by Anhelina Babii

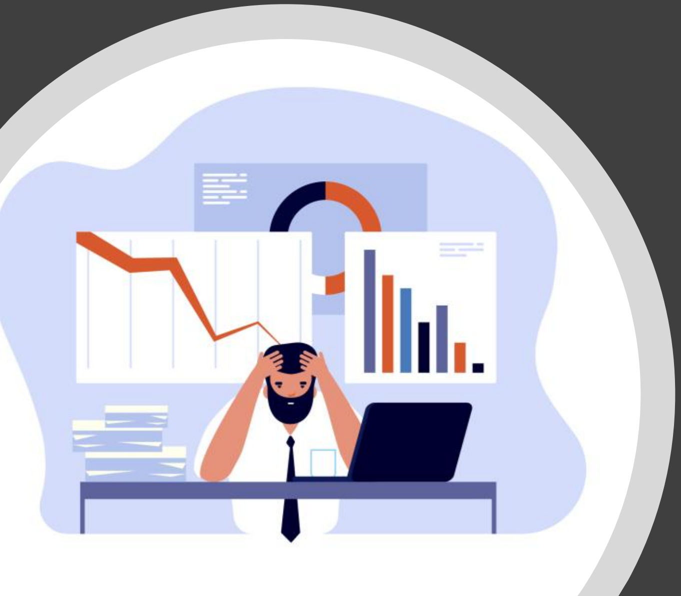
Applied Maths – 2

Computational Social Science

National University "Kyiv-Mohyla Academy"

teacher: Andrew Kurochkin

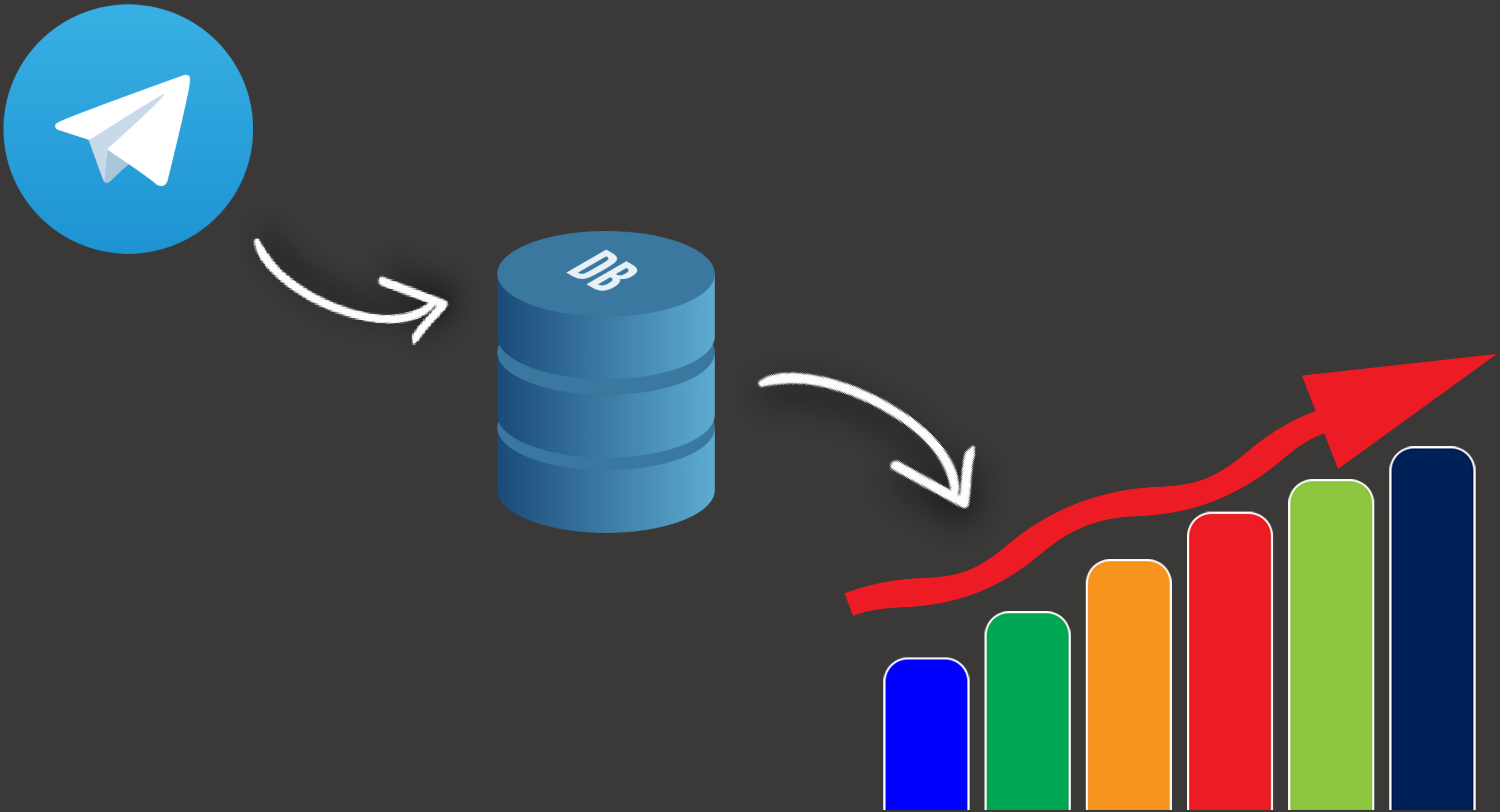
Telegram Data Analysis



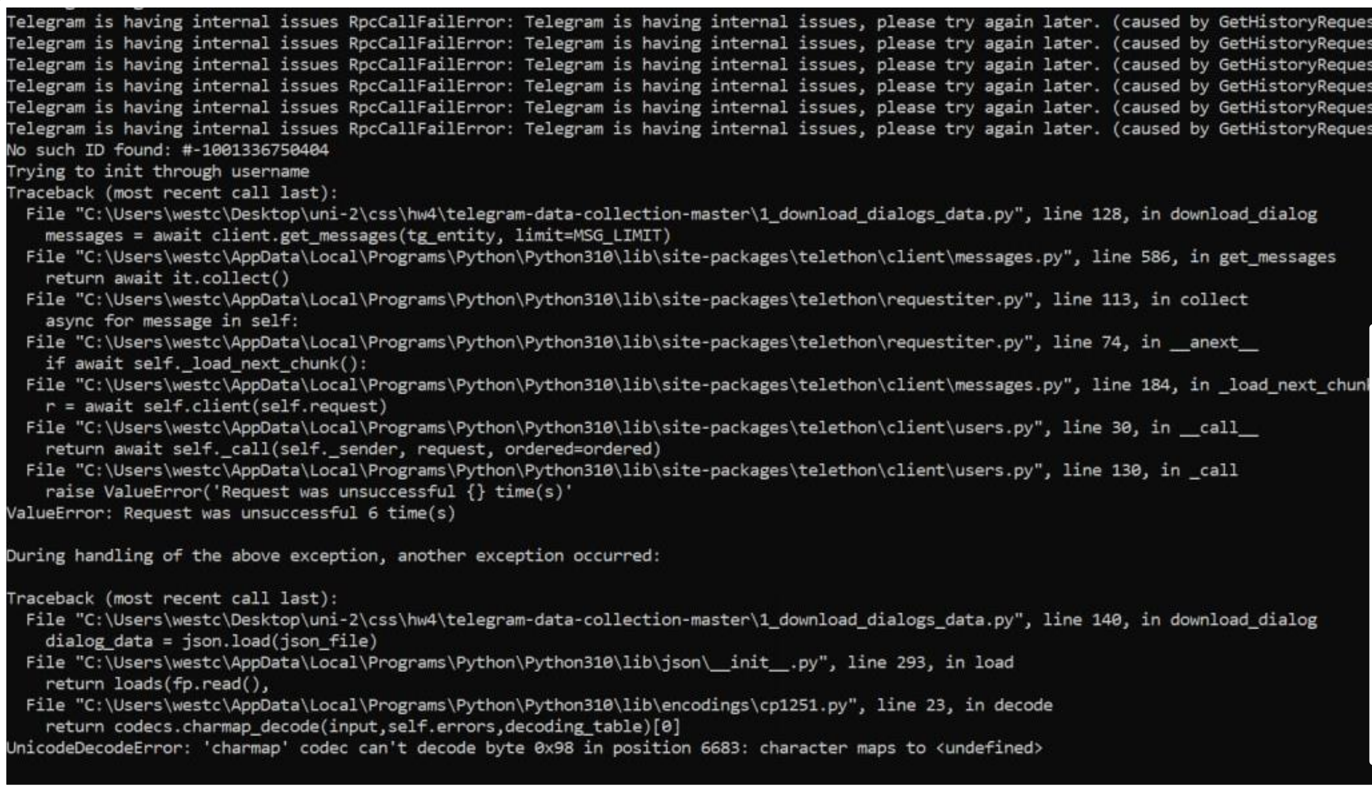
may 8th 2022

The plan of the presentation:

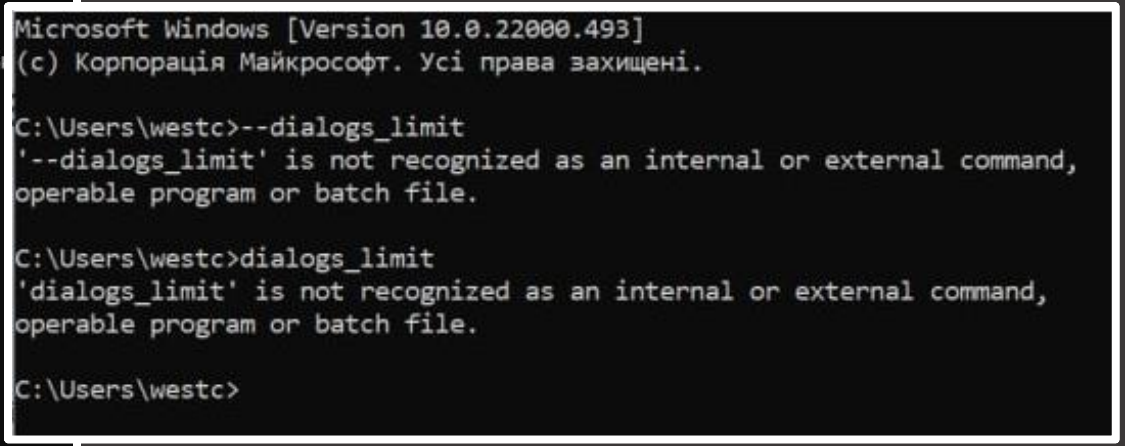
1. Intro
2. Data gathering & info
3. Project flow
4. Best results
5. Further work
6. GitHub link


