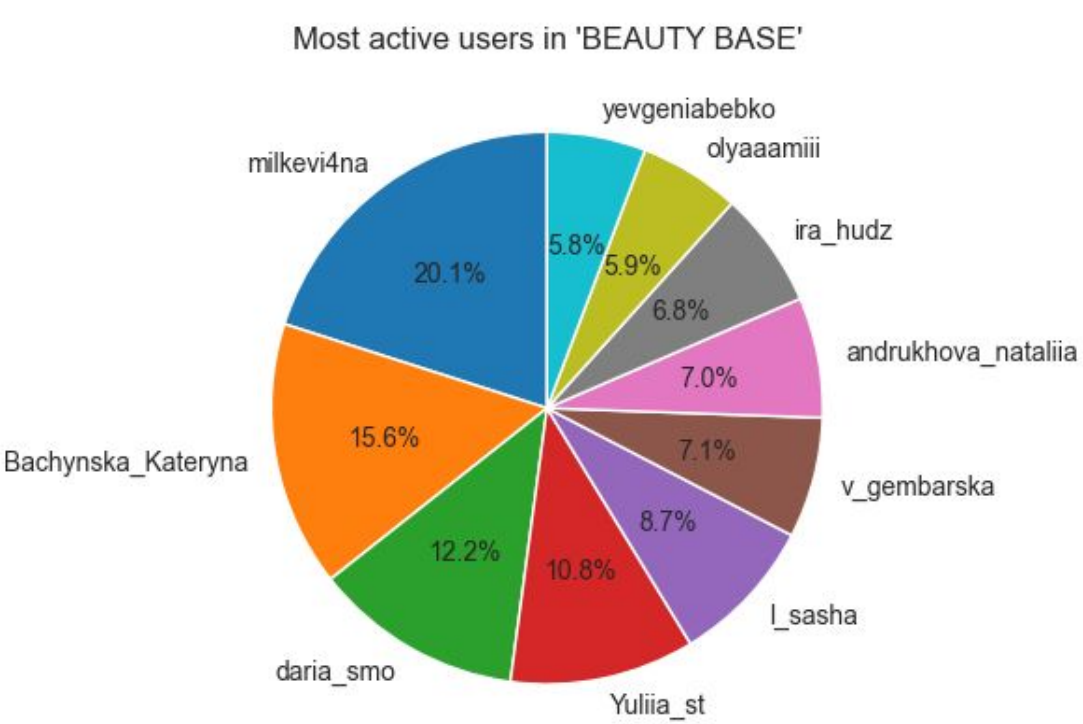
Analysis of BEAUTY BASE 🧴📝

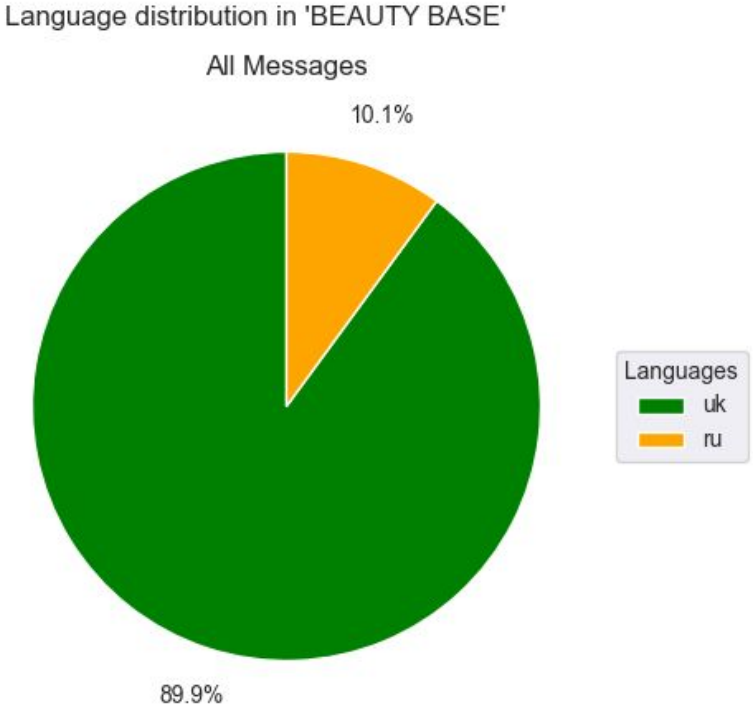
General:

