In [1]:
```
import re
from collections import defaultdict
from tqdm import tqdm
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px


import nltk
import pymorphy2
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
import sklearn.preprocessing as preprocessing

import warnings
warnings.filterwarnings("ignore")
```

Intro

We will explore our propagandistic data deeper and will make `TF-IDF` vectorization on messages text.

In [2]:
```
PATH = r"data/data.csv"
data = pd.read_csv(PATH)
data.drop(["Unnamed: 0", "date", "reactions", "to_id", "msg_entity"], axis=1, inplace=True)
```

In [3]:
```
_data = data.copy()
_data["datetime"] = pd.to_datetime(_data["datetime"])

na_messages = _data[_data["message"].isna()]
_data.drop(na_messages.index, inplace=True)
```

In [4]:
```
_data.head()
```

Out[4]:

	id	views	fwd_from	message	type	duration	channel	frw_from_title	frw_from_name	datetime	mes…e_len	reactic
0	189123.0	98413.0	NaN	ФТС России ожидает роста товарооборота с Китае...	text	NaN	rian_ru	NaN	NaN	2022-12-19 09:56:04+00:00	205	
4	189119.0	118174.0	NaN	Буэнос-Айрес наутро после праздника	video	10.0	rian_ru	NaN	NaN	2022-12-19 09:51:57+00:00	35	
6	189117.0	224975.0	NaN	В СК сообщили, что жизни рабочих, пострадавших...	photo	NaN	rian_ru	NaN	NaN	2022-12-19 09:10:44+00:00	141	
7	189116.0	226171.0	NaN	Самолет с пострадавшим при покушении главой Ру...	video	30.0	rian_ru	NaN	NaN	2022-12-19 09:09:39+00:00	116	
8	189115.0	256663.0	NaN	Норвежский король Харальд V (85 лет) госпитали...	text	NaN	rian_ru	NaN	NaN	2022-12-19 08:50:57+00:00	116	

In [5]:
```
_data.shape
```

Out[5]:
```
(7016265, 17)
```

In [6]:
```
messages = _data[["datetime", "message"]]
messages.head()
```

Out[6]:

	datetime	message
0	2022-12-19 09:56:04+00:00	ФТС России ожидает роста товарооборота с Китае...
4	2022-12-19 09:51:57+00:00	Буэнос-Айрес наутро после праздника
6	2022-12-19 09:10:44+00:00	В СК сообщили, что жизни рабочих, пострадавших...
7	2022-12-19 09:09:39+00:00	Самолет с пострадавшим при покушении главой Ру...
8	2022-12-19 08:50:57+00:00	Норвежский король Харальд V (85 лет) госпитали...

In [7]:
```
def to_vector_preprocessing(text, stop_words = []):
    if not stop_words:
        stop_words = stopwords.words("english")
#     morph = pymorphy2.MorphAnalyzer()
#     morph.parse(word)[0].normal_form
#     stemmer = PorterStemmer()
#     stemmer.stem(word)
    text_array = word_tokenize(re.sub('[\W\s\d]', ' ', text.lower()))
    processed_text = ' '.join(
            [word for word in text_array
            if (len(word) > 2) and (word not in stop_words)
            ])
    return processed_text
```

In [8]:
```
def tfidf_vectorizer(_corpus):
    vectorizer = TfidfVectorizer()
    X = vectorizer.fit_transform(_corpus)
    sparse_matrix = pd.DataFrame(X.todense(), columns=vectorizer.get_feature_names())
    return sparse_matrix
```

In [9]:
```
stop_words = [
    "и", "в", "во", "не", "что", "он", "на", "я", "с", "со", "как", "а", "то", "все", "она", "так", "его",
    "но", "да", "ты", "к", "у", "же", "вы", "за", "бы", "по", "только", "ее", "мне", "было", "вот", "от",
    "меня", "еще", "нет", "о", "из", "ему", "теперь", "когда", "даже", "ну", "вдруг", "ли", "если", "уже",
    "или", "ни", "быть", "был", "него", "до", "вас", "нибудь", "опять", "уж", "вам", "ведь", "там", "потом"
    "себя", "ничего", "ей", "может", "они", "тут", "где", "есть", "надо", "ней", "для", "мы", "тебя", "их",
    "чем", "была", "сам", "чтоб", "без", "будто", "чего", "раз", "тоже", "себе", "под", "жизнь", "впрочем",
    "хорошо", "всю", "эти", "тогда", "были", "та", "бывает", "лучше", "это", "http"]
langs = ['russian']
for lang in langs:
    stop_words += stopwords.words(lang)
stop_words = set(stop_words)
```

Because we have a large number of text data, we will evaluate the time for preprocessing and also get a sample (`frac=0.1`) due to the limited computation resources.

In [10]:
```
%%time
n_test = 1000
processed_text = messages["message"][:n_test].map(lambda row: to_vector_preprocessing(row, stop_words))
```

```
Wall time: 204 ms
```

In [11]:
```
print(f"The number of words in first {n_test} messages:\t {sum(messages['message'][:n_test].map(len))}")
```

```
The number of words in first 1000 messages:      199967
```

Evaluation of computation for sample:

$$287559504 \times 204 / (199967 * 60000) \approx 4.88 \text{ min}$$

In [12]:
```
messages_10_per = messages.sample(frac=0.1)
messages_10_per.head()
```

Out[12]:

	datetime	message
1449181	2022-07-26 16:47:16+00:00	Польский центр языка прокомментировал многочис...
6691776	2022-09-22 14:54:28+00:00	Честно говоря, не ожидал такого разворота собы...
2519768	2018-07-16 15:24:45+00:00	Трамп: "Я продолжаю традиции дипломатического ...
4316816	2021-02-16 09:03:52+00:00	Подростки решили пострелять из пистолета сре...
7203281	2020-03-05 05:41:47+00:00	Безотносительно самого сайта, весьма поучитель...

In [13]:
```
messages_10_per.shape
```

Out[13]: (701626, 2)

In [14]:
```
messages_10_per_1 = messages_10_per.sort_values(by="datetime")
messages_10_per_1.head()
```

Out[14]:

	datetime	message
5282279	2015-09-24 19:35:13+00:00	Мне нужна визуализация вашей любви
1920194	2015-09-29 05:34:30+00:00	Проблемы Wi-Fi в московском метро стали темой ...
1920193	2015-09-29 16:09:16+00:00	Первый иск в Мосгорсуд о пожизненной блокировк...
1920190	2015-09-30 05:33:00+00:00	Рассмотрение дел об экстремизме поднимут на вы...
1920188	2015-09-30 07:44:23+00:00	Путин попросил Совфед разрешить использование ...

In [15]:
```
print(f"The number of words in sampled data:\t {sum(messages_10_per_1['message'].map(len))}")
```

```
The number of words in sampled data:     287559504
```

In [16]:
```
%%time
processed_mes = messages_10_per_1["message"].map(lambda row: to_vector_preprocessing(row, stop_words))
```

```
Wall time: 3min 22s
```

In [17]:
```
processed_mes_1 = pd.DataFrame(processed_mes).set_index(messages_10_per_1["datetime"])
processed_mes_1
```

Out[17]:

datetime	message
2015-09-24 19:35:13+00:00	нужна визуализация вашей любви
2015-09-29 05:34:30+00:00	проблемы московском метро стали темой игры izv...
2015-09-29 16:09:16+00:00	первый иск мосгорсуд пожизненной блокировке по...
2015-09-30 05:33:00+00:00	рассмотрение дел экстремизме поднимут высший у...
2015-09-30 07:44:23+00:00	путин попросил совфед разрешить использование ...
...	...
2022-12-26 08:06:48+00:00	поздравление деда мороза снегурочки стали доро...
2022-12-26 08:09:12+00:00	действия командующего группировкой украине сур...
2022-12-26 09:07:38+00:00	сша будут повышать ставки предела военный эксп...
2022-12-26 10:26:21+00:00	министерство обороны республики корея сообщает...
2022-12-26 10:44:22+00:00	южнокорейское минобороны отчиталось сегодняшне...