DB

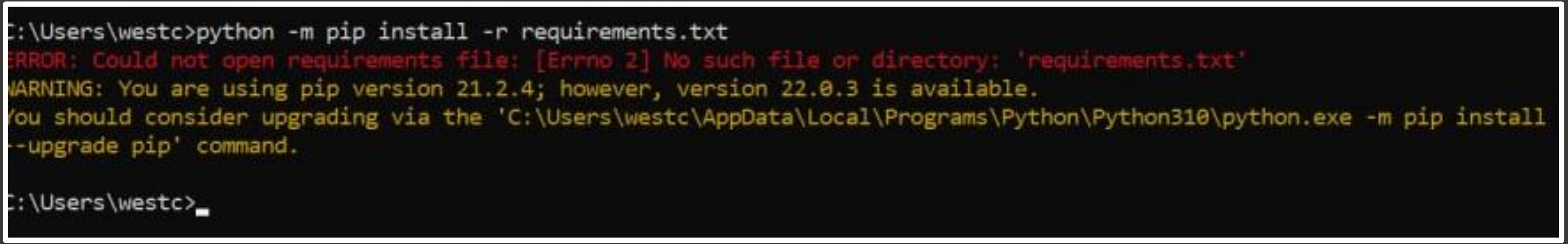
Getting the data



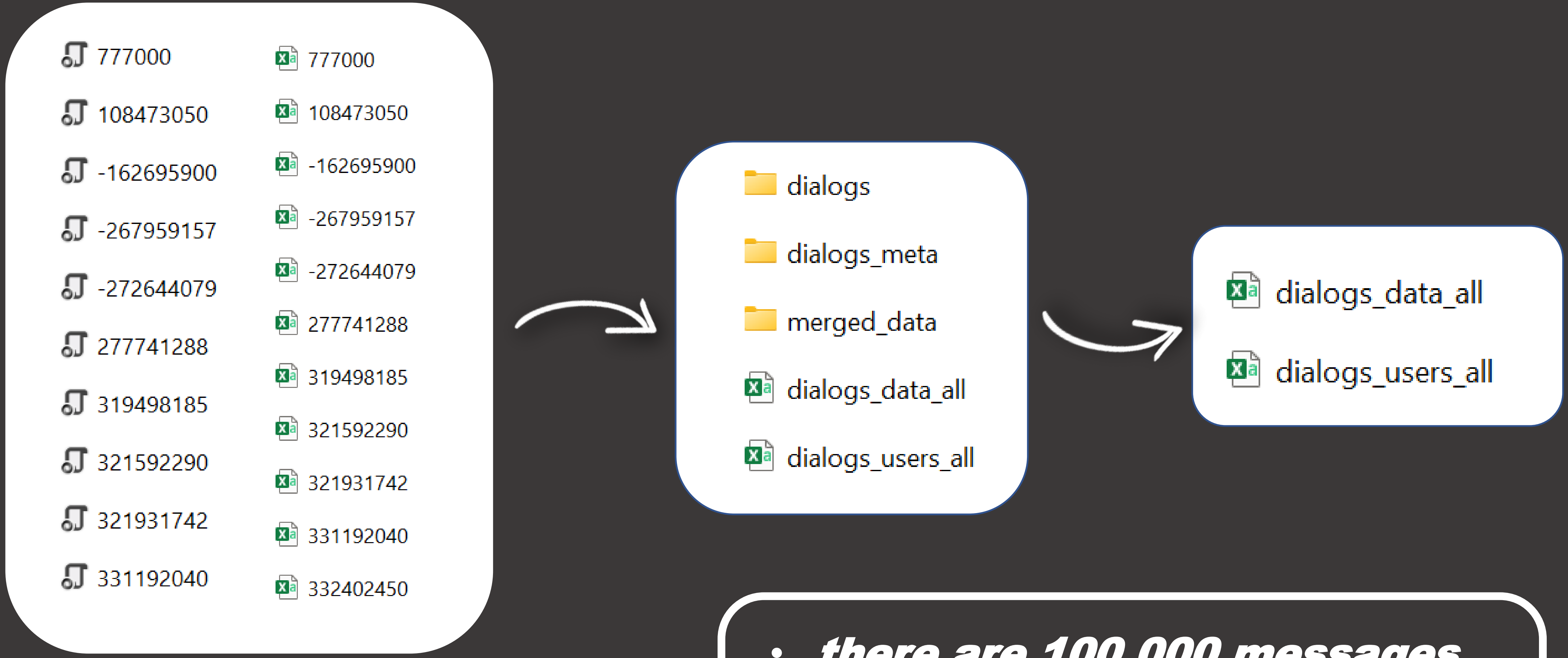
```
Telegram is having internal issues RpcCallFailError: Telegram is having internal issues, please try again later. (caused by GetHistoryRequest)
Telegram is having internal issues RpcCallFailError: Telegram is having internal issues, please try again later. (caused by GetHistoryRequest)
Telegram is having internal issues RpcCallFailError: Telegram is having internal issues, please try again later. (caused by GetHistoryRequest)
Telegram is having internal issues RpcCallFailError: Telegram is having internal issues, please try again later. (caused by GetHistoryRequest)
Telegram is having internal issues RpcCallFailError: Telegram is having internal issues, please try again later. (caused by GetHistoryRequest)
Telegram is having internal issues RpcCallFailError: Telegram is having internal issues, please try again later. (caused by GetHistoryRequest)
No such ID found: #-1001336750404
Trying to init through username
Traceback (most recent call last):
  File "C:\Users\westc\Desktop\uni-2\css\hw4\telegram-data-collection-master\1_download_dialogs_data.py", line 128, in download_dialog
    messages = await client.get_messages(tg_entity, limit=MSG_LIMIT)
  File "C:\Users\westc\AppData\Local\Programs\Python\Python310\lib\site-packages\telethon\client\messages.py", line 586, in get_messages
    return await it.collect()
  File "C:\Users\westc\AppData\Local\Programs\Python\Python310\lib\site-packages\telethon\requestiter.py", line 113, in collect
    async for message in self:
  File "C:\Users\westc\AppData\Local\Programs\Python\Python310\lib\site-packages\telethon\requestiter.py", line 74, in __anext__
    if await self._load_next_chunk():
  File "C:\Users\westc\AppData\Local\Programs\Python\Python310\lib\site-packages\telethon\client\messages.py", line 184, in _load_next_chun
    r = await self.client(self.request)
  File "C:\Users\westc\AppData\Local\Programs\Python\Python310\lib\site-packages\telethon\client\users.py", line 30, in __call__
    return await self._call(self._sender, request, ordered=ordered)
  File "C:\Users\westc\AppData\Local\Programs\Python\Python310\lib\site-packages\telethon\client\users.py", line 130, in _call
    raise ValueError('Request was unsuccessful {} time(s)'
ValueError: Request was unsuccessful 6 time(s)

During handling of the above exception, another exception occurred:

Traceback (most recent call last):
  File "C:\Users\westc\Desktop\uni-2\css\hw4\telegram-data-collection-master\1_download_dialogs_data.py", line 140, in download_dialog
    dialog_data = json.load(json_file)
  File "C:\Users\westc\AppData\Local\Programs\Python\Python310\lib\json\__init__.py", line 293, in load
    return loads(fp.read(),
  File "C:\Users\westc\AppData\Local\Programs\Python\Python310\lib\encodings\cp1251.py", line 23, in decode
    return codecs.charmap_decode(input,self.errors,decoding_table)[0]
UnicodeDecodeError: 'charmap' codec can't decode byte 0x98 in position 6683: character maps to <undefined>
```



```
Microsoft Windows [Version 10.0.22000.493]
(c) Корпорація Майкрософт. Усі права захищені.

C:\Users\westc>--dialogs_limit
'--dialogs_limit' is not recognized as an internal or external command,
operable program or batch file.

C:\Users\westc>dialogs_limit
'dialogs_limit' is not recognized as an internal or external command,
operable program or batch file.

C:\Users\westc>
```



```
:\Users\westc>python -m pip install -r requirements.txt
RROR: Could not open requirements file: [Errno 2] No such file or directory: 'requirements.txt'
ARNING: You are using pip version 21.2.4; however, version 22.0.3 is available.
ou should consider upgrading via the 'C:\Users\westc\AppData\Local\Programs\Python\Python310\python.exe -m pip install
-upgrade pip' command.

:\Users\westc>_
```



777000
108473050
-162695900
-267959157
-272644079
277741288
319498185
321592290
321931742
331192040

777000
108473050
-162695900
-267959157
-272644079
277741288
319498185
321592290
321931742
331192040
332402450

