Hypotesis

Hypothesis:
“different news channels carry different agendas”

Hypothesis justification:
some channels like to write about politics and positivity, some play with negativity and emotions. Some embellish the news, some describe the truth as it is.

Dataset overview

Ввод [16]:
```
df
```

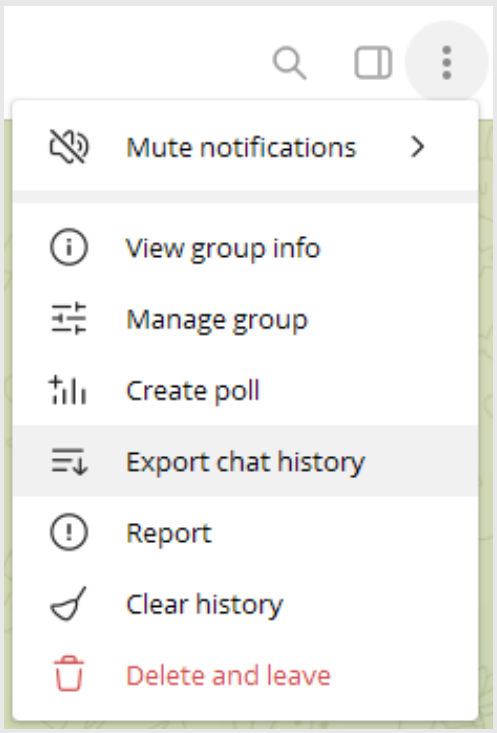
Out[16]:

	id	type	date	date_unixtime	actor	actor_id	action	title	text	text_entities	...	height	photo	mess
0	1	service	2020-09-07T10:30:29	1599453029	ORDA	channel1499924082	create_channel	ORDA.KZ		[]	...	NaN	NaN	
1	3	service	2020-10-31T13:43:36	1604130216	ORDA	channel1499924082	edit_group_title	ORDA.		[]	...	NaN	NaN	
2	4	message	2020-11-11T08:01:03	1605060063	NaN	NaN	NaN	NaN	[{'type': 'bold', 'text': '🚀 Оперативные новост...	[{'type': 'bold', 'text': '🚀 Оперативные новост...	...	NaN	NaN	
3	6	message	2020-11-11T10:42:48	1605069768	NaN	NaN	NaN	NaN	[{'type': 'bold', 'text': 'В каком случае могу...	[{'type': 'bold', 'text': 'В каком случае могу...	...	NaN	NaN	
4	7	message	2020-11-11T11:09:37	1605071377	NaN	NaN	NaN	NaN	[{'type': 'bold', 'text': 'Алматинский экстрем...	[{'type': 'bold', 'text': 'Алматинский экстрем...	...	352.0	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
37189	38746	message	2022-10-27T22:35:28	1666888528	NaN	NaN	NaN	NaN		[]	...	642.0	(File not included. Change data exporting sett...	
37190	38747	message	2022-10-27T22:35:28	1666888528	NaN	NaN	NaN	NaN		[]	...	711.0	(File not included. Change data exporting sett...	
37191	38748	message	2022-10-27T23:25:39	1666891539	NaN	NaN	NaN	NaN	[{'type': 'bold', 'text': 'Корову пришлось спа...	[{'type': 'bold', 'text': 'Корову пришлось спа...	...	485.0	(File not included. Change data exporting sett...	
37192	38749	message	2022-10-27T23:25:39	1666891539	NaN	NaN	NaN	NaN		[]	...	488.0	(File not included. Change data exporting sett...	
37193	38750	message	2022-10-27T23:25:39	1666891539	NaN	NaN	NaN	NaN		[]	...	485.0	(File not included. Change data exporting sett...	