701626 rows × 1 columns

We also add some threshhold (`threshhold=50`) preprocessing to omit unvaluable messages with few words.

In [18]:
```
threshhold = 50
processed_mes_2 = processed_mes_1[processed_mes_1["message"].apply(len) > threshhold]
processed_mes_2
```

Out[18]:

datetime	message
2015-09-29 05:34:30+00:00	проблемы московском метро стали темой игры izv...
2015-09-29 16:09:16+00:00	первый иск мосгорсуд пожизненной блокировке по...
2015-09-30 05:33:00+00:00	рассмотрение дел экстремизме поднимут высший у...
2015-09-30 07:44:23+00:00	путин попросил совфед разрешить использование ...
2015-09-30 08:42:28+00:00	американском штате джорджия впервые лет казнил...
...	...
2022-12-26 08:06:48+00:00	поздравление деда мороза снегурочки стали доро...
2022-12-26 08:09:12+00:00	действия командующего группировкой украине сур...
2022-12-26 09:07:38+00:00	сша будут повышать ставки предела военный эксп...
2022-12-26 10:26:21+00:00	министерство обороны республики корея сообщает...
2022-12-26 10:44:22+00:00	южнокорейское минобороны отчиталось сегодняшне...

615878 rows × 1 columns

In [19]:
```
%%time

top_words_by_message = []
for idx in range(processed_mes_2.shape[0]):
    sm_idx = tfidf_vectorizer(processed_mes_2.iloc[idx].tolist())
    sm_idx.index = [processed_mes_2.index[idx]]
    top_words_by_message.append(sm_idx)
```

```
Wall time: 16min 43s
```

In [20]:
```
len(top_words_by_message)
```

Out[20]: 615878

Besides, we will resample data for 3 months period for convenient interpretation of results and RAM limit if the the period will be too high:

```
slicing_threshhold_3M = the number of messages / data period in months = 615878 / (7 x 4) ≈ 22000
```

Note: we resample data in the average number of messages for 3 months and thus periods can actually be inequal to define.

In [158]:
```
%%time

counter, slicing_threshhold = 0, 22000 # 3M period
top_words_by_3M = []
df_3M = defaultdict(float)
for row_idx in tqdm(range(len(top_words_by_message))):
    counter += 1
    current_message = top_words_by_message[row_idx].iloc[0].to_dict()
    for key, item in current_message.items():
        df_3M[key] += item
    if counter > slicing_threshhold:
        top_words_by_3M.append((top_words_by_message[row_idx - counter + 1].index, df_3M))
        df_3M = defaultdict(float)
        counter = 0
```

```
100%|████████████████████████████████████████████████████████████████████████████████
████| 615878/615878 [04:45<00:00, 2155.99it/s]

Wall time: 4min 45s
```

In [159]:
```
%%time

top_30_words_3M = []
for row_idx in tqdm(range(len(top_words_by_3M))):
    top_30_words_3M.append((top_words_by_3M[row_idx][0],
                            sorted(top_words_by_3M[row_idx][1].items(), key=lambda item: item[1], reverse=T
```

```
100%|████████████████████████████████████████████████████████████████████████████████
██████████████| 27/27 [00:05<00:00,  4.70it/s]

Wall time: 5.75 s
```

In [160]:
```
def plot_results_idx(top_words_by_period, idx):
    words_series = pd.Series(dict(top_words_by_period[idx][1]))
    return px.bar(words_series.reset_index().rename(columns={0: "value"}),
       x="index",
       y="value",
       color="value",
       color_continuous_scale='Agsunset').data
```

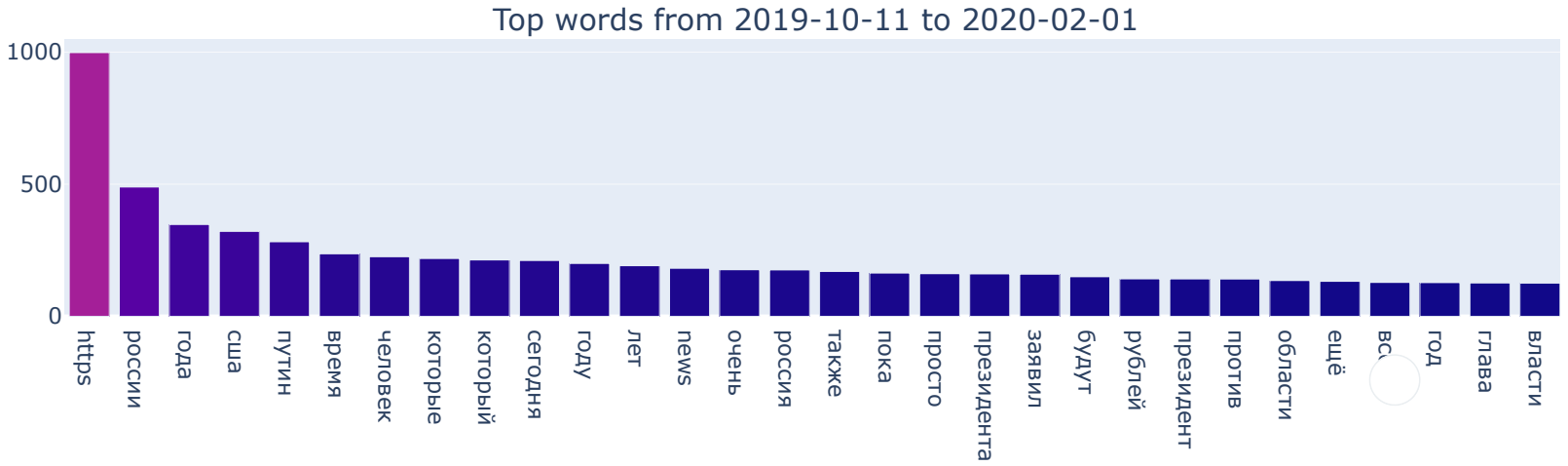
In [161]:
```
from plotly.subplots import make_subplots

titles = []
for row_idx in range(len(top_30_words_3M) - 1):
    start = top_30_words_3M[row_idx][0].strftime("%Y-%m-%d").values[0]
    end = top_30_words_3M[row_idx + 1][0].strftime("%Y-%m-%d").values[0]
    titles.append(f"Top words from {start} to {end}")

fig = make_subplots(rows=len(top_30_words_3M), cols=1, subplot_titles=titles)
```