Amount of messages: 100 000

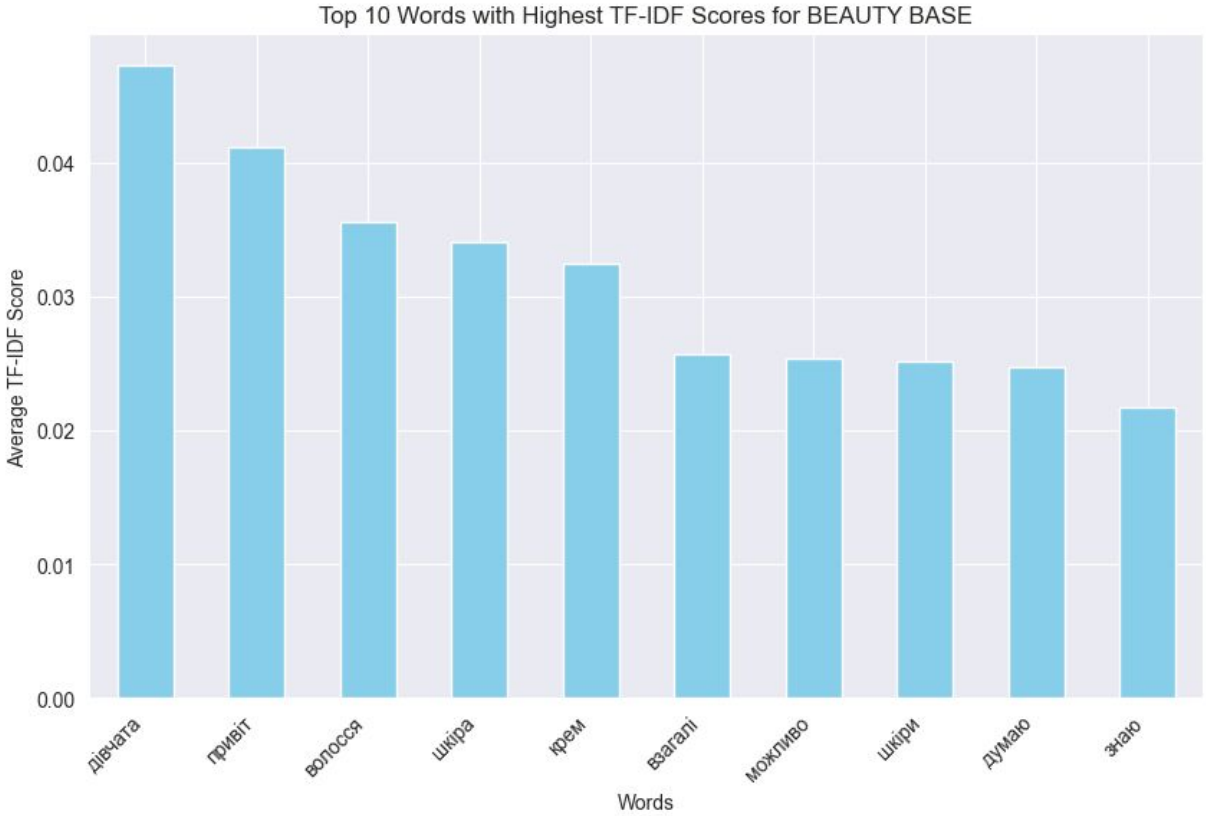
Amount of users: 1808

Analysis of BEAUTY BASE 🧴📝





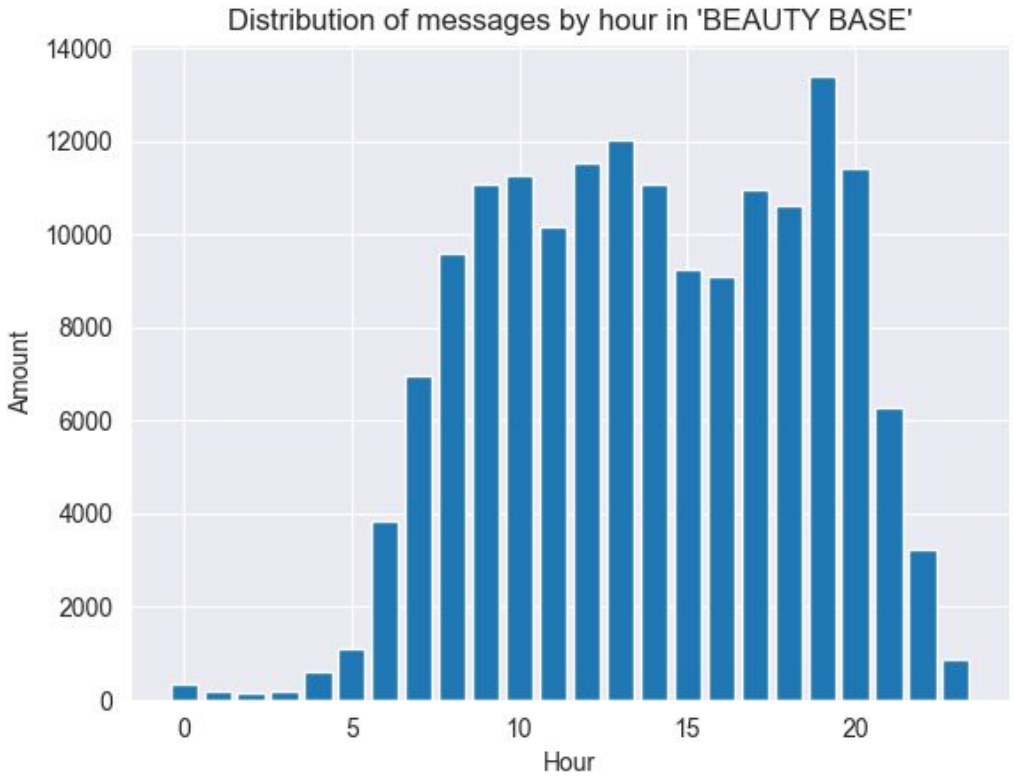
