Разговоры о семье в социальных сетях: анализ группы семейной тематики в сети Вконтакте

Большой МИР маленьких ДЕТЕЙ

https://vk.com/mir.detey

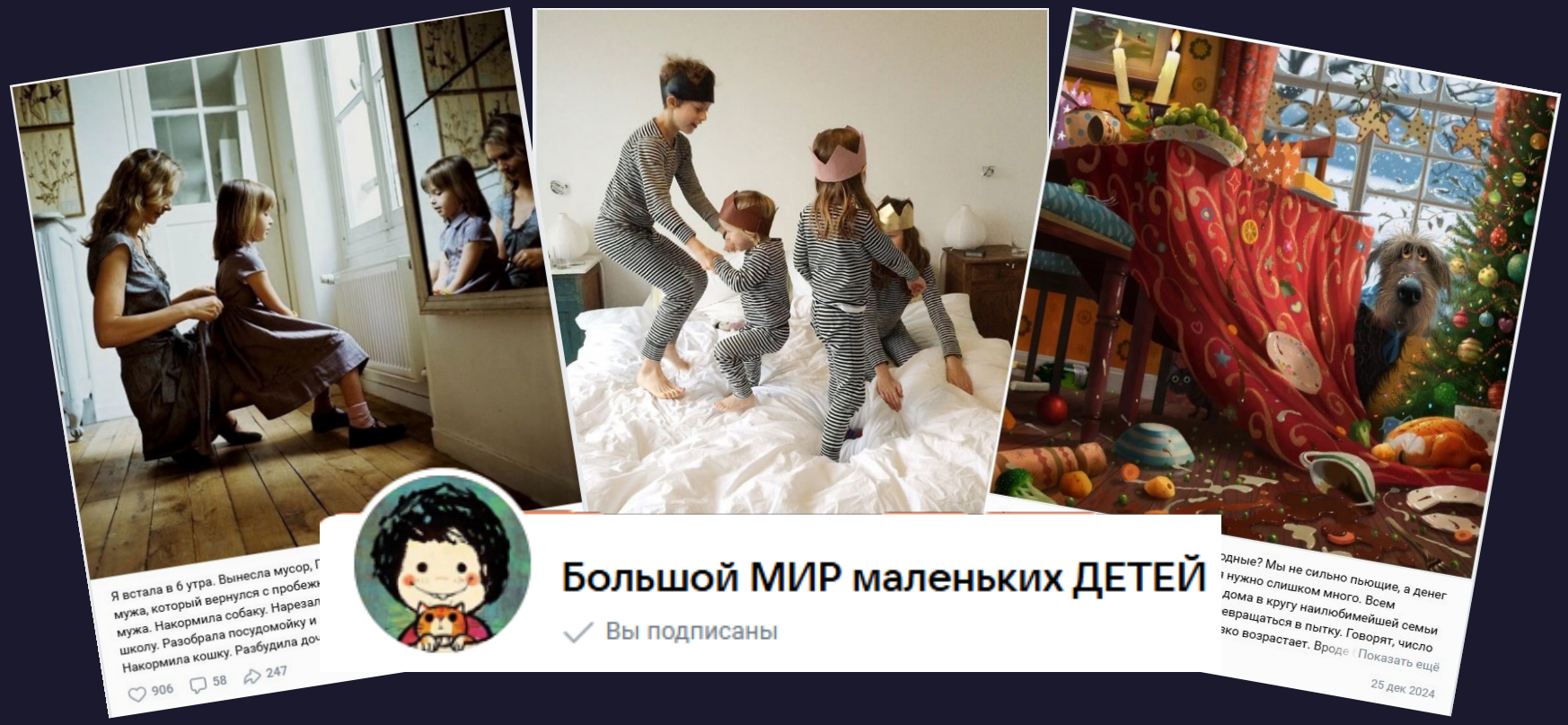
190 892 подписчиков

Задачи проекта

1. Собрать данные группы ВКонтакте “Большой МИР маленьких ДЕТЕЙ”
2. Структурировать данные для дальнейшей удобной работы с ними: работа с pandas
3. Выполнить частотный анализ постов группы, включающий:

- вывод самых частотных слов
- подсчет частоты для разных частей речи (существительные, прилагательные, глаголы)
- вывод самых частотных биграмм и триграмм
- вывод самых частотных словосочетаний с определенным ключом