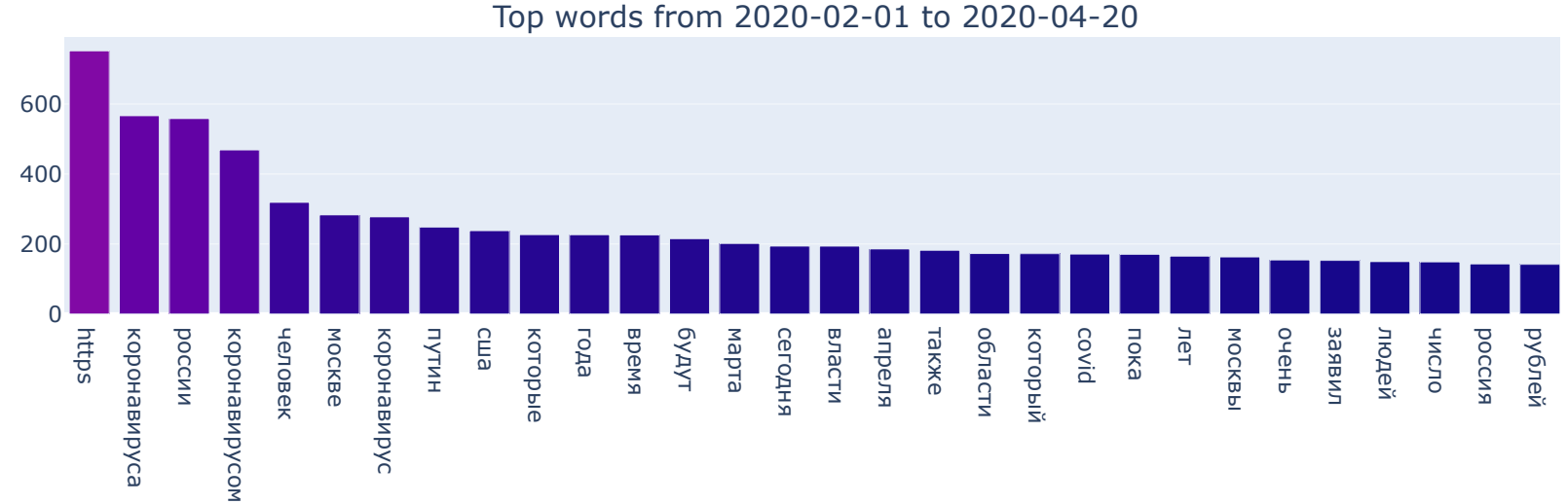
In [162]:
```
for row_idx in range(len(top_30_words_3M)):
    for trace in plot_results_idx(top_30_words_3M, row_idx):
        fig.add_trace(trace, row=row_idx + 1, col=1)
fig.update_layout(height=8000, width=1000,
                  showlegend=False)
fig.show()
```

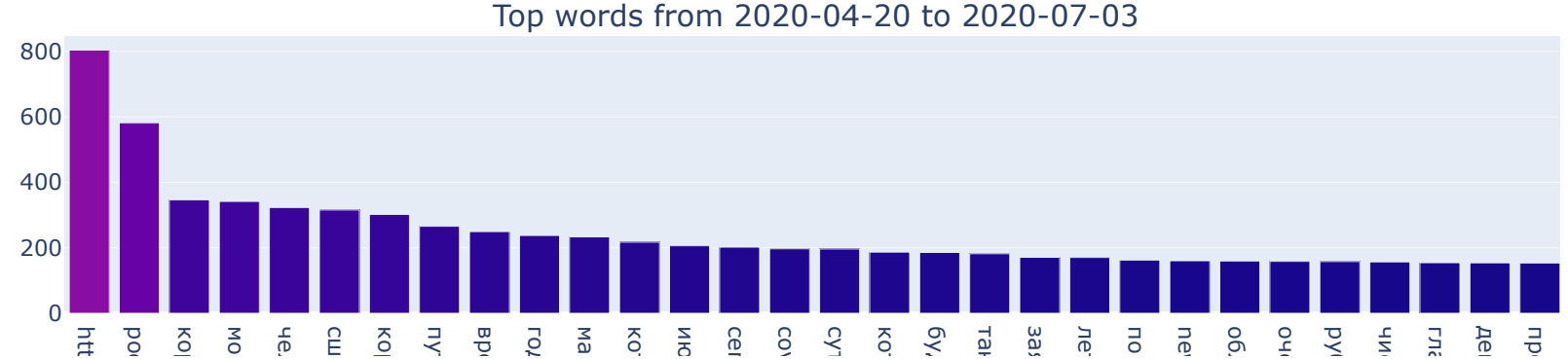

Top words from 2015-09-29 to 2017-11-07
2000
1000
0
https
www
россии
com
news
html
сша
goo
года
telegram
ria
utm_source
человек
время
очень
сегодня
москве
путин
который
ntv
novosti
которые
лет
просто
против
россия
видео
году
почему
читать
Top words from 2017-11-07 to 2018-06-27
1000
500
0
https
россии
news
html
www
года
сша
com
сегодня
путин
который
время
которые
человек
лет
очень
goo
просто
telegram
году
москве
пока
рублей
всё
президента
россия
сми
области
будут
против
Top words from 2018-06-27 to 2019-01-09
1000