Analysis of BEAUTY BASE 🧴📝



Analysis of BEAUTY BASE 🧴📝

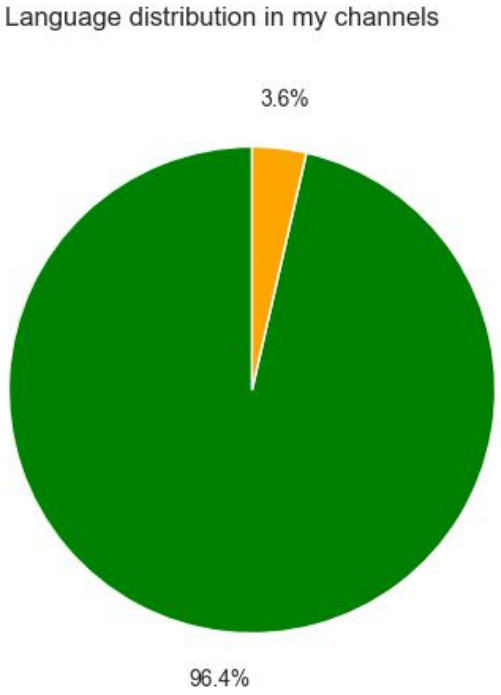
Top-10 emojis of 'BEAUTY BASE'