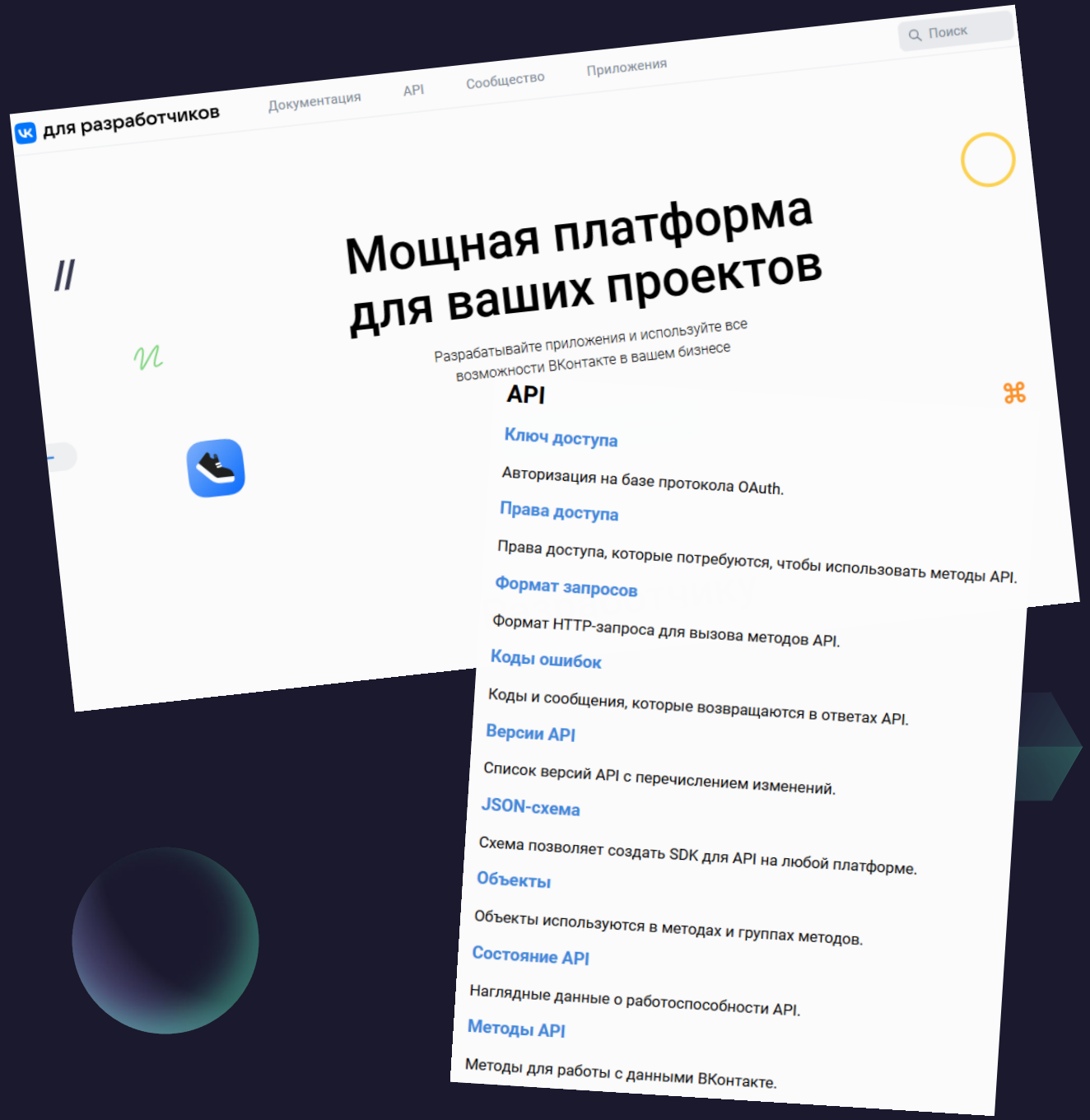
4. Найти ключевые слова в постах группы при помощи TF-IDF
5. Проанализировать активность пользователей по годам
6. Выполнить сентимент-анализ комментариев к постам

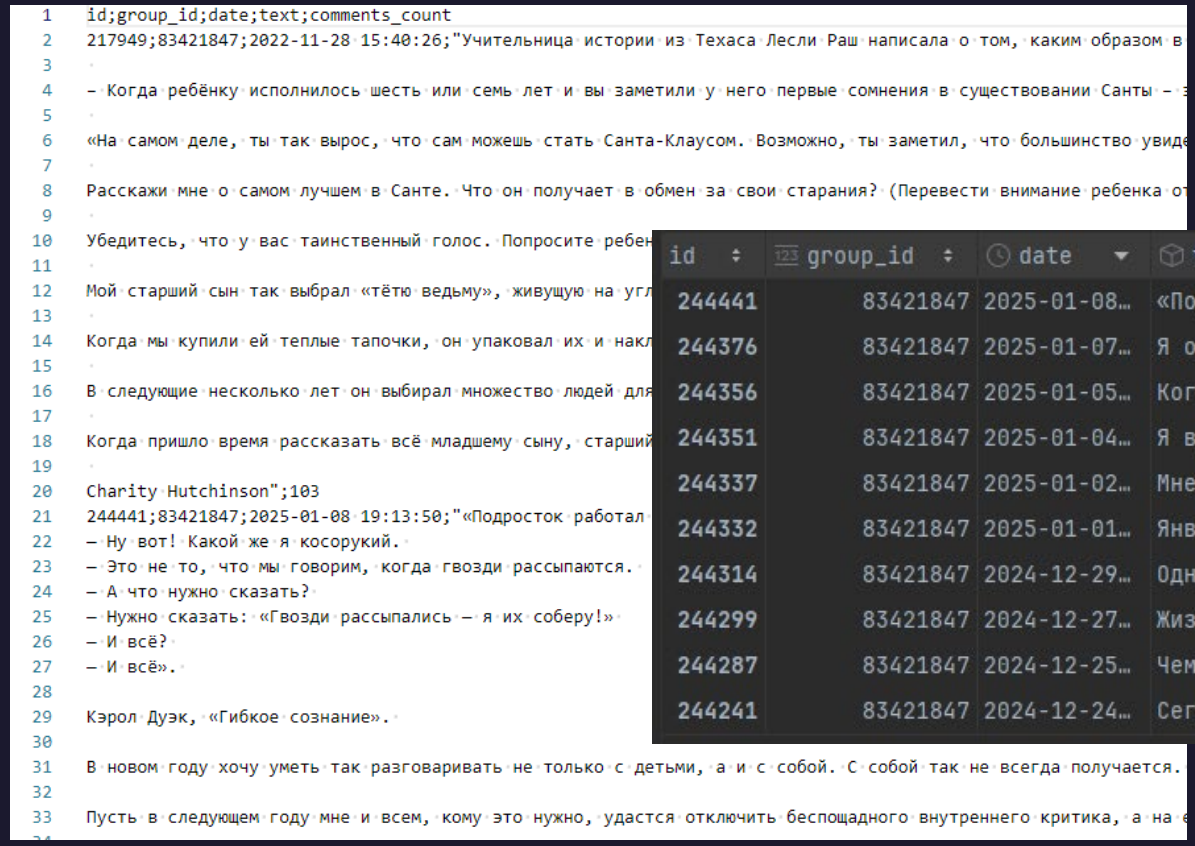


Задача I
Сбор данных

- Извлечение данных с использованием средств разработчика в виде API ВКонтакте:
  a. Создание приложения
  b. Авторизация через Oauth
  c. Получение данных о группе
  d. Получение постов и комментариев