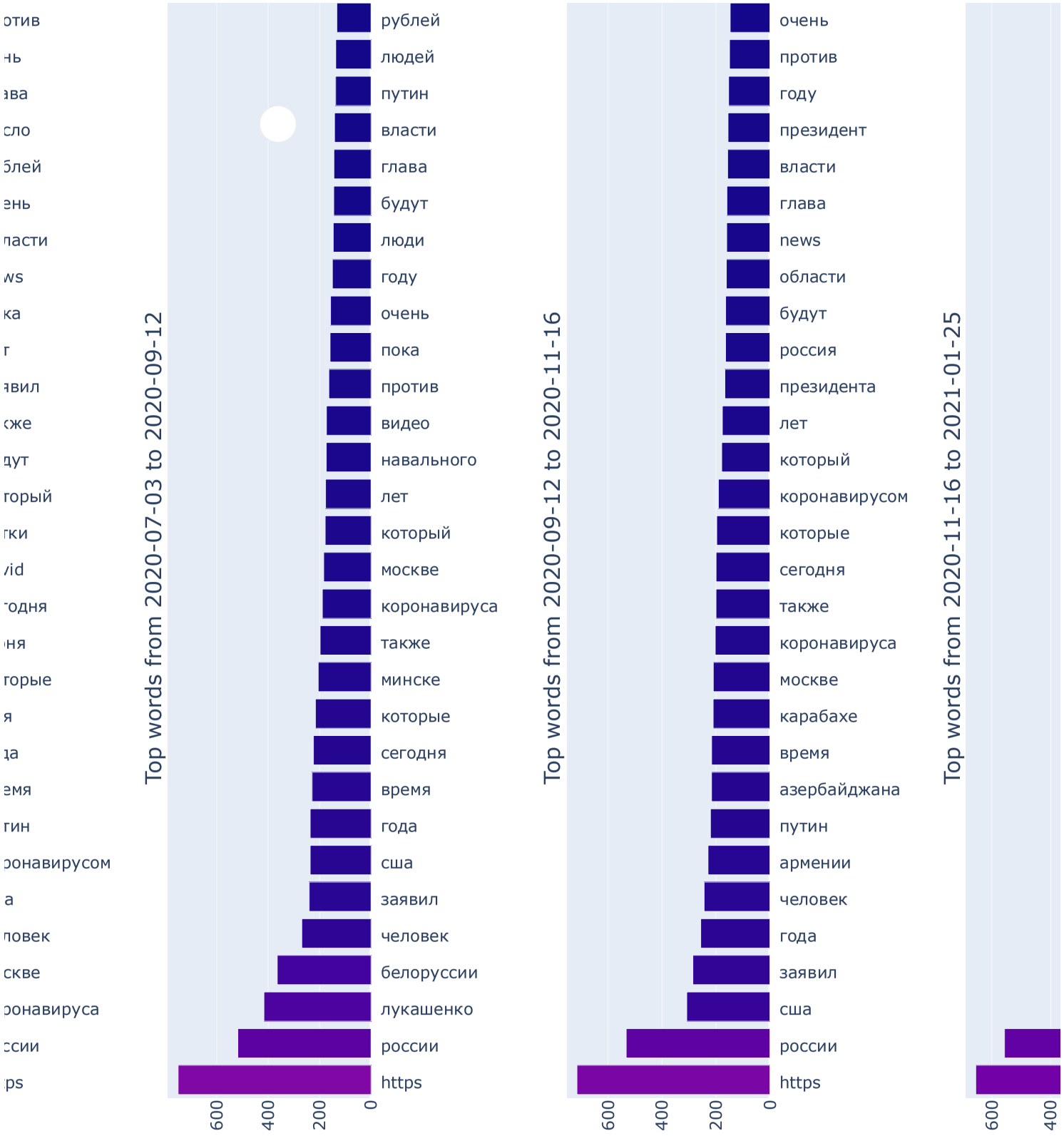
Top words from 2019-01-09 to 2019-06-15

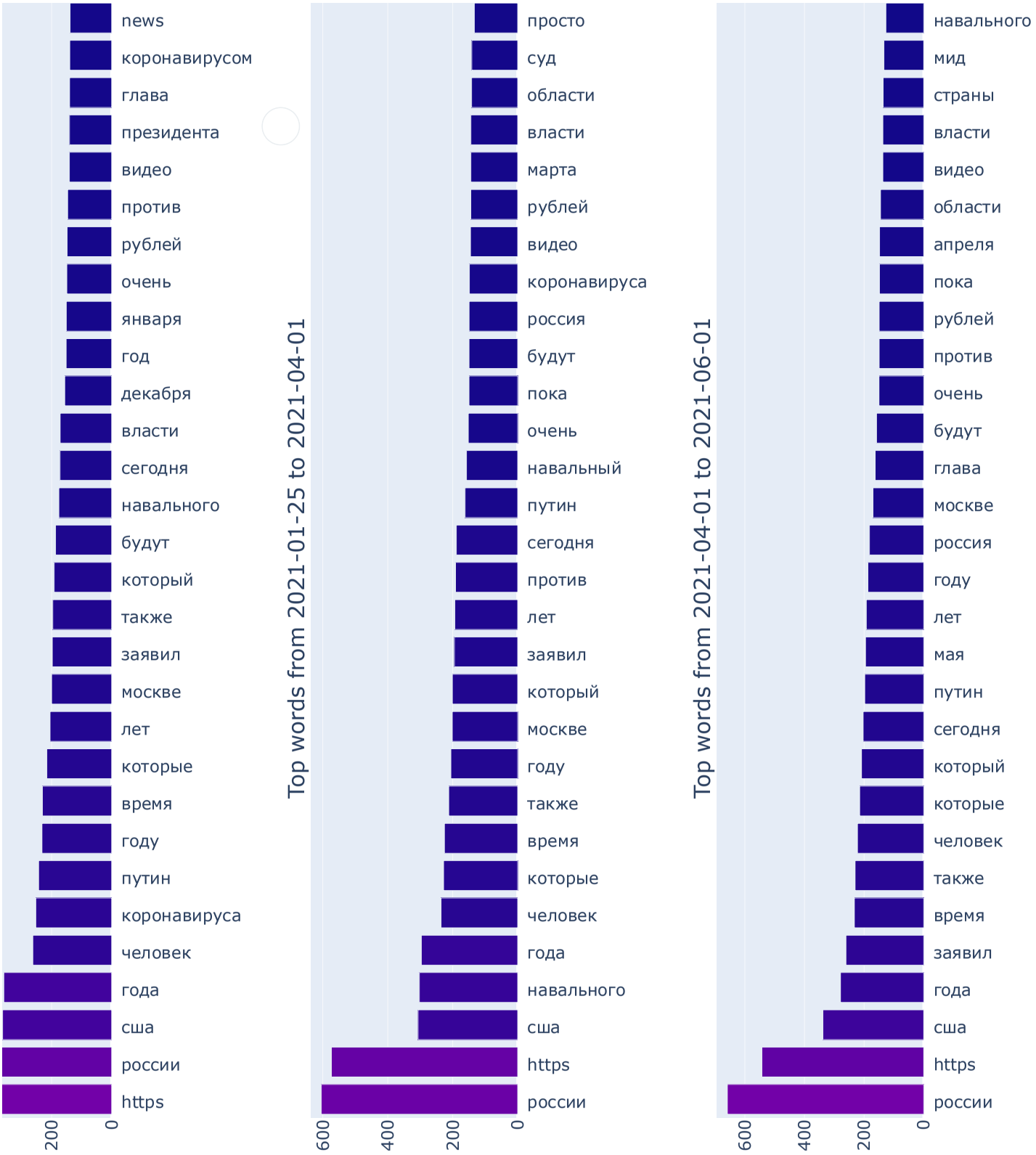
Top words from 2019-06-15 to 2019-10-11

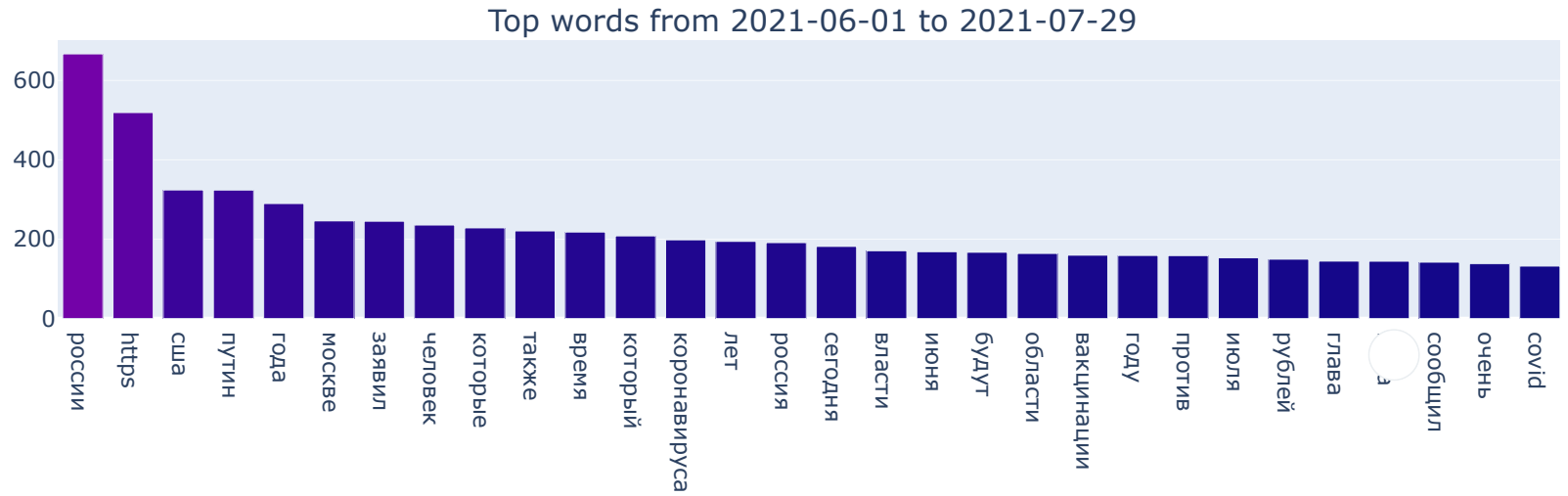

Top words from 2019-10-11 to 2020-02-01
1000
500
0
https
россии
года
сша
путин
время
человек
которые
который
сегодня
году
лет
news
очень
россия
также
пока
просто
президента
заявил
будут
рублей
президент
против
области
ещё
год
глава
власти