dialogs
dialogs_meta
merged_data
dialogs_data_all
dialogs_users_all

dialogs_data_all
dialogs_users_all

- ***there are 100 000 messages downloaded to the dataset***

- ***The merged dataset is 176 MB.***

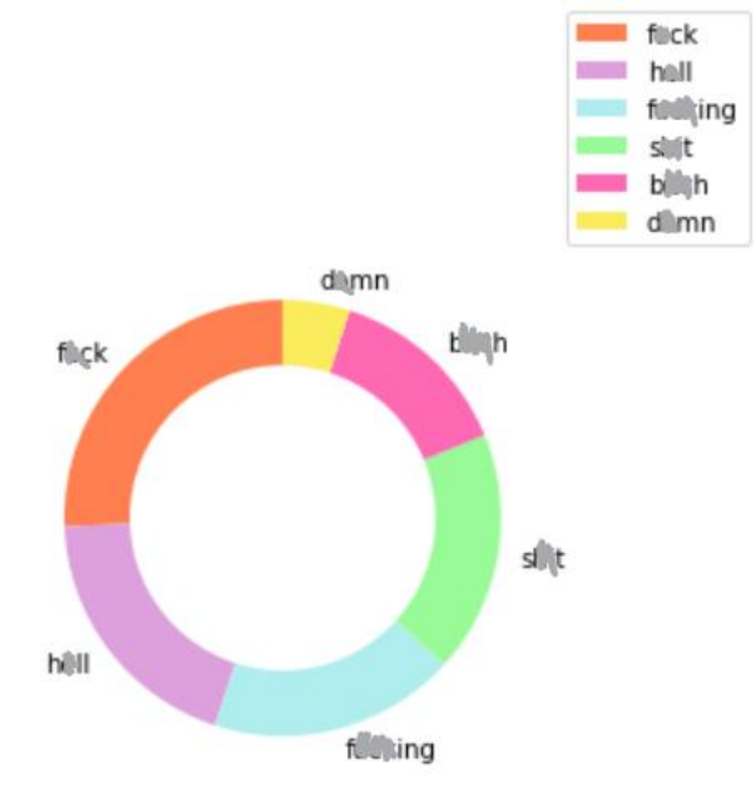
How it all went

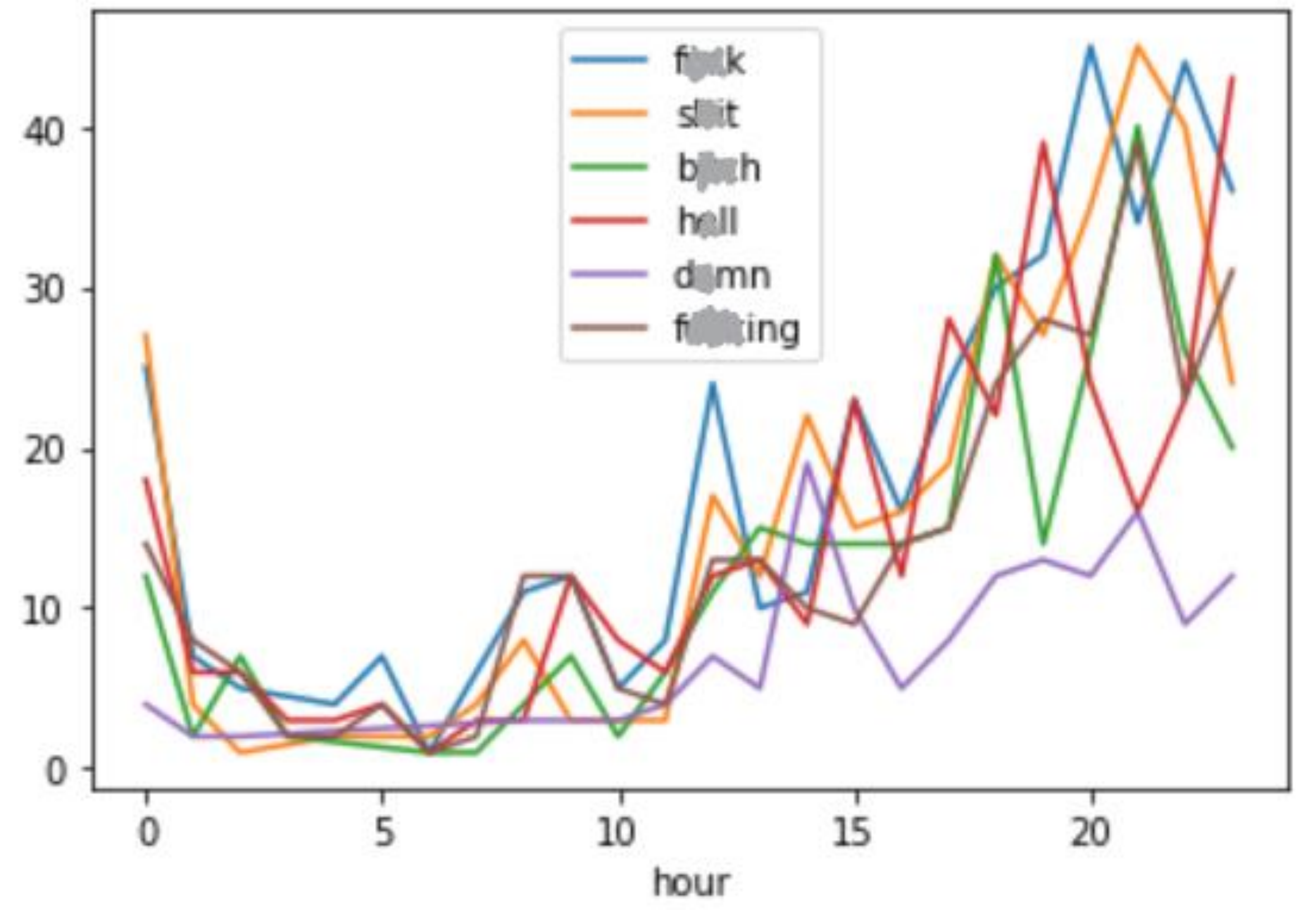
Some of the plots I created and WHY

1. *General stats plots (top-ns, word count, sorting etc)*
2. *Link plot*
3. *Swear word plot*
4. *Days in a week plot*

Some of the plots I never used and WHY

1. *Gender diversity plot*
2. *Languages plot*
3. *Activity/day in a month*

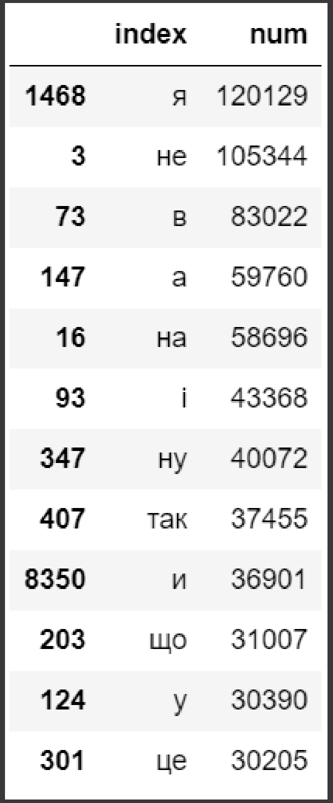

f[redacted]ck
h[redacted]ll
f[redacted]ing
s[redacted]t
b[redacted]h
d[redacted]mn
d[redacted]mn
f[redacted]ck
b[redacted]h
s[redacted]t
h[redacted]ll
f[redacted]ing


f[redacted]k
s[redacted]t
b[redacted]h
h[redacted]ll
d[redacted]mn
f[redacted]ing
40
30
20
10
0
0
5
10
15
20
hour

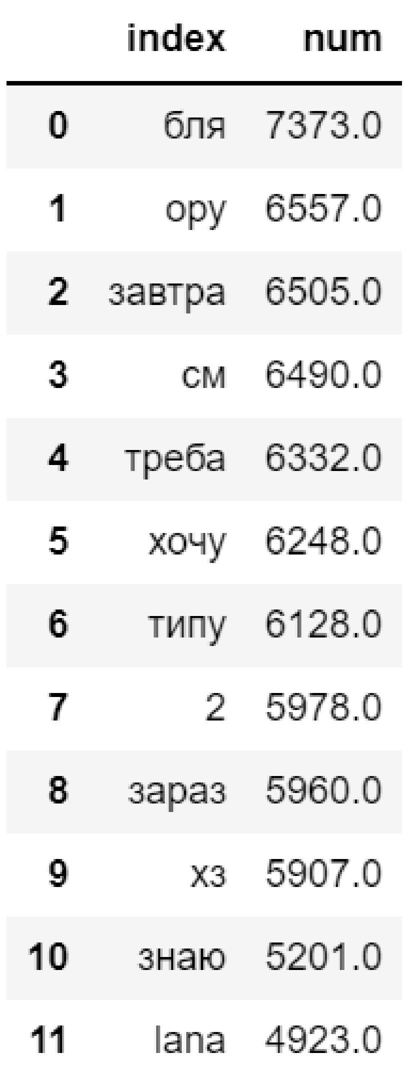
5 best research plots

by me

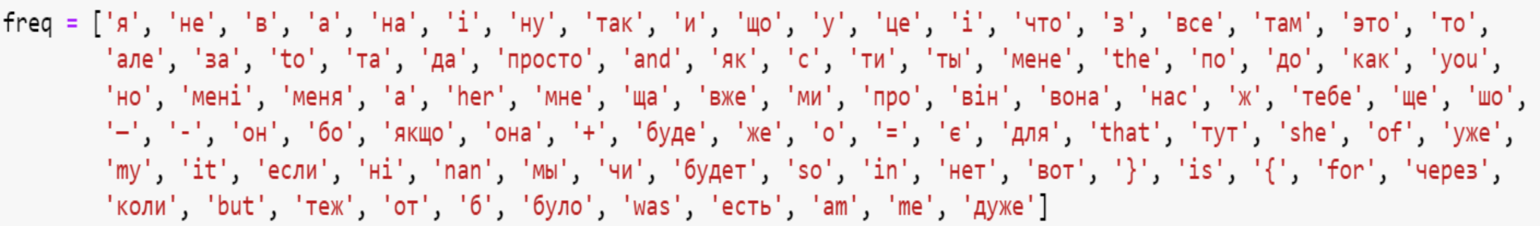
1. Top-12 words I used + improved



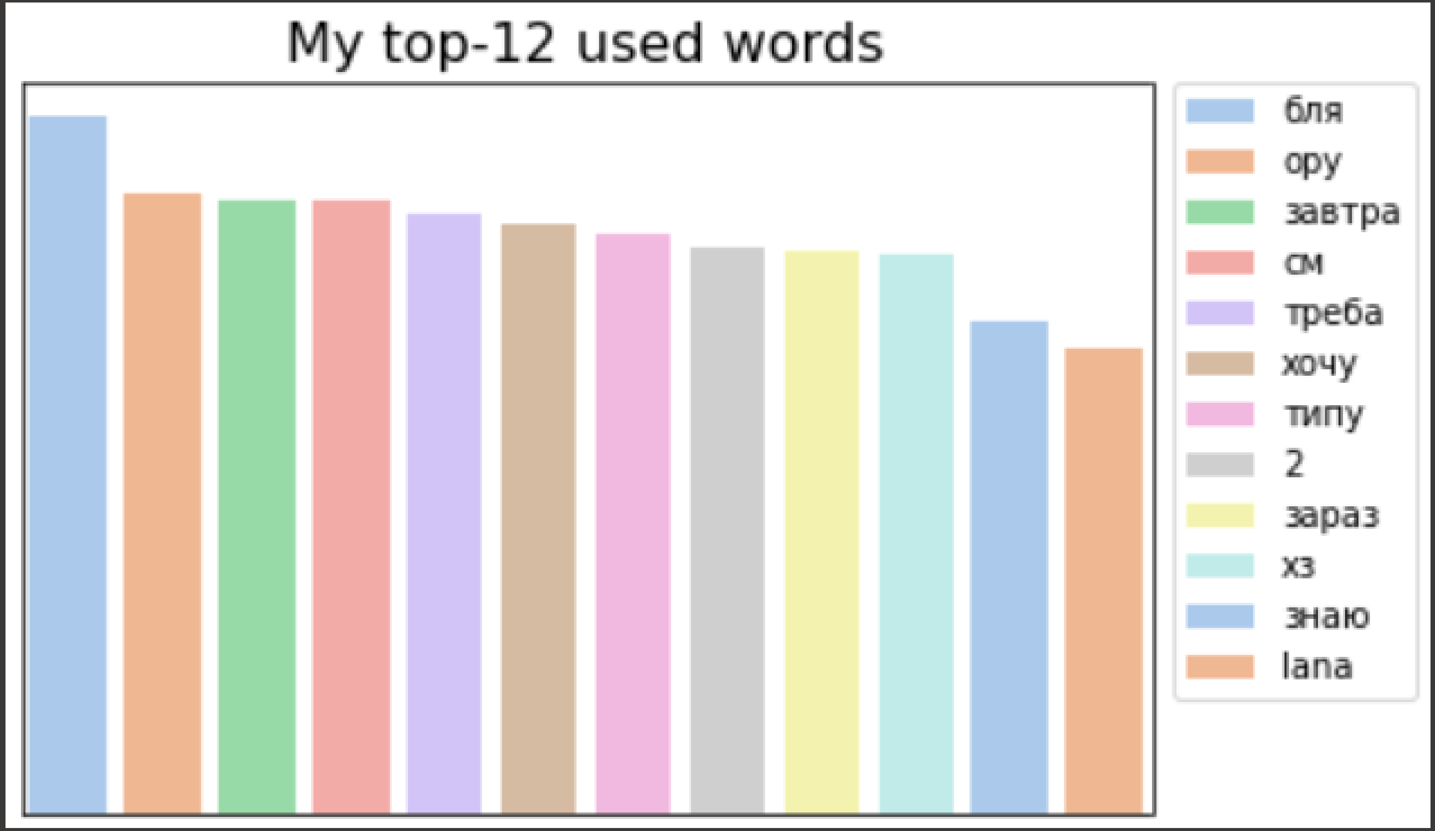
	index	num
1468	я	120129
3	не	105344
73	в	83022
147	а	59760
16	на	58696
93	і	43368
347	ну	40072
407	так	37455
8350	и	36901
203	що	31007
124	у	30390
301	це	30205