- Сохранение данных в CSV файлы



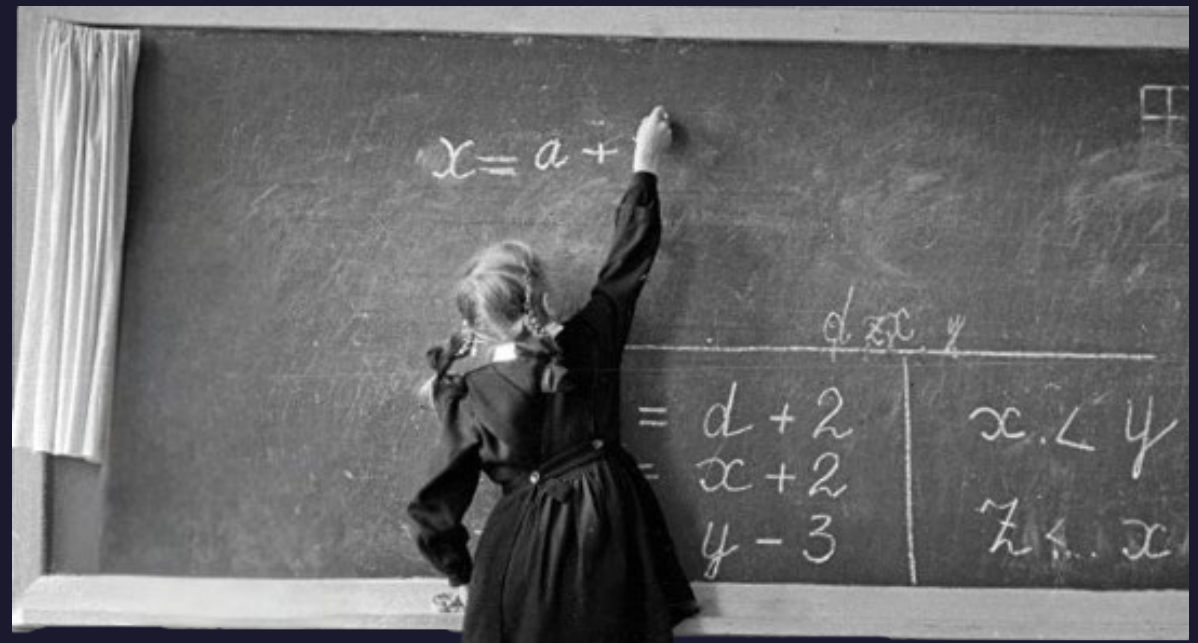


```
id;group_id;date;text;comments_count
217949;83421847;2022-11-28 15:40:26;"Учительница истории из Техаса Лесли Раш написала о том, каким образом в

– Когда ребёнку исполнилось шесть или семь лет и вы заметили у него первые сомнения в существовании Санты – з

«На самом деле, ты так вырос, что сам можешь стать Санта-Клаусом. Возможно, ты заметил, что большинство увиде

Расскажи мне о самом лучшем в Санте. Что он получает в обмен за свои старания? (Перевести внимание ребенка от

Убедитесь, что у вас таинственный голос. Попросите ребен

Мой старший сын так выбрал «тётю ведьму», живущую на угл

Когда мы купили ей теплые тапочки, он упаковал их и накл

В следующие несколько лет он выбирал множество людей для

Когда пришло время рассказать всё младшему сыну, старший

Charity Hutchinson";103
244441;83421847;2025-01-08 19:13:50;"«Подросток работал
– Ну вот! Какой же я косорукий.
– Это не то, что мы говорим, когда гвозди рассыпаются.
– А что нужно сказать?
– Нужно сказать: «Гвозди рассыпались – я их соберу!»
– И всё?
– И всё».

Кэрол Дуэк, «Гибкое сознание».

В новом году хочу уметь так разговаривать не только с детьми, а и с собой. С собой так не всегда получается.

Пусть в следующем году мне и всем, кому это нужно, удастся отключить беспощадного внутреннего критика, а на
```

id	group_id	date	text	comments_count
244441	83421847	2025-01-08…	«Подросток работал вместе с отцом и нечаянно опрокинул коробку…	0
244376	83421847	2025-01-07…	Я очень, очень уважаю родителей, которые в разборках с внешним…	64
244356	83421847	2025-01-05…	Когда я в миноре, спасаюсь вот этой картиной. 📌Для меня это …	16
244351	83421847	2025-01-04…	Я вдруг подумала: а как подвести итоги года в материнстве? 📌📌…	5
244337	83421847	2025-01-02…	Мне было семь, когда прямо 31 декабря скорая увезла меня в бол…	8
244332	83421847	2025-01-01…	Январь начинается с дома. 📌С маминого холодца. С бабушкиного …	0
244314	83421847	2024-12-29…	Однажды моя мама спросила меня о моих самых лучших воспоминани…	18
244299	83421847	2024-12-27…	Жизнь такая интересная… 📌Сначала ты рождаешься и тебе все рав…	18
244287	83421847	2024-12-25…	Чем заняться в длинные зимние выходные? Мы не сильно пьющие, а…	3
244241	83421847	2024-12-24…	Сегодня я видела, как рождается травма. 📌На детском празднике…	39