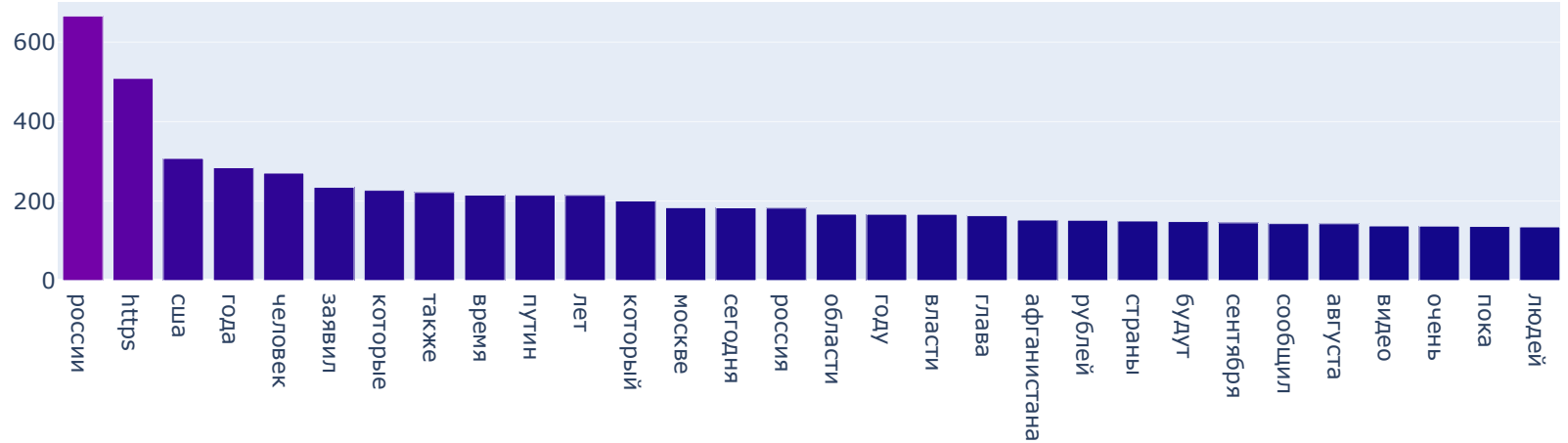

Top words from 2020-02-01 to 2020-04-20
600
400
200
0
https
коронавируса
россии
коронавирусом
человек
москве
коронавирус
путин
сша
которые
года
время
будут
марта
сегодня
власти
апреля
также
области
который
covid
пока
лет
москвы
очень
заявил
людей
число
россия
рублей

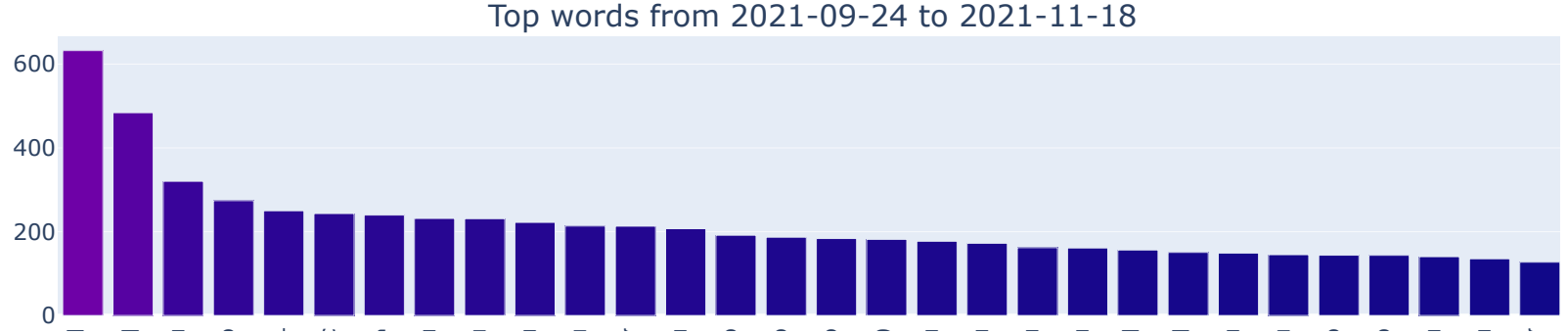

Top words from 2020-04-20 to 2020-07-03
800
600
400
200
0


Top words from 2020-07-03 to 2020-09-12
рублей
людей
путин
власти
глава
будут
люди
году
очень
пока
против
видео
навального
лет
который
москве
коронавируса
также
минске
которые
сегодня
время
года
сша
заявил
человек
белоруссии
лукашенко
россии
https
600
400
200
0
Top words from 2020-09-12 to 2020-11-16
очень
против
году
президент
власти
глава
news
области
будут
россия
президента
лет
который
коронавирусом
которые
сегодня
также
коронавируса
москве
карабахе
время
азербайджана
путин
армении
человек
года
заявил
сша
россии
https
600
400
200
0
Top words from 2020-11-16 to 2021-01-25
600
400

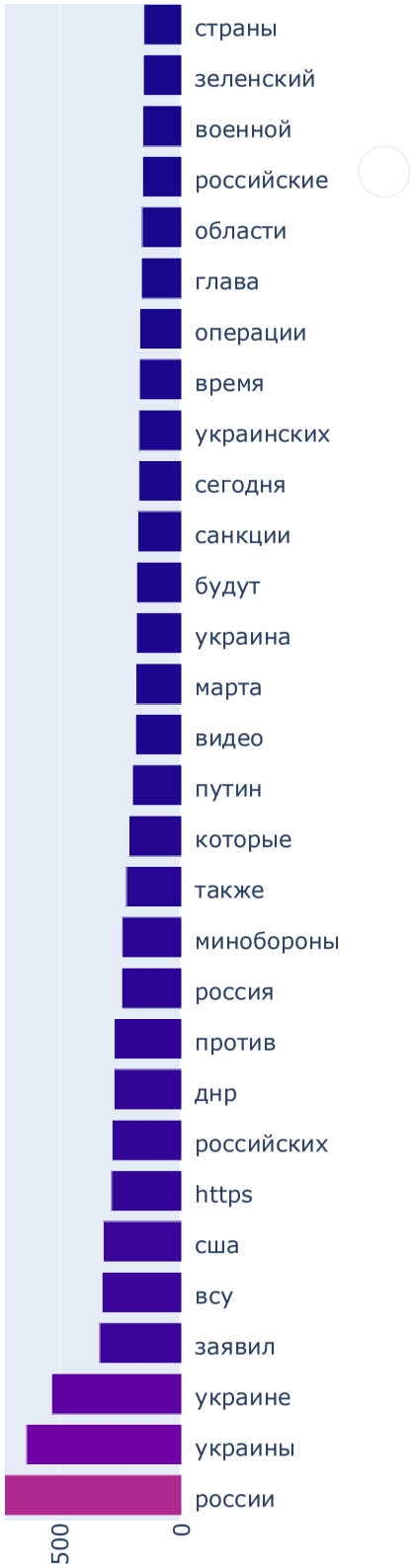

news
коронавирусом
глава
президента
видео
против
рублей
очень
января
год
декабря
власти
сегодня
навального
будут
который
также
заявил
москве
лет
которые
время
году
путин
коронавируса
человек
года
сша
россии
https
200
0
Top words from 2021-01-25 to 2021-04-01
просто
суд
области
власти
марта
рублей
видео
коронавируса
россия
будут
пока
очень
навальный
путин
сегодня
против
лет
заявил
который
москве
году
также
время
которые
человек
года
навального
сша
https
россии
600
400
200
0
Top words from 2021-04-01 to 2021-06-01
навального
мид
страны
власти
видео
области
апреля
пока
рублей
против
очень
будут
глава
москве
россия
году
лет
мая
путин
сегодня
который
которые
человек
также
время
заявил
года
сша
https
россии
600
400
200
0