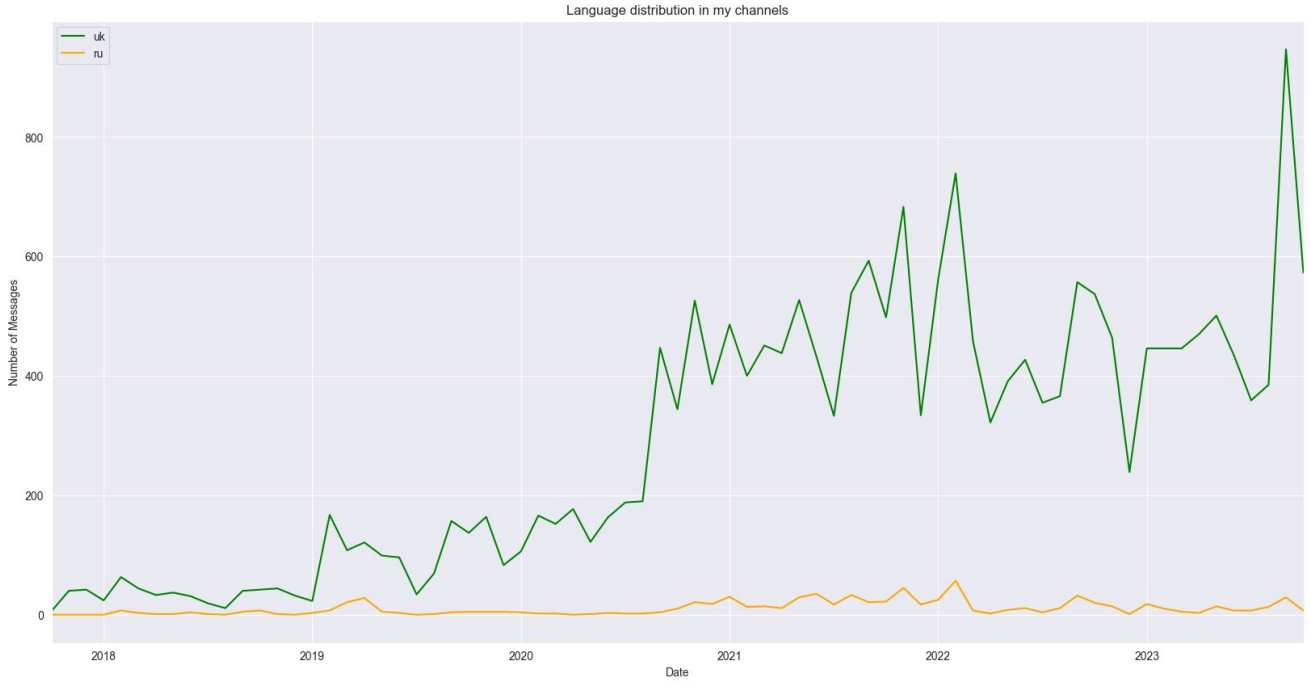
Analysis of BEAUTY BASE 🧴📝



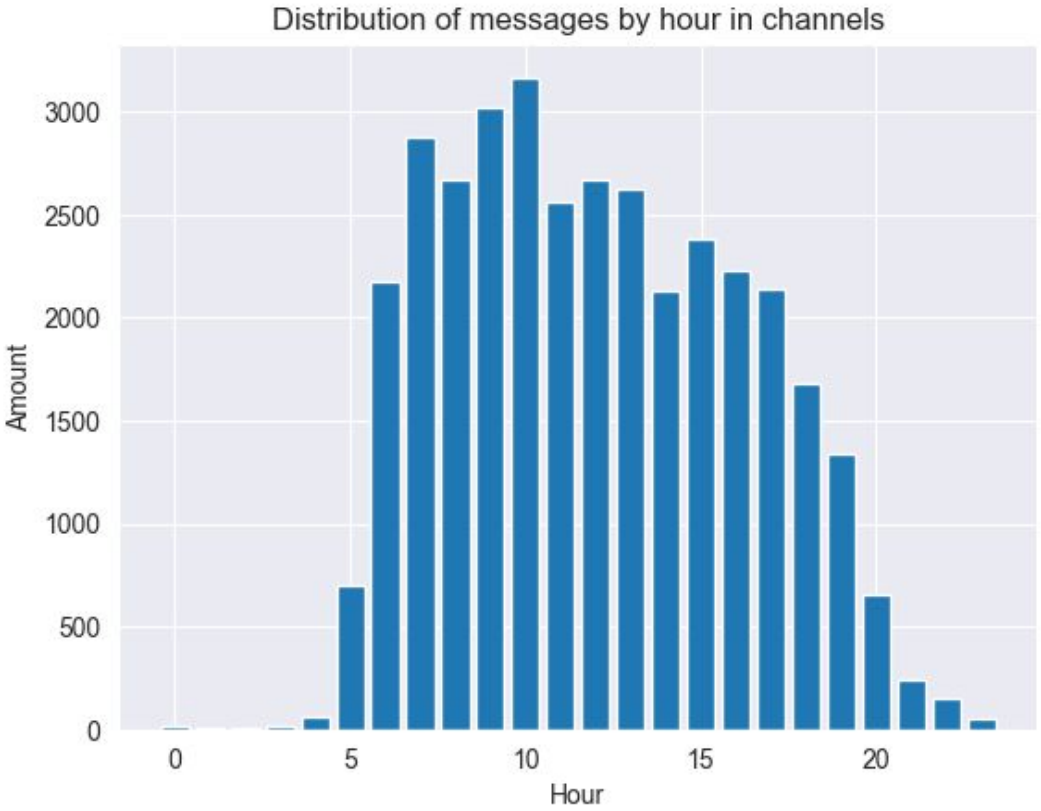
Analysis of channels

Analysis of channels

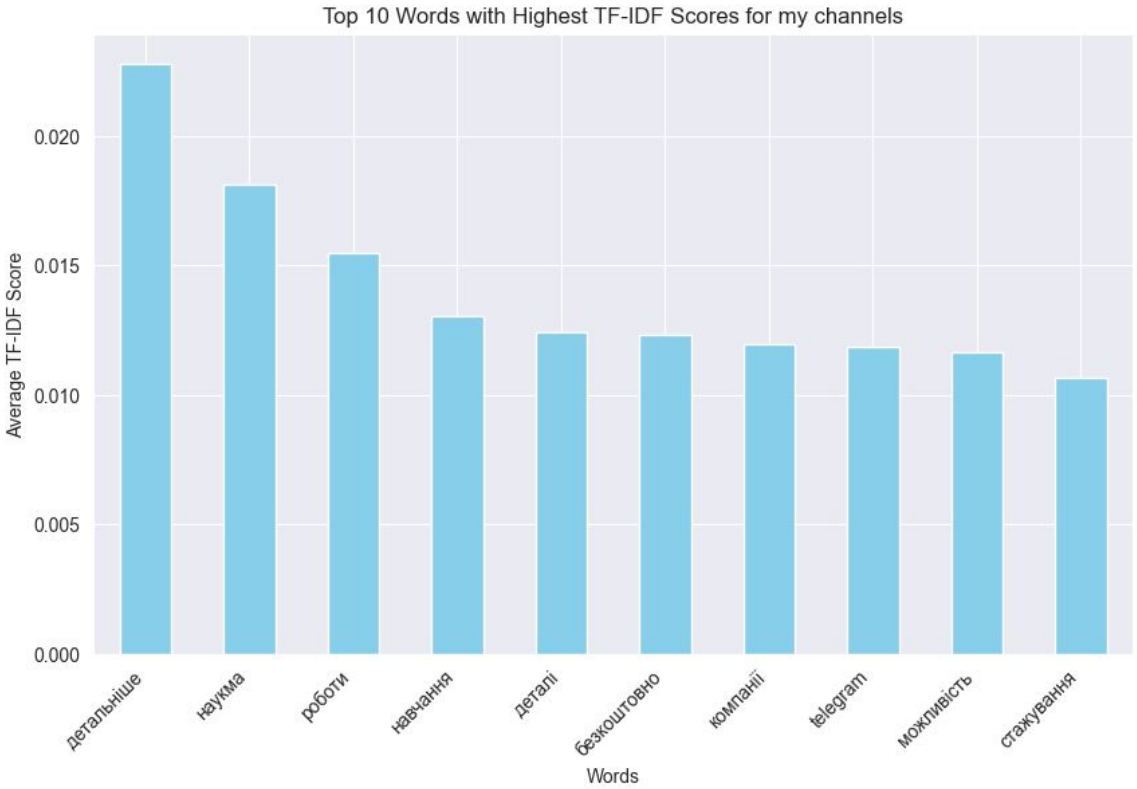




Analysis of channels



Analysis of channels



Further work

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

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

Git Repository