Задача 2
Структурирование данных

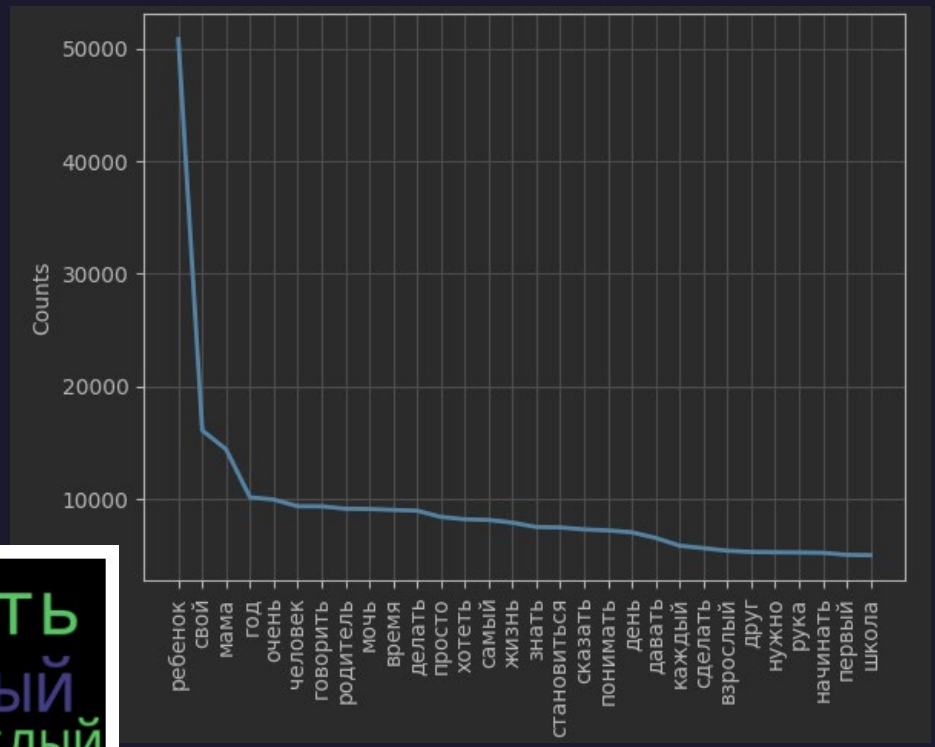
Задача 3
Частотный анализ

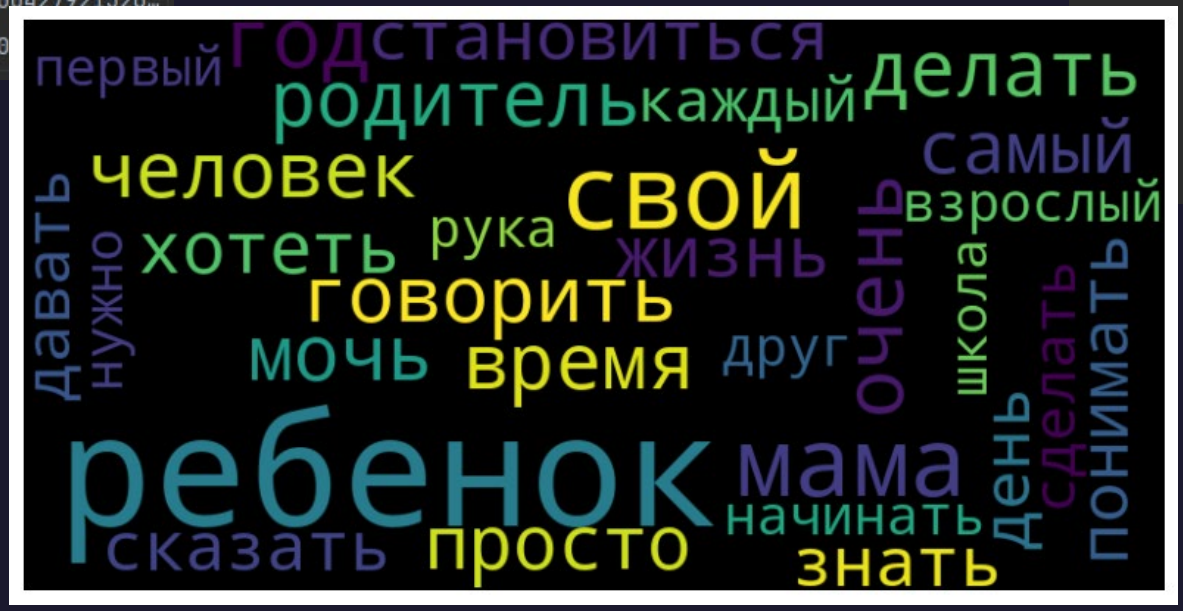


Топ-30 самых частотных слов

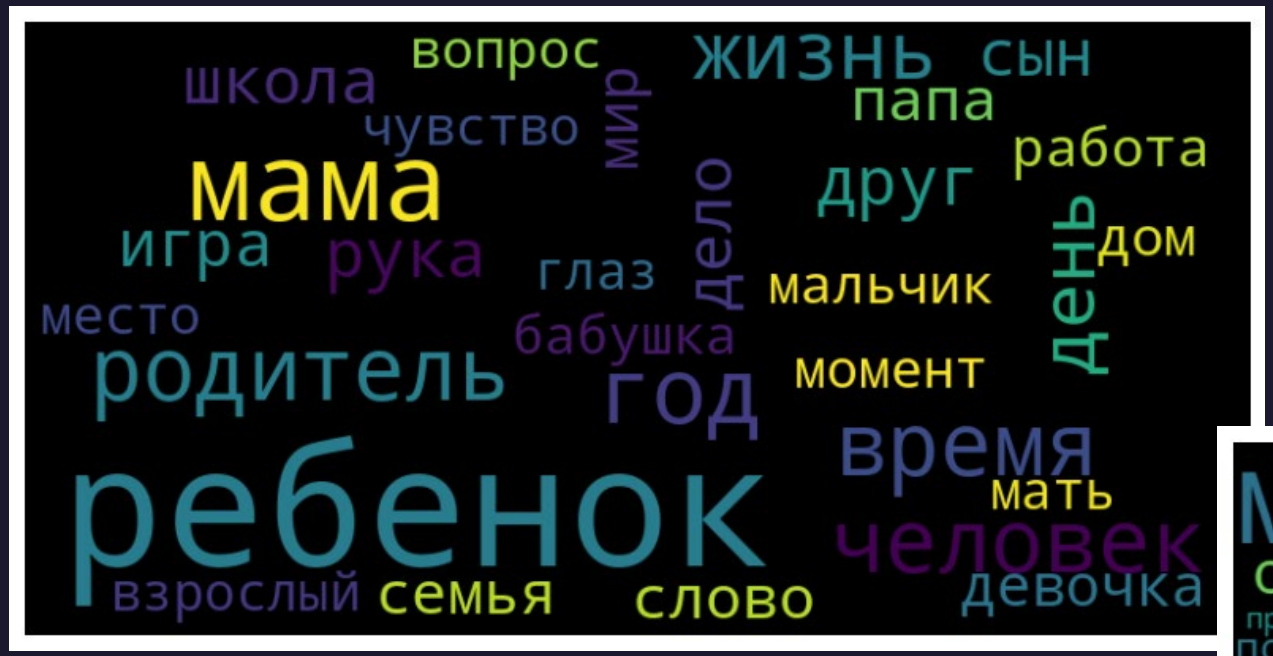
	obj_…	absolute…	relative…
0	ребенок	50799	0.02390365695…
1	свой	16071	0.00756226836…
2	мама	14388	0.00677032650…
3	год	10126	0.00476482667…
4	очень	9924	0.00466977483…
5	человек	9335	0.00439261870…
6	говорить	9326	0.00438838372…
7	родитель	9114	0.00428862634…
8	мочь	9094	0.00427921526…
9	время	9004	0.0…