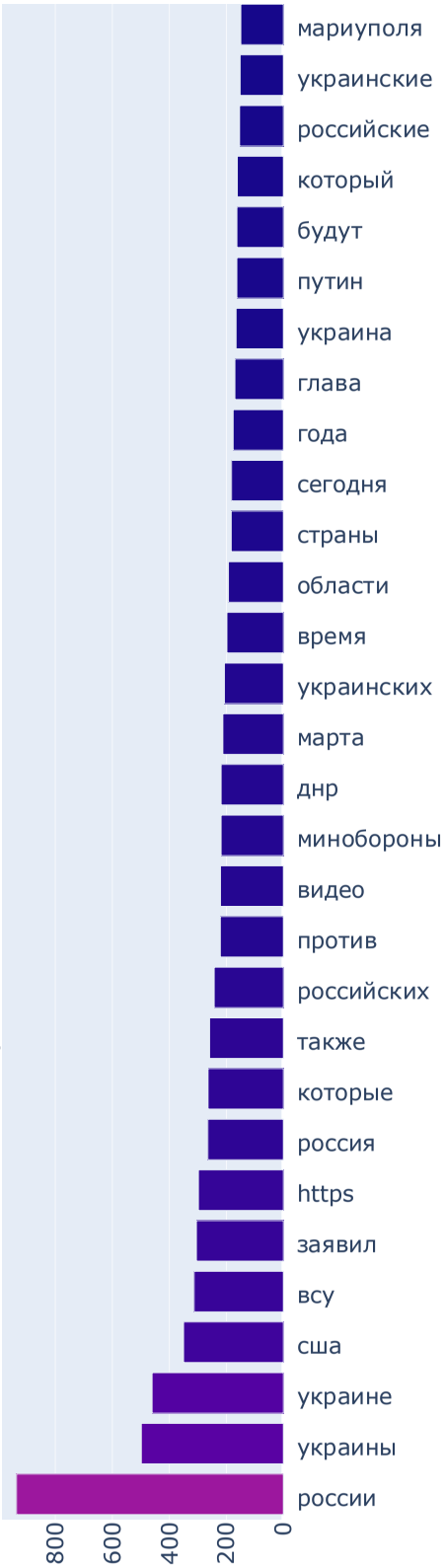

Top words from 2021-06-01 to 2021-07-29
600
400
200
0
россии
https
сша
путин
года
москве
заявил
человек
которые
также
время
который
коронавируса
лет
россия
сегодня
власти
июня
будут
области
вакцинации
году
против
июля
рублей
глава
сообщил
очень
covid

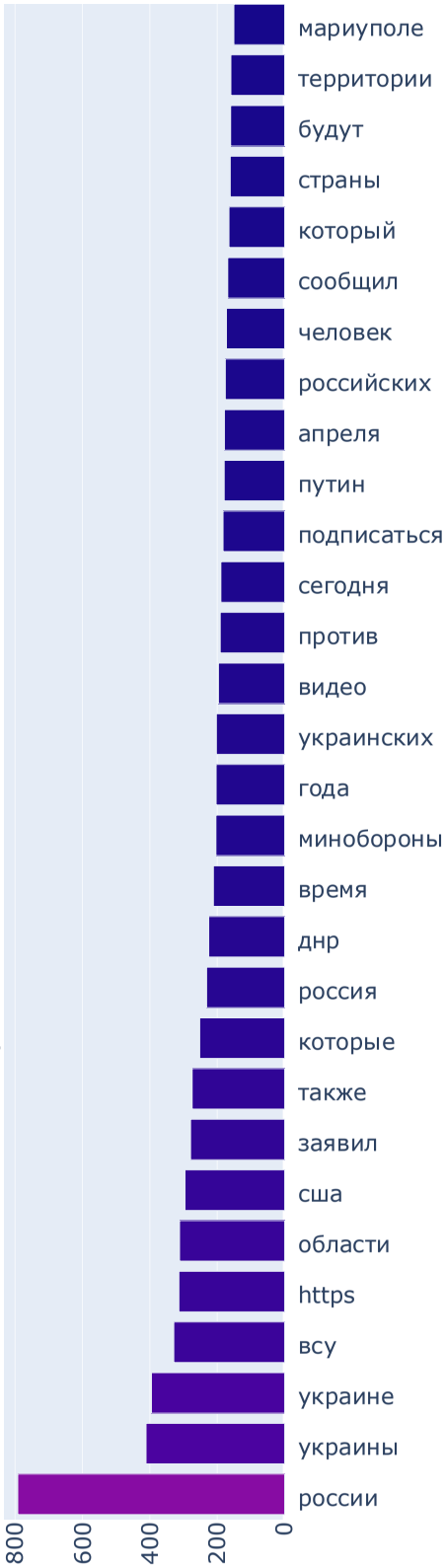

Top words from 2021-07-29 to 2021-09-24
600
400
200
0
россии
https
сша
года
человек
заявил
которые
также
время
путин
лет
который
москве
сегодня
россия
области
году
власти
глава
афганистана
рублей
страны
будут
сентября
сообщил
августа
видео
очень
пока
людей