37194 rows × 30 columns

Dataset overview

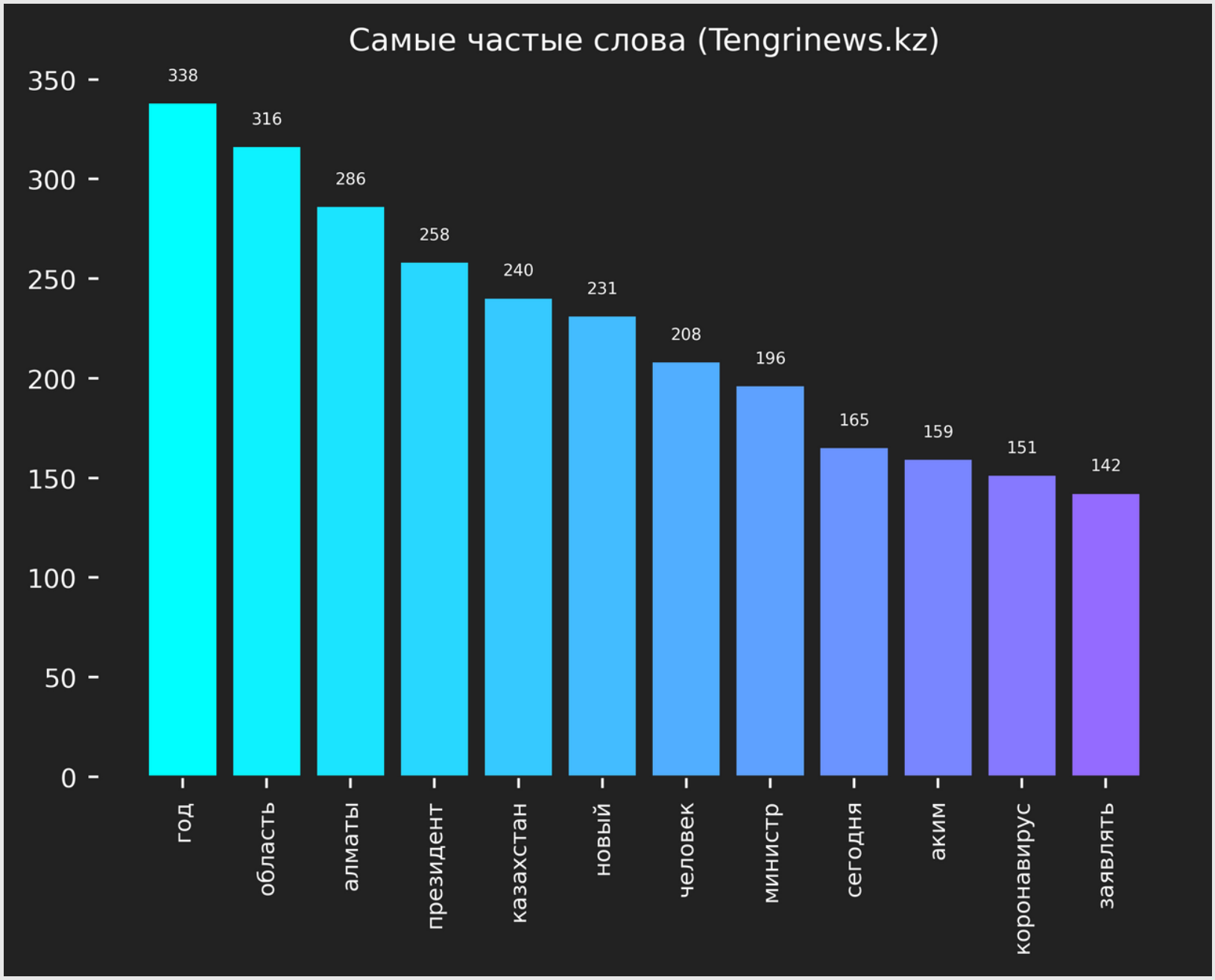
37000 rows data exported from ORDA telegram channel. Others channels data also available




```
Ввод [46]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37194 entries, 0 to 37193
Data columns (total 30 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   id                    37194 non-null  int64
 1   type                  37194 non-null  object
 2   date                  37194 non-null  object
 3   date_unixtime         37194 non-null  object
 4   actor                 657 non-null    object
 5   actor_id              657 non-null    object
 6   action                657 non-null    object
 7   title                 5 non-null      object
 8   text                  37194 non-null  object
 9   text_entities         37194 non-null  object
 10  edited                30370 non-null  object
 11  edited_unixtime       30370 non-null  object
 12  from                  36537 non-null  object
 13  from_id               36537 non-null  object
 14  file                  4591 non-null   object
 15  thumbnail             4552 non-null   object
 16  media_type            4577 non-null   object
 17  mime_type             4588 non-null   object
 18  duration_seconds      4574 non-null   float64
 19  width                 25575 non-null  float64
 20  height                25575 non-null  float64
 21  photo                 21002 non-null  object
 22  message_id            652 non-null    float64
 23  reply_to_message_id   9439 non-null   float64
 24  sticker_emoji         2 non-null      object
 25  poll                  102 non-null    object
 26  forwarded_from        744 non-null    object
 27  duration              1 non-null      float64
 28  performer             1 non-null      object
 29  author                1 non-null      object
dtypes: float64(6), int64(1), object(23)
memory usage: 8.5+ MB
```

diagramms overview



python code overview

```
russian_stopwords = stopwords.words("russian")
def preprocess_text(text):
    tokens = nltk.tokenize.word_tokenize(text)
    tokens = [token for token in tokens if token not in russian_stopwords\
              and token != " " \
              and token !='это'\
              and token.strip() not in punctuation]
    text = " ".join(tokens)
    text = [x for x in m.lemmatize(text) if x != ' ']
    return text
def pop_msg(x):
    messages_only = ""
    only_one = df.loc[df['from']==x]
    for i in only_one.itertuples():
            if(isinstance(i.text,str)==True):
                messages_only += "\n"
                messages_only += (i.text.lower())

    messages_only = preprocess_text(messages_only)
    fdist = FreqDist(messages_only)
    return fdist.most_common()[:10]

def pop_msg2(x):
    messages_only = ""
    only_one = df.loc[df['from']==x]
    for i in only_one.itertuples():
            if(isinstance(i.text,str)==True):
                messages_only += "\n"
                messages_only += (i.text.lower())

    messages_only = preprocess_text(messages_only)
```

map overview