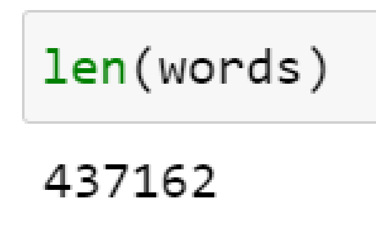
	index	num
0	бля	7373.0
1	ору	6557.0
2	завтра	6505.0
3	см	6490.0
4	треба	6332.0
5	хочу	6248.0
6	типу	6128.0
7	2	5978.0
8	зараз	5960.0
9	хз	5907.0
10	знаю	5201.0
11	lana	4923.0



```
freq = ['я', 'не', 'в', 'а', 'на', 'і', 'ну', 'так', 'и', 'що', 'у', 'це', 'і', 'что', 'з', 'все', 'там', 'это', 'то',
        'але', 'за', 'to', 'та', 'да', 'просто', 'and', 'як', 'с', 'ти', 'ты', 'мене', 'the', 'по', 'до', 'как', 'you',
        'но', 'мені', 'меня', 'a', 'her', 'мне', 'ща', 'вже', 'ми', 'про', 'він', 'вона', 'нас', 'ж', 'тебе', 'ще', 'шо',
        '—', '-', 'он', 'бо', 'якщо', 'она', '+', 'буде', 'же', 'о', '=', 'є', 'для', 'that', 'тут', 'she', 'of', 'уже',
        'my', 'it', 'если', 'ні', 'nan', 'мы', 'чи', 'будет', 'so', 'in', 'нет', 'вот', '}', 'is', '{', 'for', 'через',
        'коли', 'but', 'теж', 'от', 'б', 'було', 'was', 'есть', 'am', 'me', 'дуже']
```



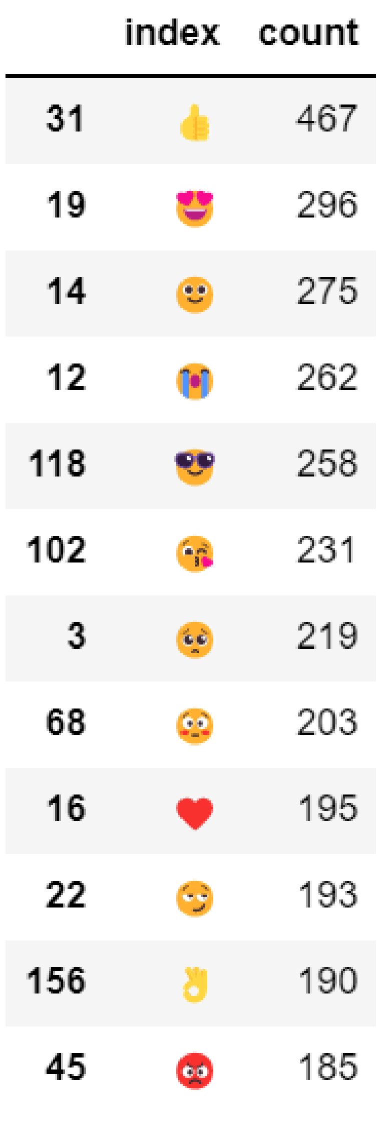
unique



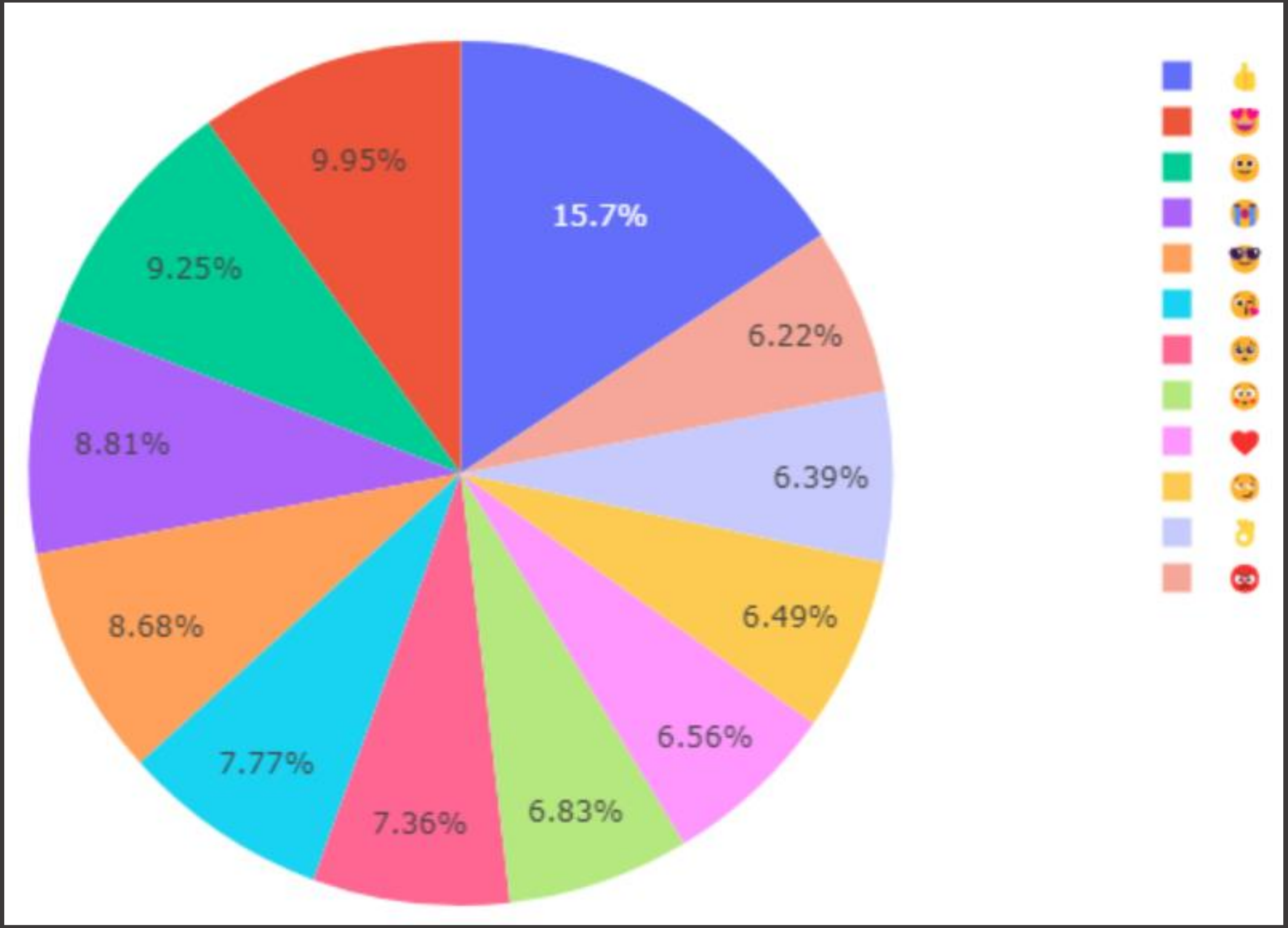
```
len(words)
```

```
437162
```

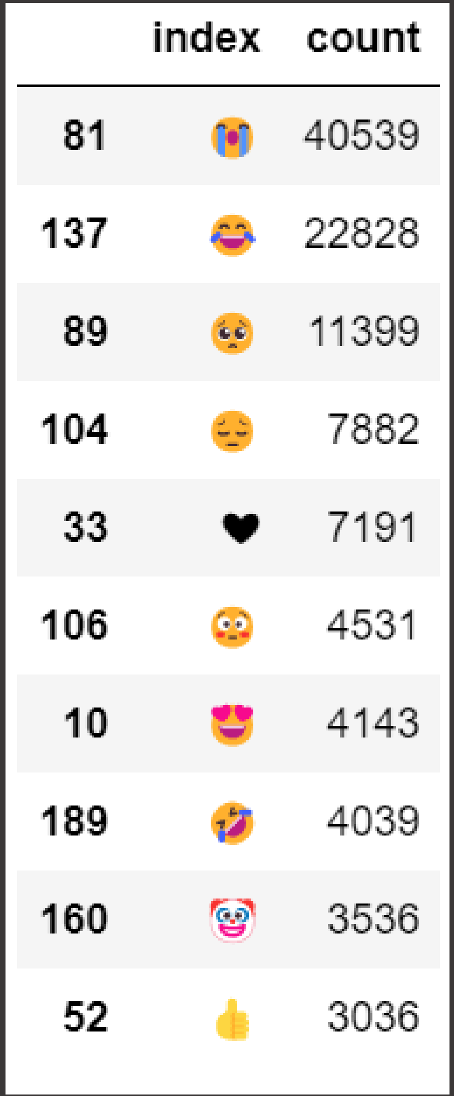
Did the same with stickers:



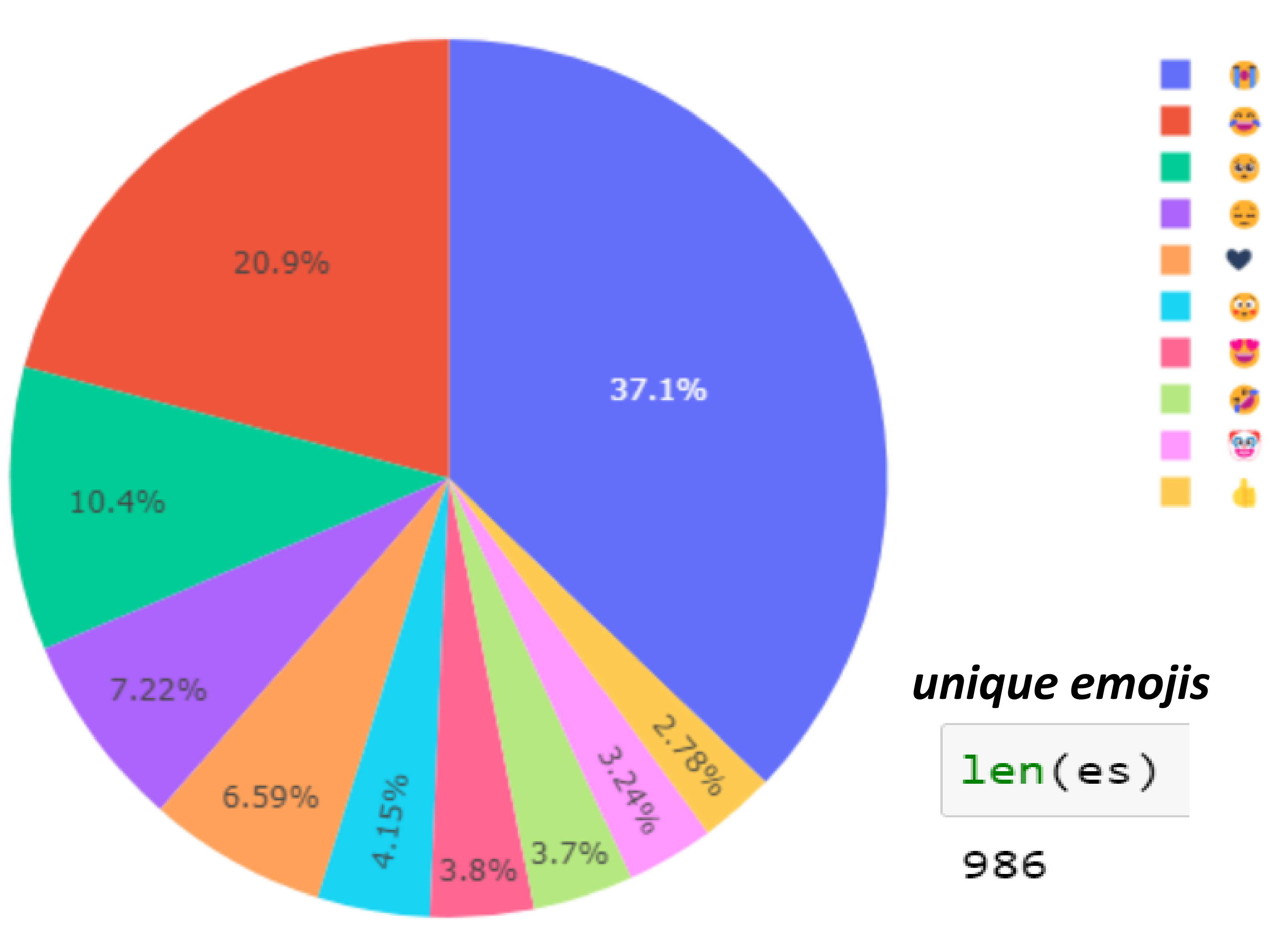
	index	count
31	👍	467
19	😍	296
14	🙂	275
12	😭	262
118	😎	258
102	😘	231
3	🥺	219
68	😳	203
16	❤️	195
22	🙄	193
156	👌	190
45	😡	185



and emojis:



	index	count
81	😭	40539
137	😂	22828
89	🥺	11399
104	😔	7882
33	🖤	7191
106	😳	4531
10	😍	4143
189	🤣	4039
160	🤡	3536
52	👍	3036