39	папа	4231



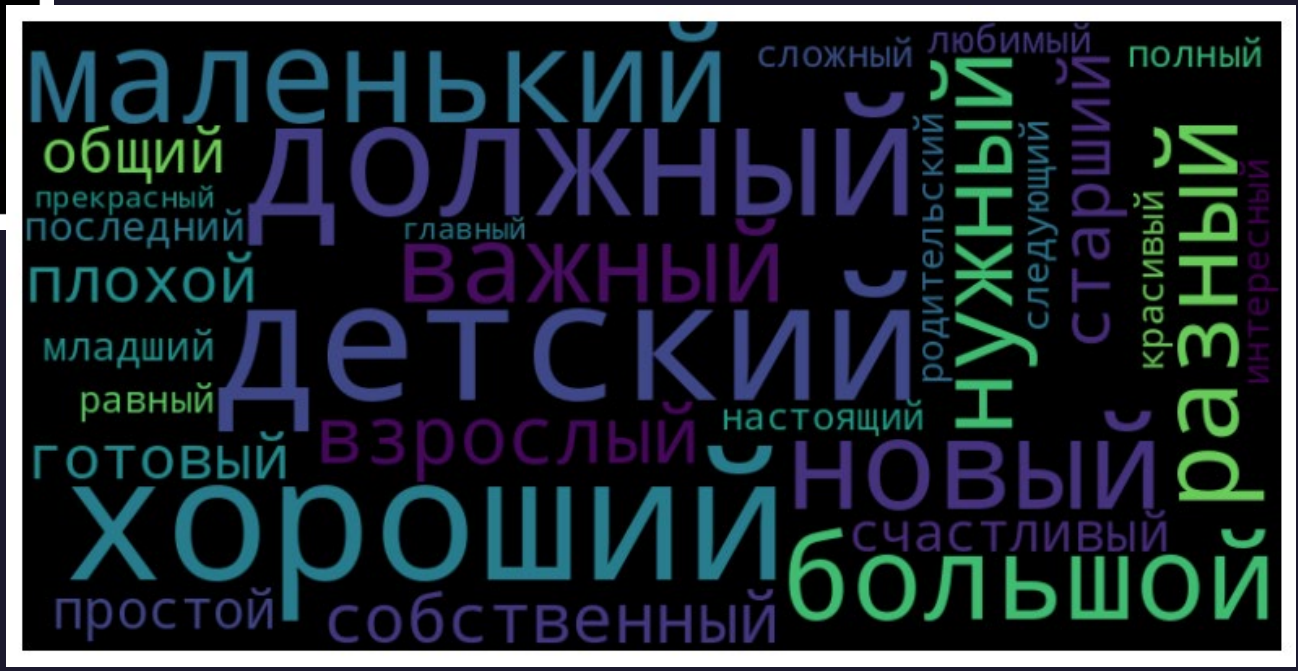


Частота по разным частям речи



самые частотные существительные

самые частотные прилагательные



Биграммы с метриками

	bigram	frequency	pmi	chi_sq	likelihood_ratio	student_t
0	('учительница', 'история')	6	3.77507…	70.65264162…	20.34982960504584	2.270567534441…
1	('история', 'техас')	5	9.58885…	3843.521222…	60.59733374875851	2.233164281809…
2	('техас', 'лесли')	5	18.2269…	1533990.624…	130.45447888896598	2.236060689106…
3	('лесли', 'раш')	5	18.9050…	2454388.0	141.03949069949374	2.236063422254…
4	('раш', 'написать')	5	10.7072…	8354.649172…	74.23435707250113	2.234730557421…
5	('написать', 'образ')	5	2.88329…	27.59848139…	11.370050227620222	1.933008587772…
6	('образ', 'семья')	8	2.08558…	19.88283768…	10.941798298362885	2.162048470080…
7	('семья', 'поколение')	9	3.62008…	93.56576503…	28.803829248081563	2.756012496801…
8	('поколение', 'рассказывать')	5	3.11420…	33.92323109…	12.791647859381829	1.977832932312…
9	('рассказывать', 'правда')	9	2.29648…	28.09808120…	14.36935760585145	2.389322850883…