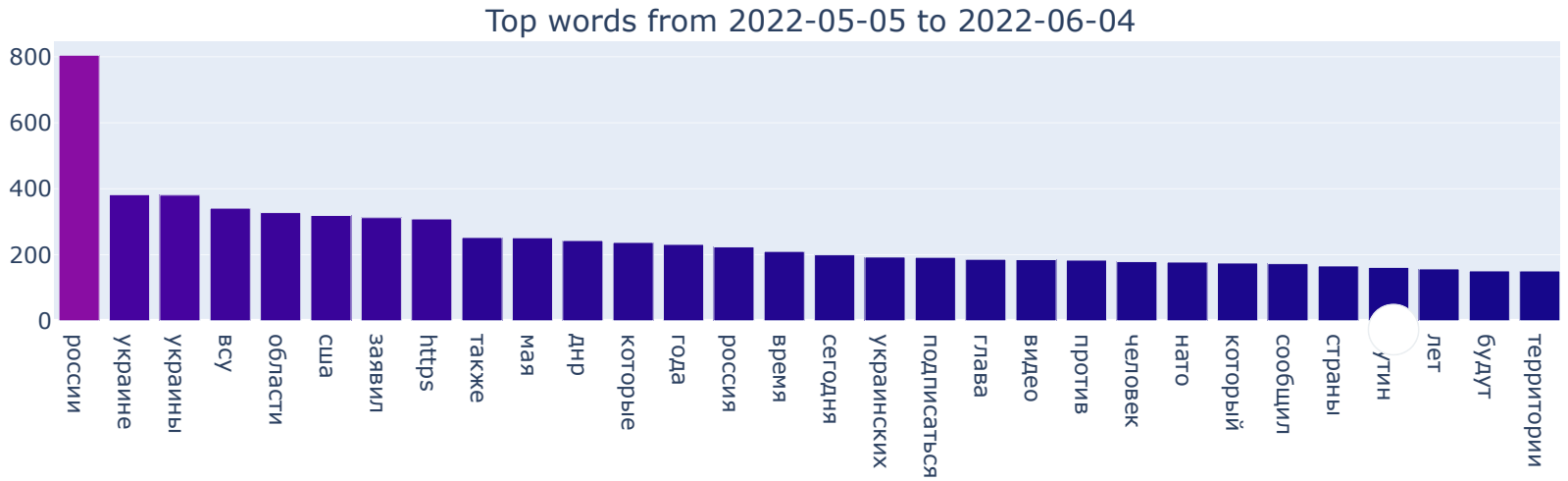

Top words from 2021-09-24 to 2021-11-18
600
400
200
0

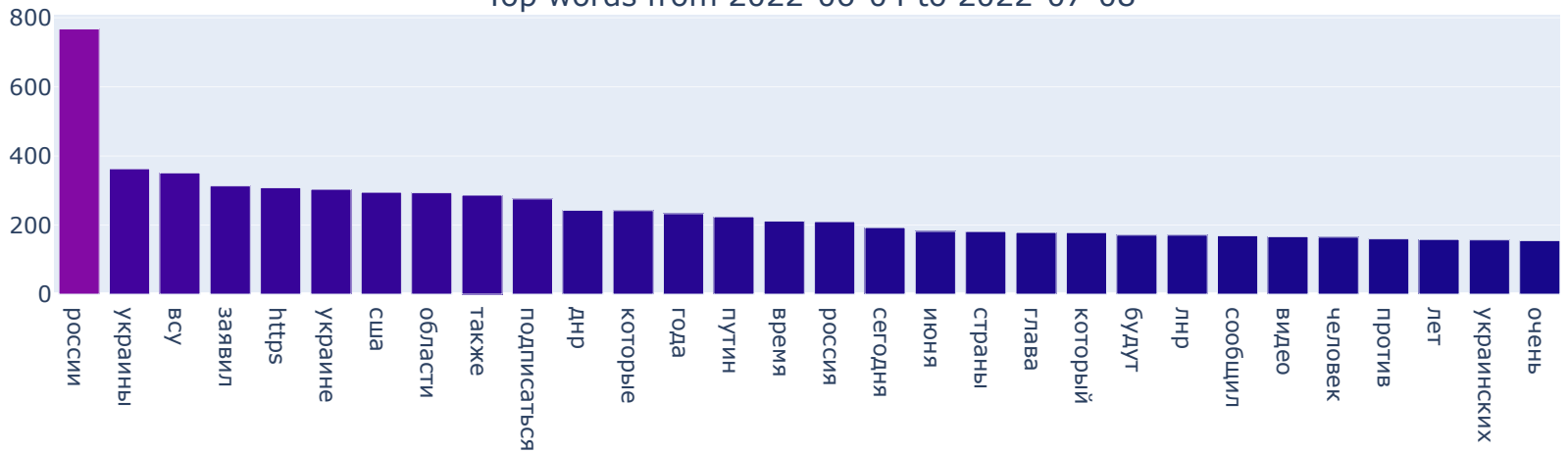

Top words from 2021-11-18 to 2022-01-12
страны
коронавируса
год
пока
области
сообщил
глава
казахстане
очень
власти
будут
москве
рублей
декабря
сегодня
против
россия
который
лет
время
которые
году
человек
также
заявил
путин
сша
года
https
россии
600
400
200
0
Top words from 2022-01-12 to 2022-02-24
человек
страны
президент
который
году
области
лет
мид
будут
безопасности
февраля
украине
сегодня
против
глава
время
украину
которые
также
нато
года
россия
лнр
путин
заявил
днр
https
украины
сша
россии
800
600
400
200
0
Top words from 2022-02-24 to 2022-03-14
1000

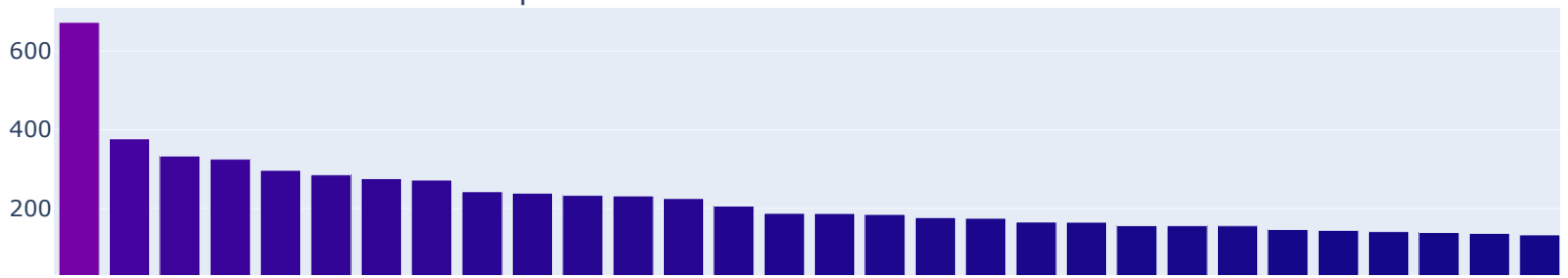

страны
зеленский
военной
российские
области
глава
операции
время
украинских
сегодня
санкции
будут
украина
марта
видео
путин
которые
также
минобороны
россия
против
днр
российских
https
сша
всу
заявил
украине
украины
россии
500
0


Top words from 2022-03-14 to 2022-04-07
мариуполя
украинские
российские
который
будут
путин
украина
глава
года
сегодня
страны
области
время
украинских
марта
днр
минобороны
видео
против
российских
также
которые
россия
https
заявил
всу
сша
украине
украины
россии
800
600
400
200
0

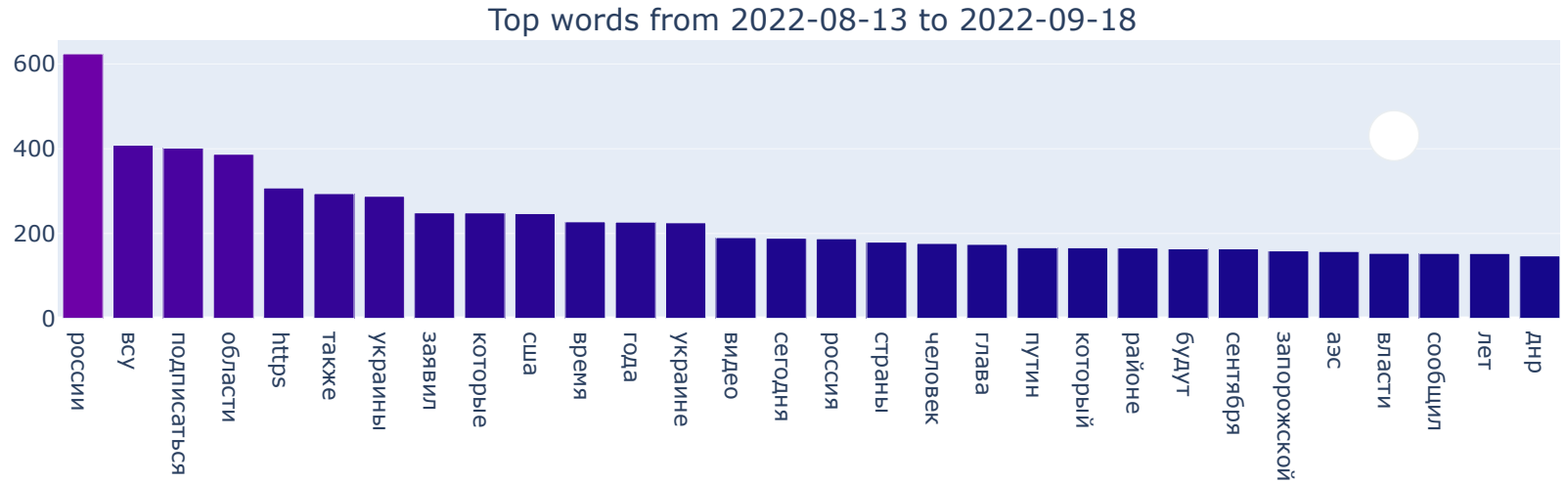

Top words from 2022-04-07 to 2022-05-05
мариуполе
территории
будут
страны
который
сообщил
человек
российских
апреля
путин
подписаться
сегодня
против
видео
украинских
года
минобороны
время
днр
россия
которые
также
заявил
сша
области
https
всу
украине
украины
россии
800
600
400
200
0