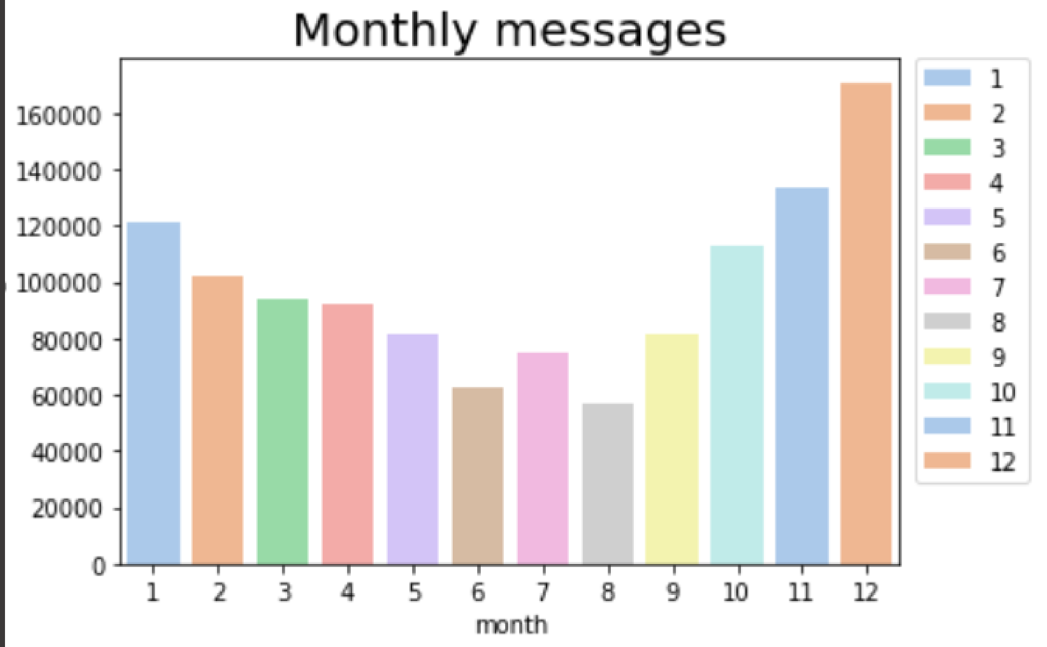


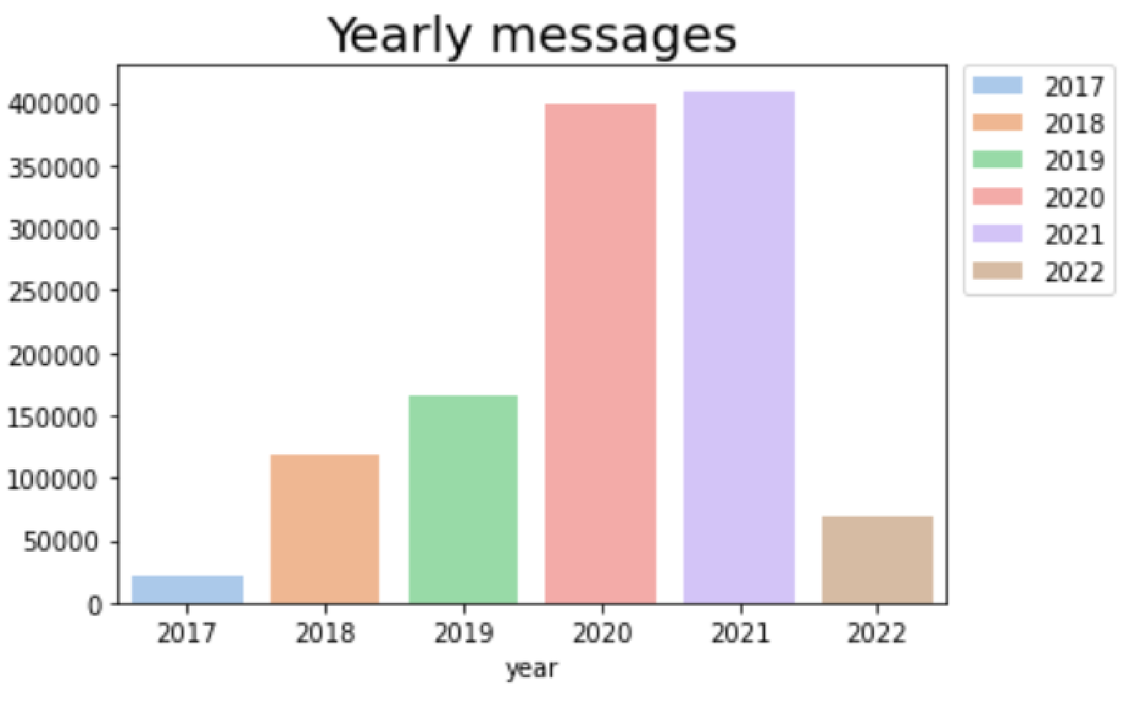
unique emojis

```
len(es)
```

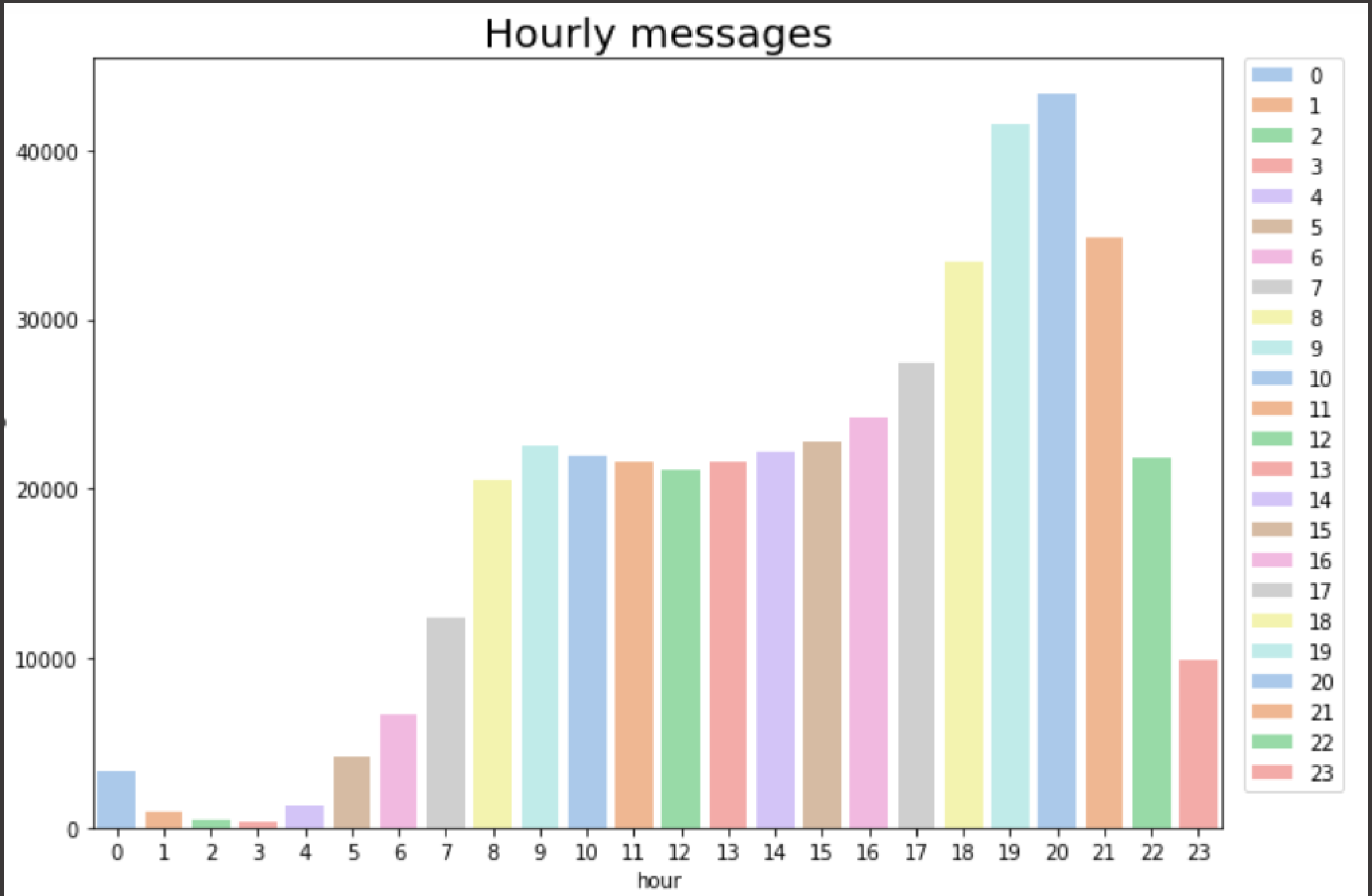
```
986
```

2. Message count stats (months/years)

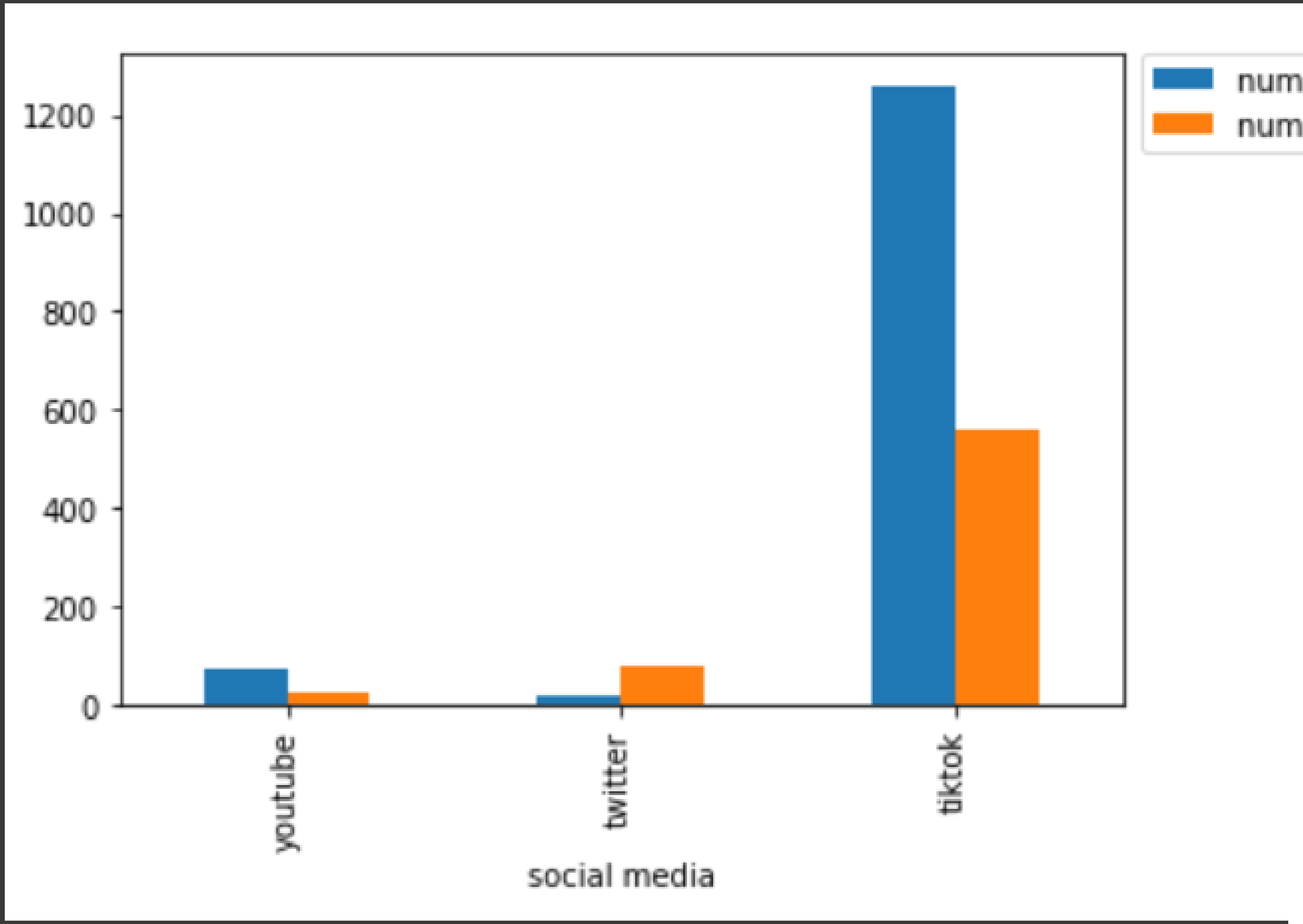


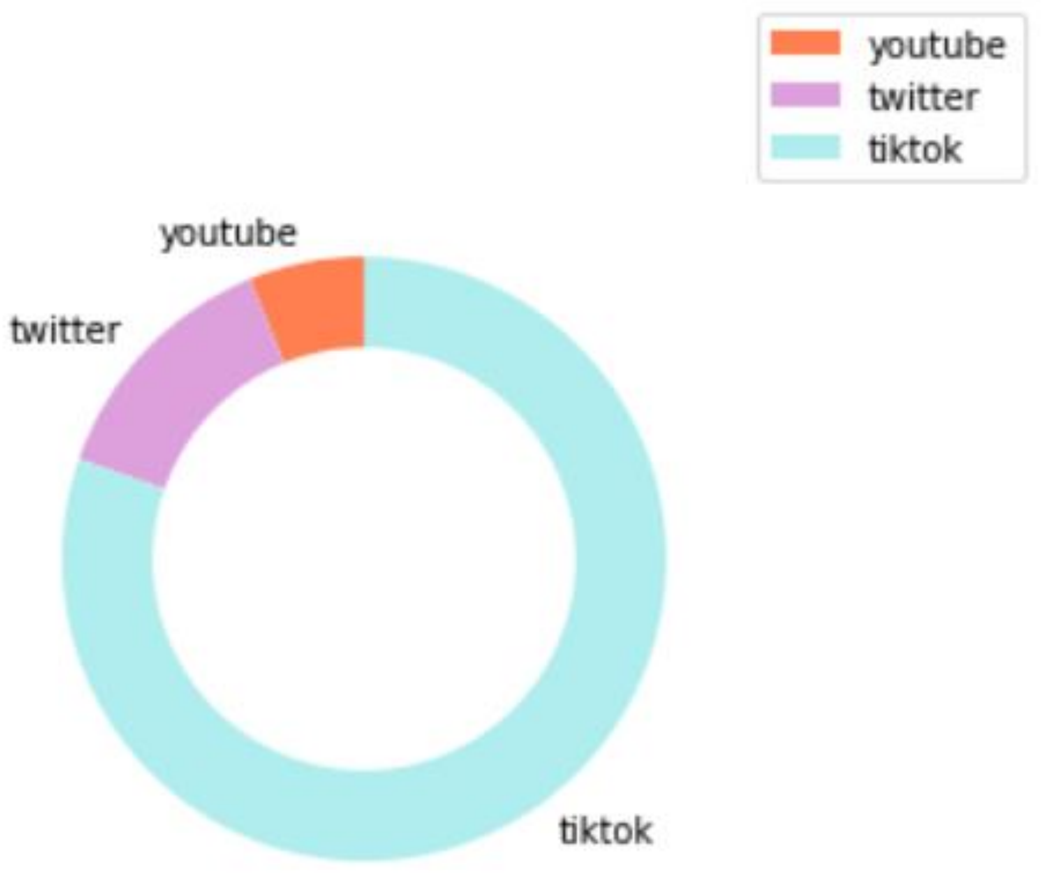