Триграммы с метриками

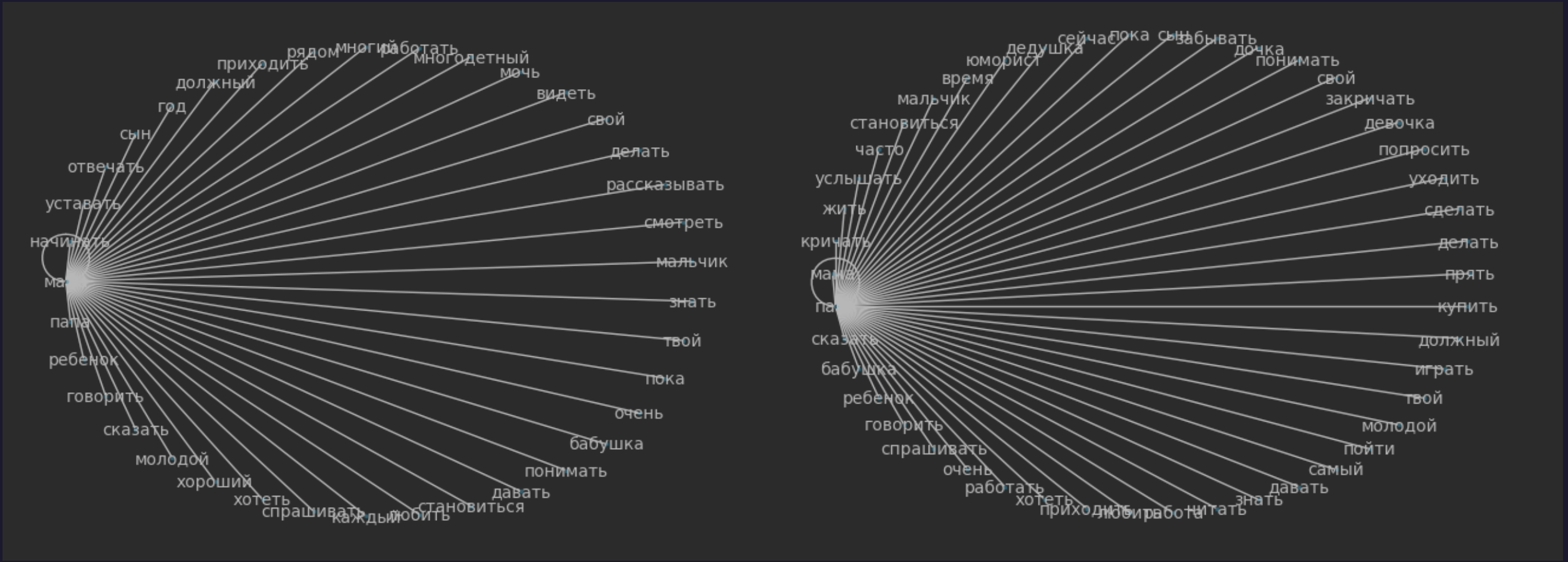
	trigram	frequency	pmi	chi_sq	likelihood_ratio	student_t
0	('учительница', 'история', 'техас')	5	21.7389…	17500611.12…	145.40374001087537	2.236067338645…
1	('история', 'техас', 'лесли')	5	28.4938…	1890066656.…	201.63682444915483	2.236067971584…
2	('техас', 'лесли', 'раш')	5	37.1319…	75300255659…	271.49396958626494	2.236067977484…
3	('лесли', 'раш', 'написать')	5	29.6122…	4103556167.…	215.27384777170593	2.236067974775…
4	('раш', 'написать', 'образ')	5	21.7882…	18109246.22…	152.40954092578403	2.236067360116…
5	('написать', 'образ', 'семья')	5	12.1148…	22170.21490…	80.80235270312305	2.235563822701…
6	('образ', 'семья', 'поколение')	5	13.8530…	73994.20881…	94.24245875645258	2.235916868432…
7	('семья', 'поколение', 'рассказывать')	5	12.3457…	26032.15293…	85.16912331432331	2.235638390291…
8	('поколение', 'рассказывать', 'правда')	5	13.8940…	76113.27960…	89.96049895866295	2.235921099286…
9	('рассказывать', 'правда', 'санта')	5	15.9749…	322038.3344…	103.31003174159312	2.236033260831…