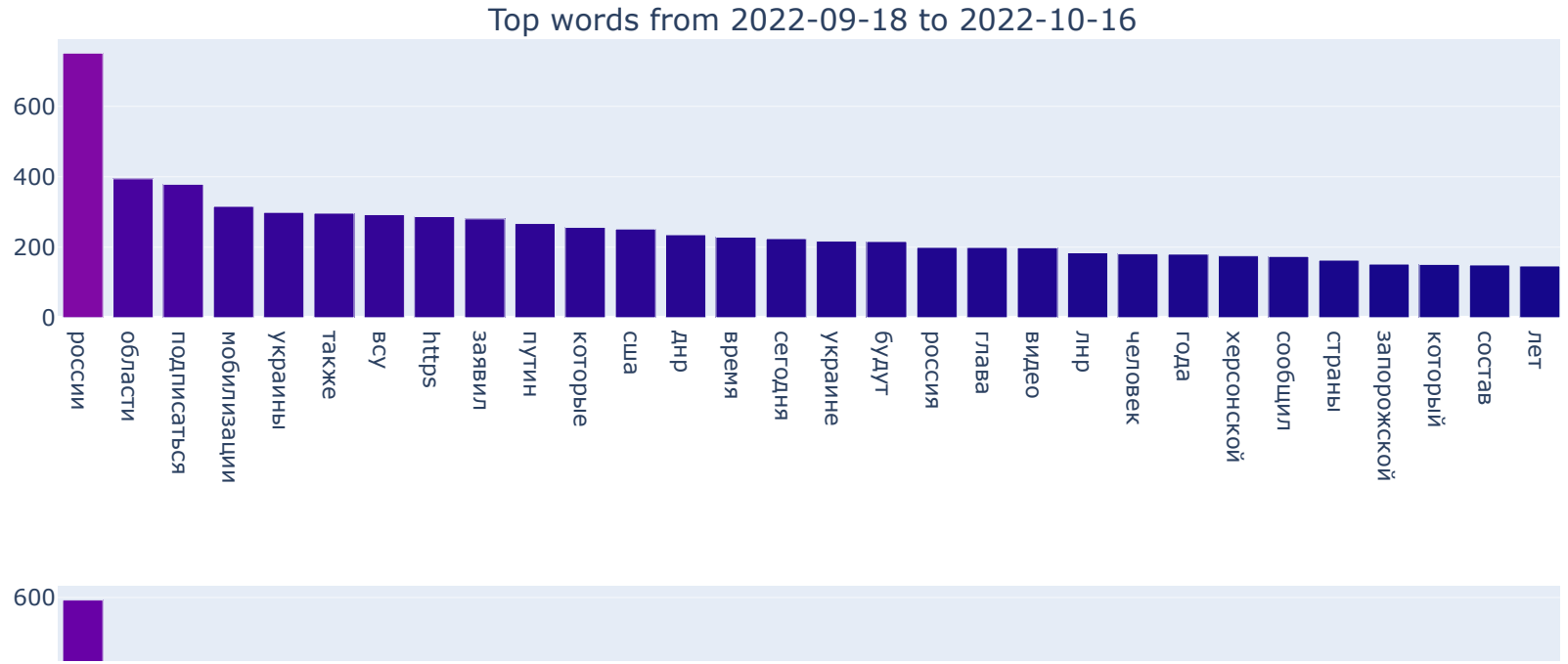

Top words from 2022-05-05 to 2022-06-04
800
600
400
200
0
россии
украине
украины
всу
области
сша
заявил
https
также
мая
днр
которые
года
россия
время
сегодня
украинских
подписаться
глава
видео
против
человек
нато
который
сообщил
страны
путин
лет
будут
территории


Top words from 2022-06-04 to 2022-07-08
800
600
400
200
0
россии
украины
всу
заявил
https
украине
сша
области
также
подписаться
днр
которые
года
путин
время
россия
сегодня
июня
страны
глава
который
будут
лнр
сообщил
видео
человек
против
лет
украинских
очень

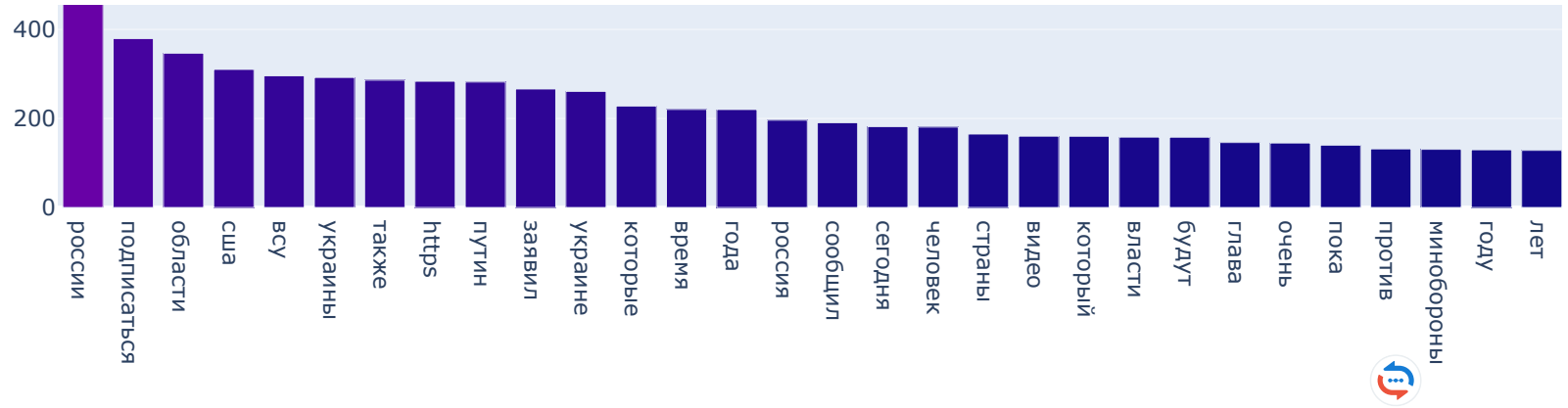

Top words from 2022-07-08 to 2022-08-13
600
400
200


0
россии
подписаться
сша
всу
украины
https
также
области
заявил
года
украине
которые
время
днр
россия
человек
сегодня
который
глава
видео
страны
будут
сообщил
лет
очень
против
путин
пока
сообщает
июля


Top words from 2022-08-13 to 2022-09-18
600
400
200
0
россии
всу
подписаться
области
https
также
украины
заявил
которые
сша
время
года
украине
видео
сегодня
россия
страны
человек
глава
путин
который
районе
будут
сентября
запорожской
аэс
власти
сообщил
лет
днр


Top words from 2022-09-18 to 2022-10-16
600
400
200
0
россии
области
подписаться
мобилизации
украины
также
всу
https
заявил
путин
которые
сша
днр
время
сегодня
украине
будут
россия
глава
видео
лнр
человек
года
херсонской
сообщил
страны
запорожской
который
состав
лет

600



From the cumulative `TF-IDF` text processing plots above, there is the overall tendency of significance russia, putin, the main russian regions, and viral events during the entire period.
There are also more significant mentions of the USA from the start of coronavirus to now.
From the beginning of the war in 2022-02, territories, Ukraine, defense, etc., have been more important.