3. Messages sent by hours

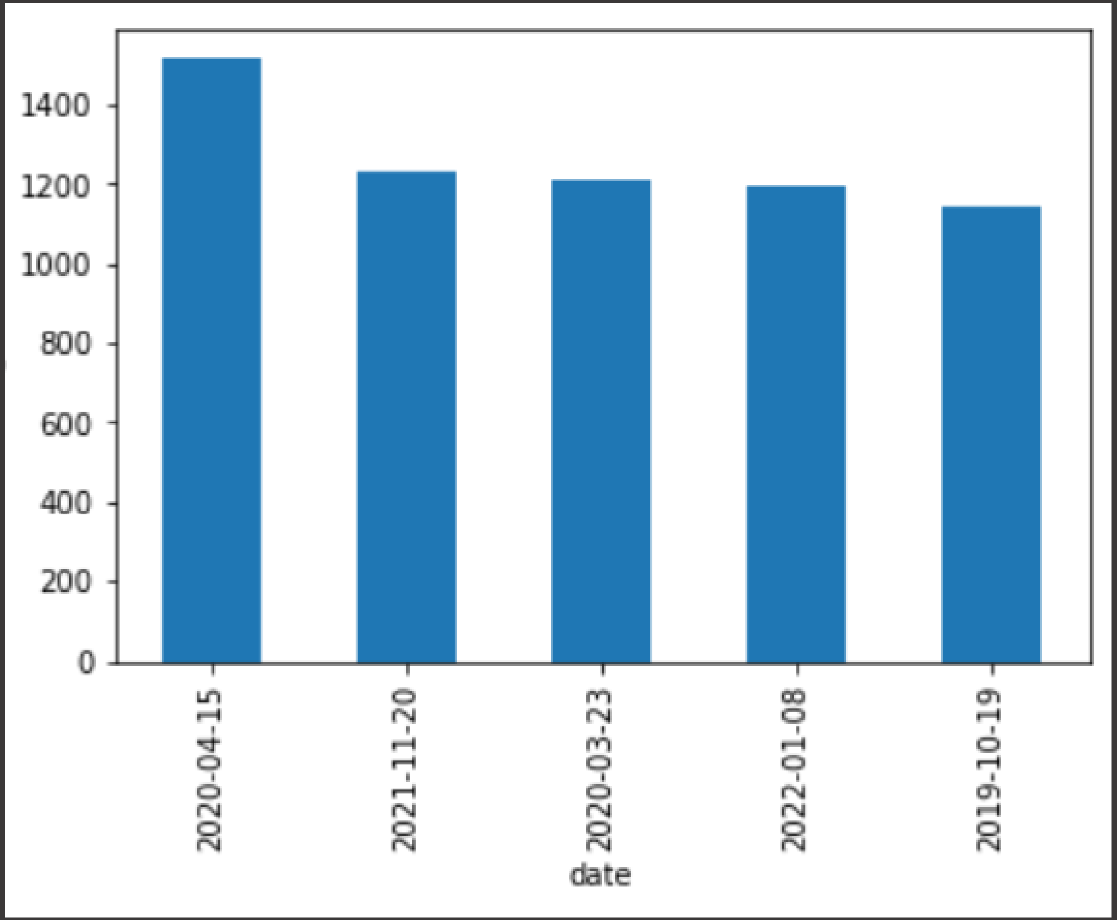


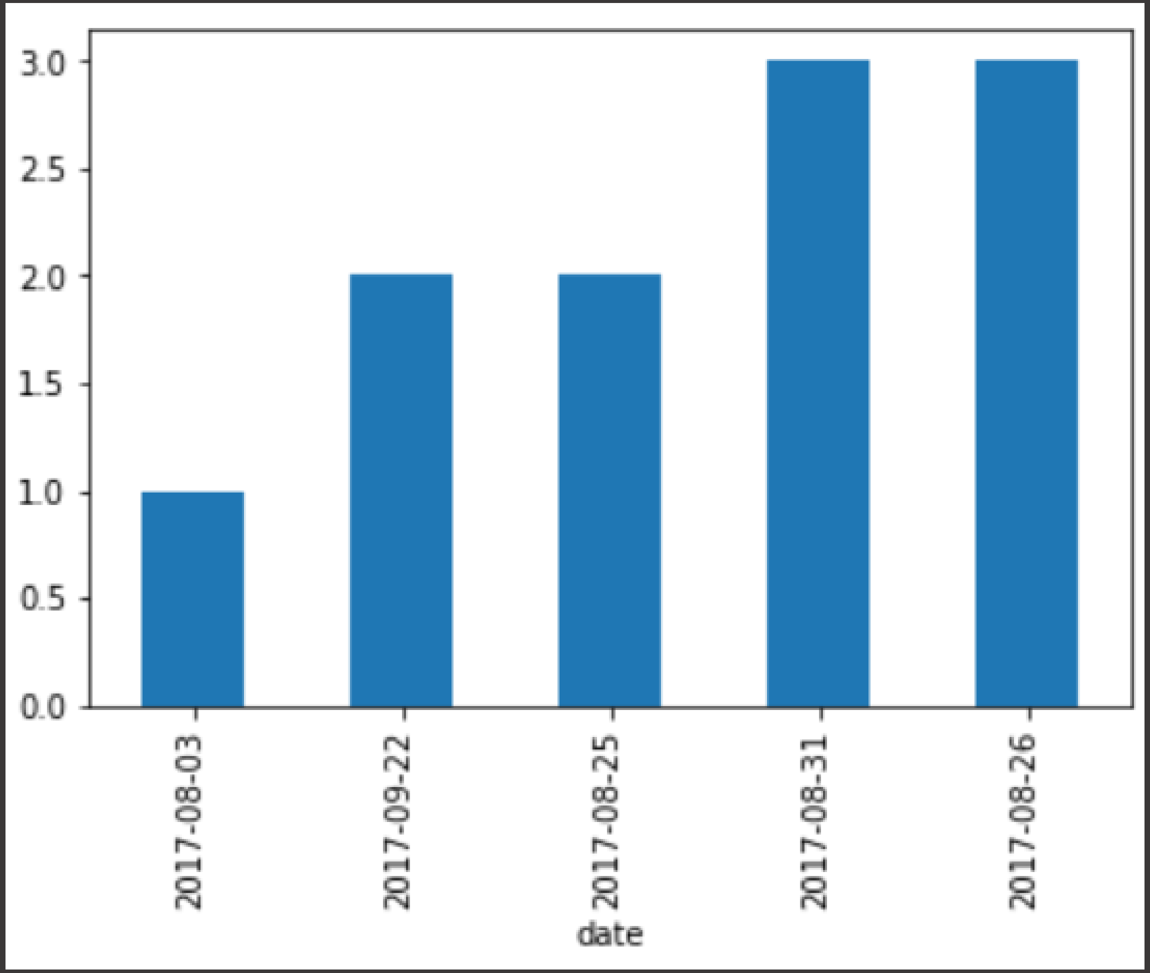
4. Links investigation

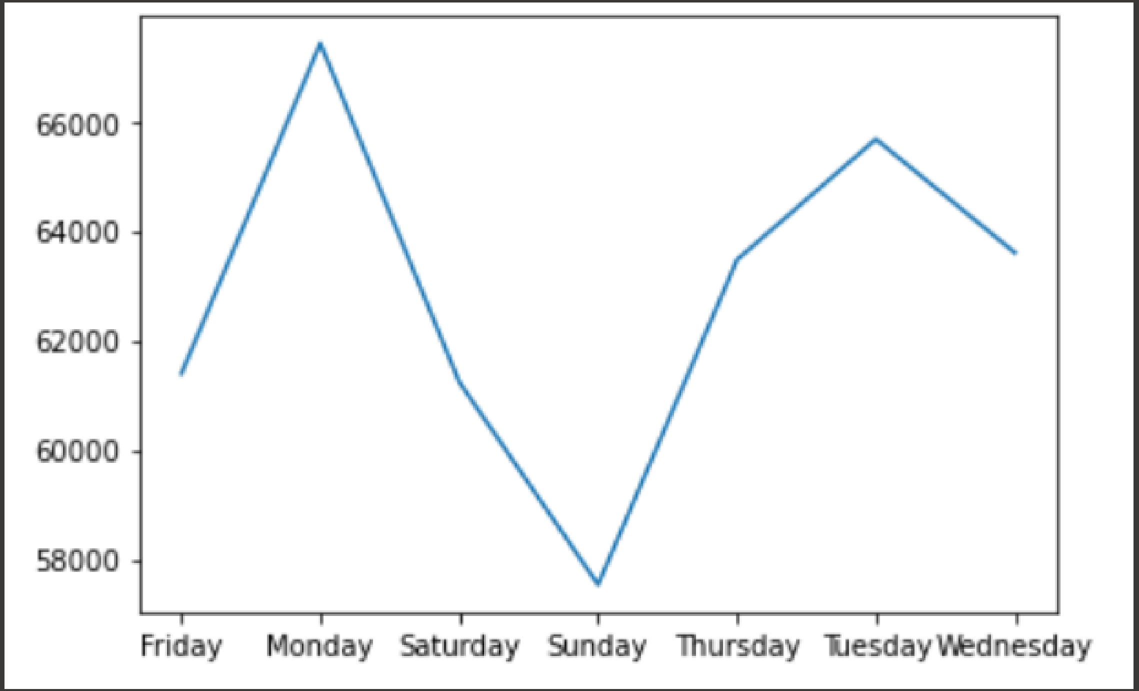


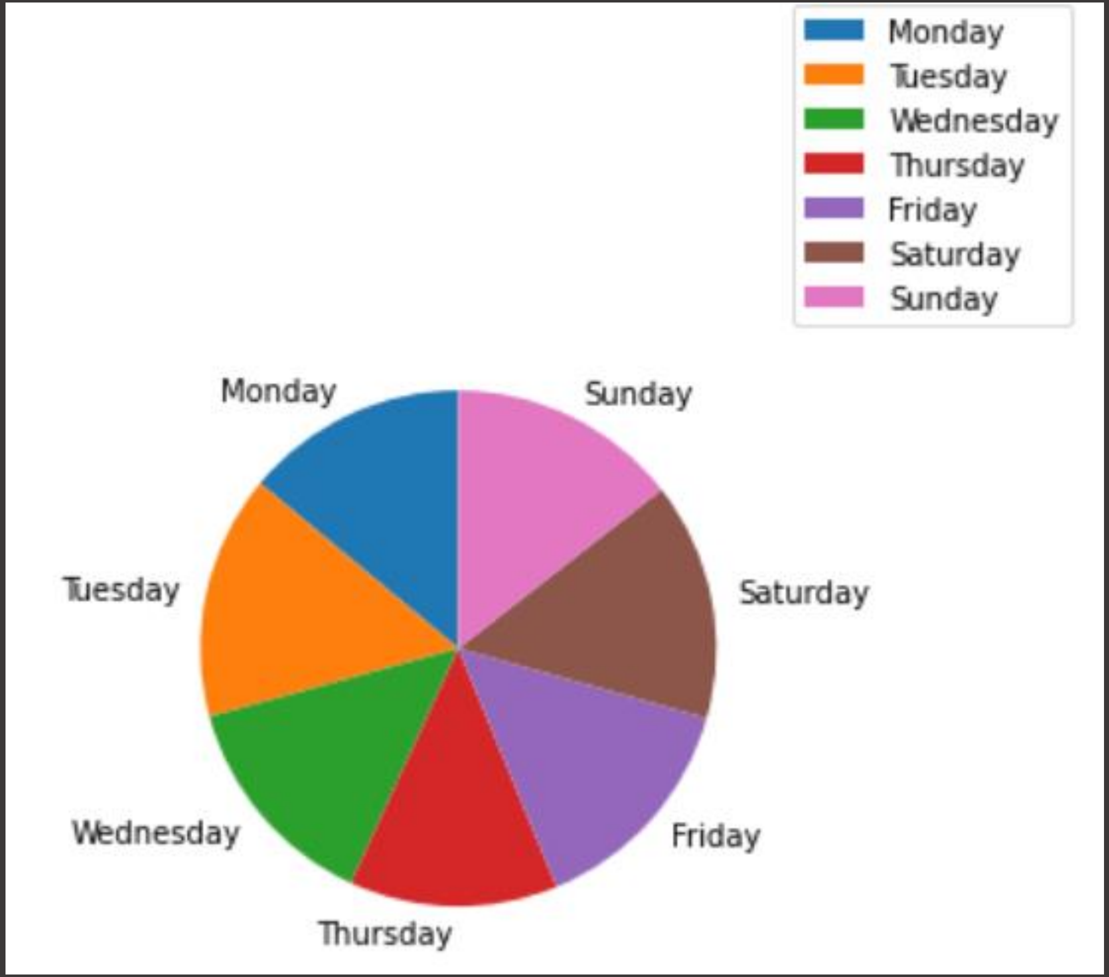


5. 5 most active and inactive days (sent messages)