Поиск словосочетаний с ключами «мама» и «папа»

	bigram	freq…	pmi	chi_sq	likelihood_ratio	student_t
0	(обратно, мама)	16	2.803…	82.476521	35.266365	3.426974
1	(приготавливать, мама)	13	2.247…	38.704159	20.215470	2.846270
2	(мама, салат)	7	2.945…	41.079138	16.648022	2.302320
3	(мама, успевать)	12	0.489…	1.404057	1.246882	0.996789
4	(мама, просто)	40	-0.29…	1.710759	1.826685	-1.442765
...	...	...	...	...	...	...
4723	(мама, немецкий)	1	0.581…	0.165631	0.143630	0.331714
4724	(мама, суетиться)	1	1.713…	1.594710	0.995077	0.695168
4725	(доводить, мама)	1	-0.36…	0.065453	0.070958	-0.289674
4726	(мама, дита)	1	2.369…	3.382591	1.692402	0.806549
4727	(мама, напыжиться)	1	7.414…	169.5858…	10.278545	0.994138

	bigram	freq…	pmi	chi_sq	likelihood_ratio	student_t
0	(настаивать, папа)	4	3.558…	39.525332	12.481208	1.830201
1	(папа, заставать)	3	4.498…	62.045115	13.083736	1.655415
2	(папа, везти)	4	2.770…	19.921862	8.573777	1.706945
3	(мандарин, папа)	3	4.765…	75.819062	14.172885	1.668354
4	(папа, заменять)	3	2.787…	15.179864	6.490900	1.481244
...	...	...	...	...	...	...
2013	(папа, привыкать)	1	-0.15…	0.011950	0.012383	-0.115332
2014	(папа, благословлять)	1	6.010…	62.578779	6.475246	0.984485
2015	(папа, второй)	1	-1.88…	1.958806	2.762378	-2.682146
2016	(папа, север)	1	4.050…	14.659999	3.762072	0.939665
2017	(папа, колдовать)	1	4.536…	21.283860	4.412214	0.956904

Визуализация биграмм с ключами «мама» и «папа» с помощью графа



Задача 4
Ключевые слова публикаций

- Подготовка данных – лемматизированные предложения постов
- Рассчёт TF-IDF и получение матрицы
- Вычисление среднего значения для каждого слова и составление списка уникальных слов
- Сортировка индексов по убыванию значений
- Выбор топ-5 ключевых слов

key_words
[рассыпаться, сказать, гвоздь, нужно, получаться]
[уметь, ребенок, хотеть, бегать, лег]
[счастие, детство, удивительно, надежда, сумерки]
[сколько, совместный, анастасия, красильникова...
[санта, готовый, ребенок, сын, человек]
[мультфильм, новогодний, серия, сборник, уолт]
[иллюстратор, зимний, волшебство]
[ребенок, гипертимный, сын, характер, интересный]
[рисовалок, учиться, считать, рисовать, учить]
[ребенок, говорить, мочь, родитель, сейчас]
...

Задача 4
Анализ активности пользователей

	год	количество публикаций	всего комментариев
0	2015	898	2144
1	2016	961	6847
2	2017	969	13373
3	2018	877	22442
4	2019	716	69940
5	2020	600	66759
6	2021	470	63884
7	2022	351	46758
8	2023	204	22171
9	2024	192	12258

Задача 6
Анализ тональности комментариев

- Скачать и сохранить словарь тональности русского языка https://www.labinform.ru/pub/rusentilex/rusentilex_2017.txt в формате csv
- Проверить разметку словаря и, при необходимости, внести изменения
- Предобработать, токенизировать и лемматизировать комментарии
- Найти слова с тэгами “positive” и “negative” и выполнить их частотный анализ
- Анализ тональности всего комментария с помощью NLTK

```
Всего комментариев: 215239
Количество нейтральных комментариев: 212162
Количество положительных комментариев: 89
Количество отрицательных комментариев: 32
```

```
sentiment_words
[(врать, negative), (ребенок, unidentified), (...
[(тайный, unidentified), (санта, unidentified)]
[(год, unidentified), (рассказывать, unidentif...
[(никто, unidentified), (рассказывать, unident...
[(очень, unidentified), (нравиться, positive),...
...
[(добрый, positive), (фото, unidentified)]
[(что-то, unidentified), (боль, negative), (зн...
[(ой, unidentified), (всякое, unidentified), (...
[(сместный, unidentified), (поездка, unidentif...
[(реальность, positive)]
```