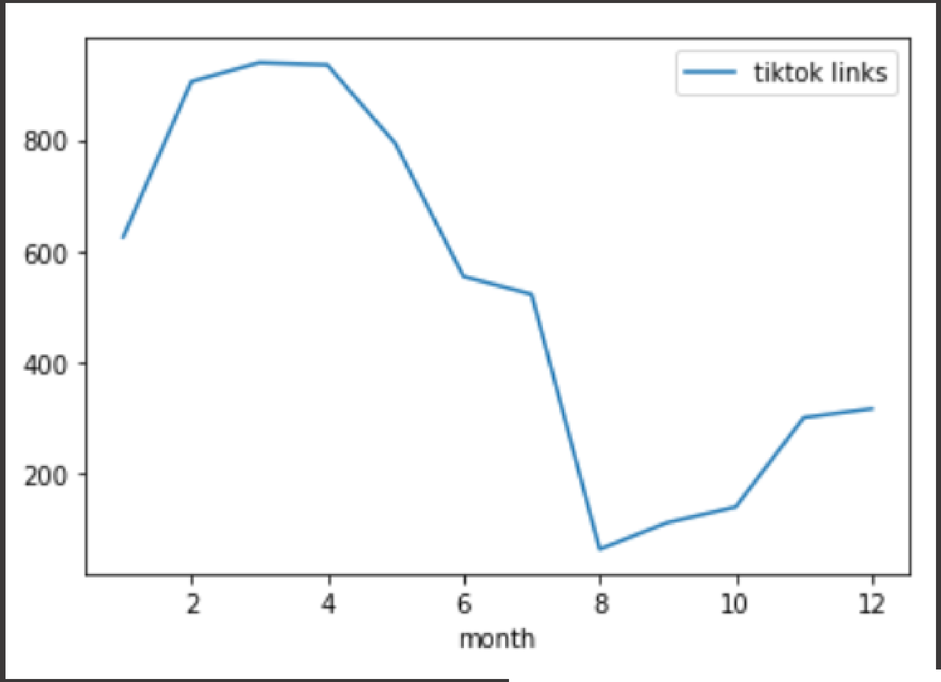


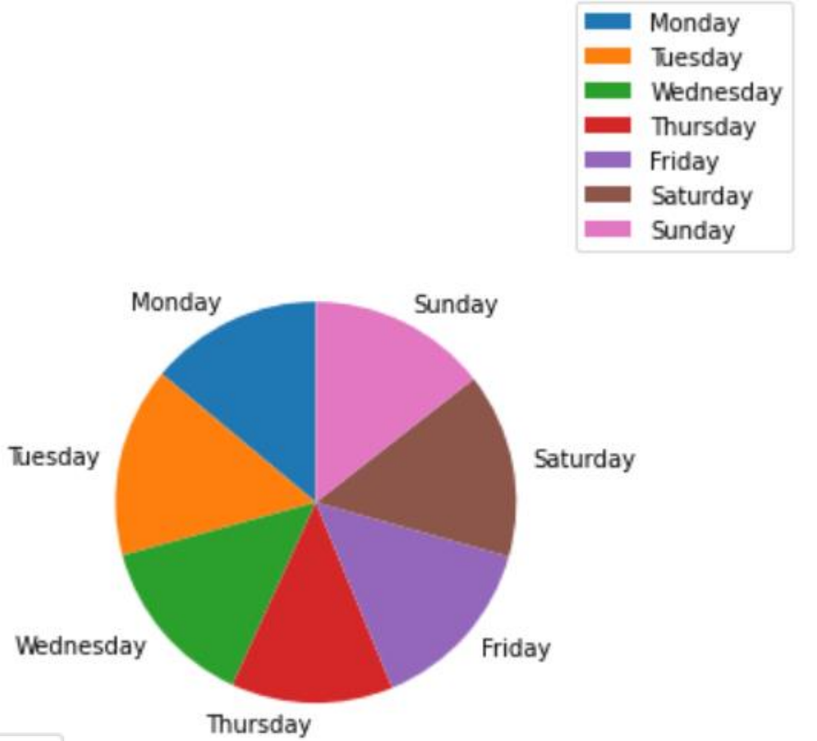



66000
64000
62000
60000
58000
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday

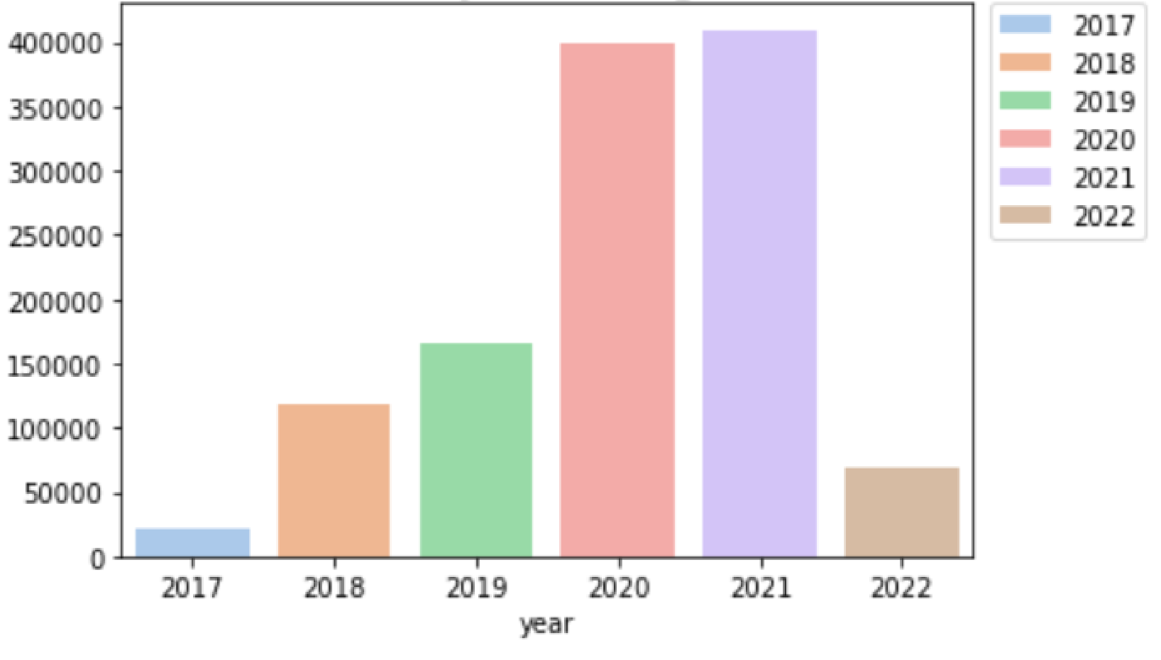

Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Monday
Sunday
Tuesday
Saturday
Wednesday
Friday
Thursday

Further work





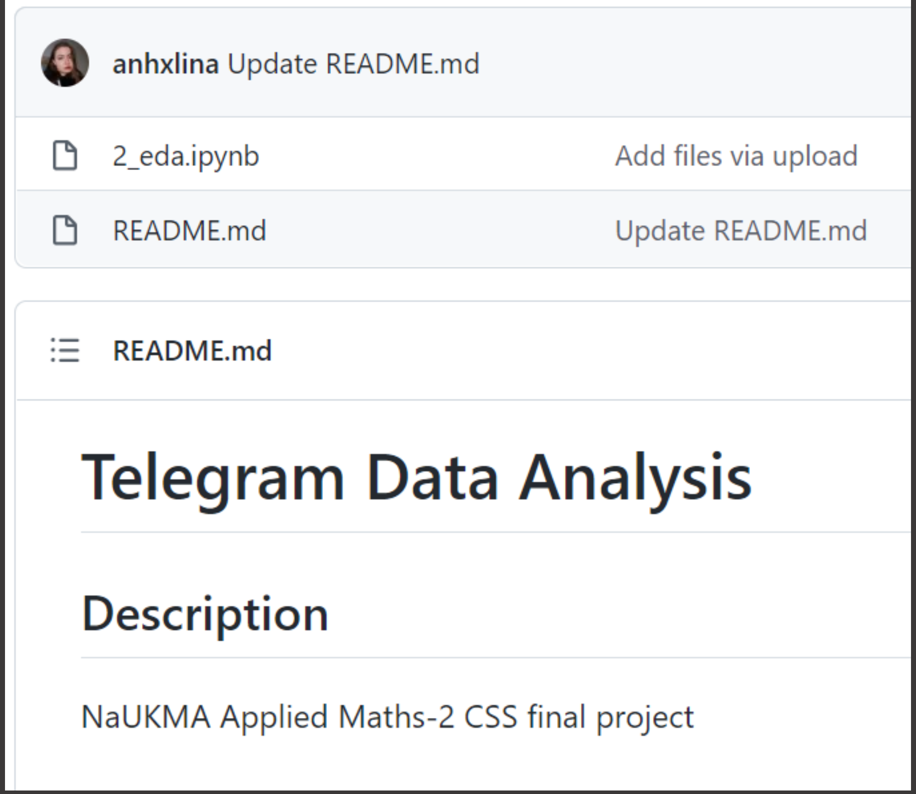
Yearly messages



Feel free to ask questions
(that are somehow connected to the project)

The full version of the project is available at:

https://github.com/anhxlina/telegram_research



anhxlina Update README.md

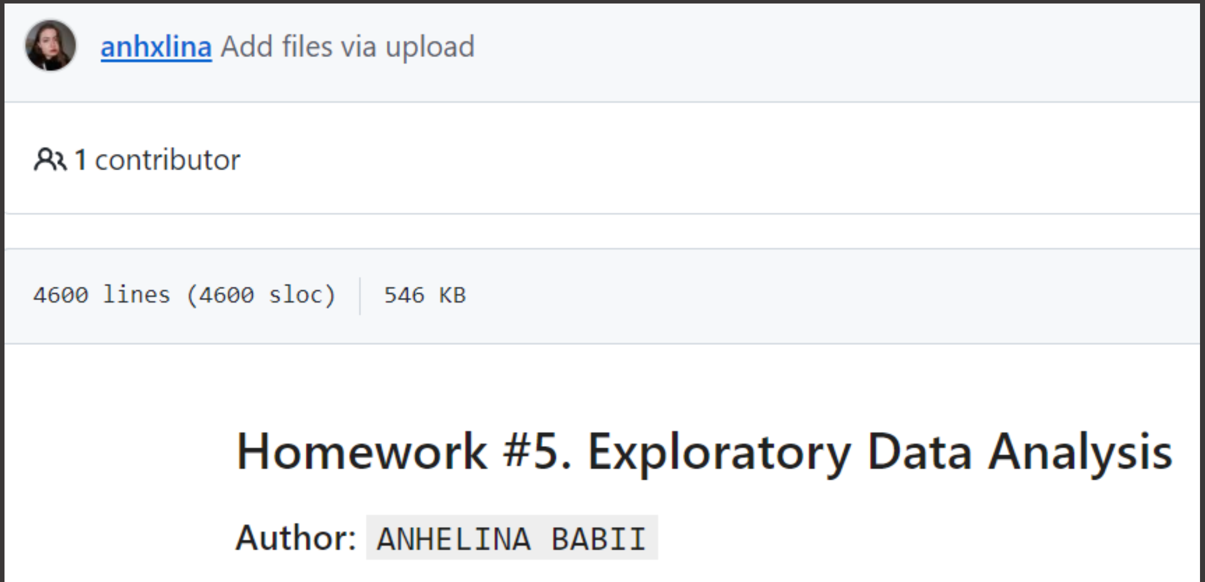
2_eda.ipynb	Add files via upload
README.md	Update README.md

README.md

Telegram Data Analysis

Description

NaUKMA Applied Maths-2 CSS final project



anhxlina Add files via upload

1 contributor

4600 lines (4600 sloc) | 546 KB

Homework #5. Exploratory Data Analysis

Author: `ANHELINA BABII`

Thank you for your attention!