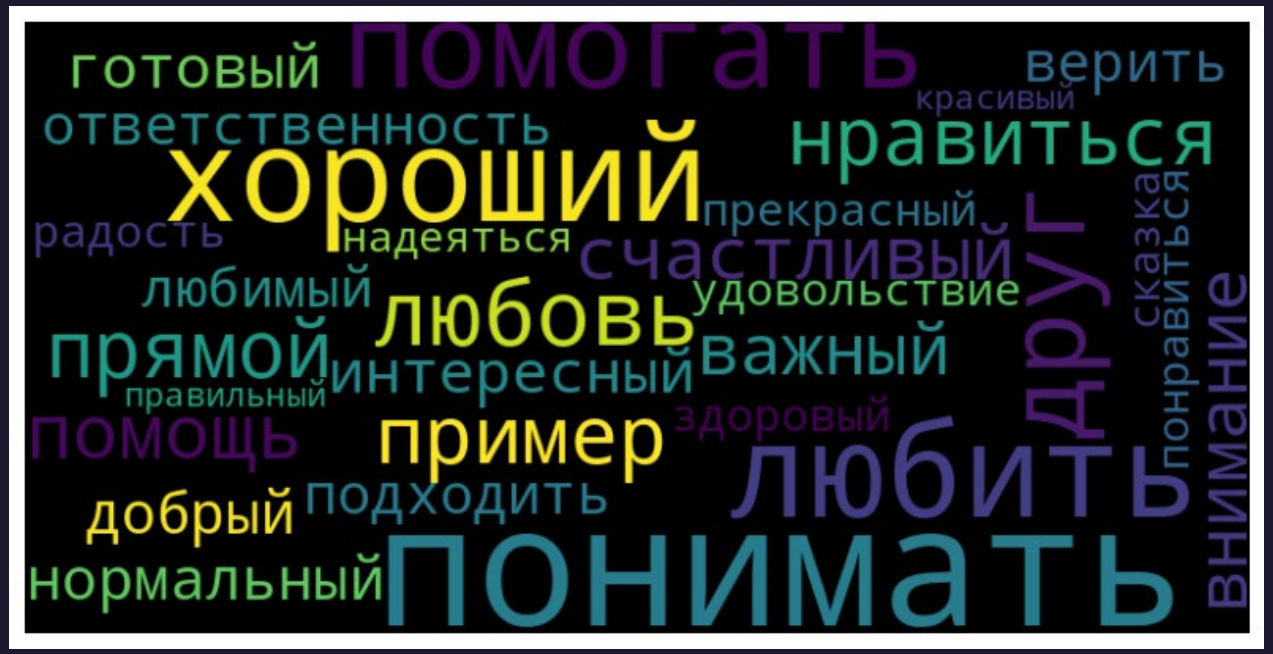
```
Относительная частота нейтральных комментариев: 0.9857042636325202
Относительная частота положительных комментариев: 0.0004134938370834282
Относительная частота отрицательных комментариев: 0.00014867194142325507
```

	lemma	absolute_freq	relative_freq
0	понимать	18020	0.07176566638125012
1	хороший	9806	0.03905294808737729
2	любить	9799	0.0390250701925566…
3	помогать	8490	0.0338119038610884…
4	друг	6574	0.0261813257930265…
5	любовь	4488	0.0178737131364622…
6	пример	3610	0.0143770286146677…
7	нравиться	3499	0.0139349648539397…
8	прямой	3480	0.0138592962822835…
9	помощь	3150	0.0125450526693084…

слова с оценкой “positive”

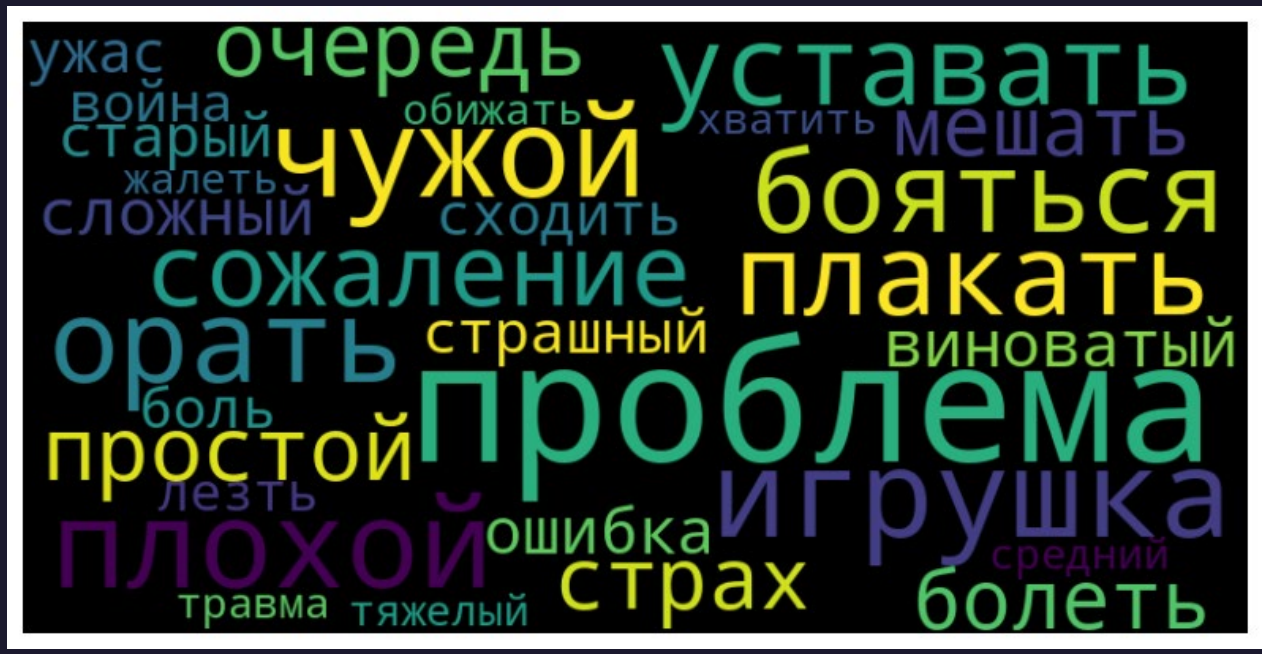
слова с оценкой “negative”

	lemma	absolute_freq	relative_freq
0	проблема	7066	0.028636039424199…
1	чужой	3415	0.013839806769549…
2	игрушка	3304	0.013389962391388…
3	плохой	3199	0.012964433925560…
4	орать	2992	0.012125534950071…
5	бояться	2771	0.011229898845804…
6	плакать	2748	0.011136687848528…
7	уставать	2637	0.010686843470367…
8	сожаление	2096	0.008494358708338…
9	простой	1606	0.006508559201141…



слова с оценкой “positive”

слова с оценкой “negative”



Спасибо за внимание!

Ссылка на GIT
https://github.com/lavrentyukann/KL_25_Project