



# 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 ВКонтakte “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 sentiment-анализ комментариев к постам







Я встала в 6 утра. Вынесла мусор, Г  
мужа, который вернулся с пробеж  
мужа. Накормила собаку. Нарезал  
школу. Разобрала посудомойку и  
Накормила кошку. Разбудила доч

906 58 247



Большой МИР маленьких ДЕТЕЙ

✓ Вы подписаны

здные? Мы не сильно пьющие, а денег  
т нужно слишком много. Всем  
дома в кругу любимейшей семьи  
е возвращаться в пытку. Говорят, число  
ко возрастает. Вроде Показать ещё

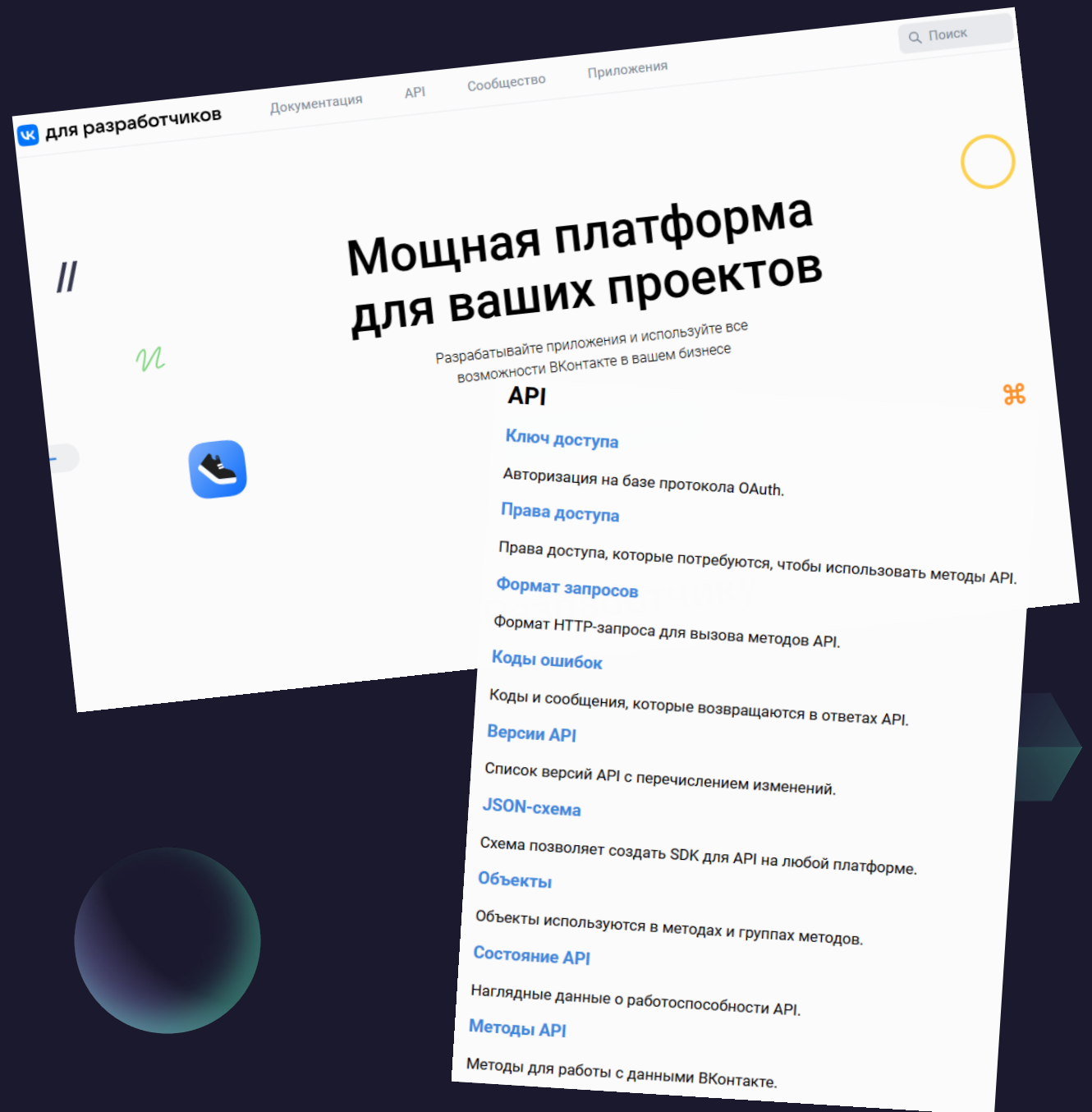
25 дек 2024

# Задача I

## Сбор данных

- ❖ Извлечение данных с использованием средств разработчика в виде API ВКонтakte:
  - а. Создание приложения
  - б. Авторизация
  - с. Получение данных о группе
  - д. Получение постов и комментариев

- ❖ Сохранение данных в CSV файлы



```

1 id;group_id;date;text;comments_count
2 217949;83421847;2022-11-28 15:40:26;"Учительница истории из Техаса Лесли Раш написала о том, каким образом в
3
4 - Когда ребёнку исполнилось шесть или семь лет и вы заметили у него первые сомнения в существовании Санты -
5
6 «На самом деле, ты так вырос, что сам можешь стать Санта-Клаусом. Возможно, ты заметил, что большинство увиде
7
8 Расскажи мне о самом лучшем в Санте. Что он получает в обмен за свои старания? (Перевести внимание ребенка от
9
10 Убедитесь, что у вас таинственный голос. Попросите ребен
11
12 Мой старший сын так выбрал «тётю ведьму», живущую на углу
13
14 Когда мы купили ей теплые тапочки, он упаковал их и накл
15
16 В следующие несколько лет он выбирал множество людей для
17
18 Когда пришло время рассказать всё младшему сыну, старший
19
20 Charity Hutchinson";103
21 244441;83421847;2025-01-08 19:13:50;"«Подросток работал
22 - Ну вот! Какой же я косорукий.
23 - Это не то, что мы говорим, когда гвозди рассыпаются.
24 - А что нужно сказать?
25 - Нужно сказать: «Гвозди рассыпались — я их соберу!»
26 - И всё?
27 - И всё».
28
29 Кэрл Дуэк, «Гибкое сознание».
30
31 В новом году хочу уметь так разговаривать не только с детьми, а и с собой. С собой так не всегда получается.
32
33 Пусть в следующем году мне и всем, кому это нужно, удастся отключить беспощадного внутреннего критика, а на е
34

```

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

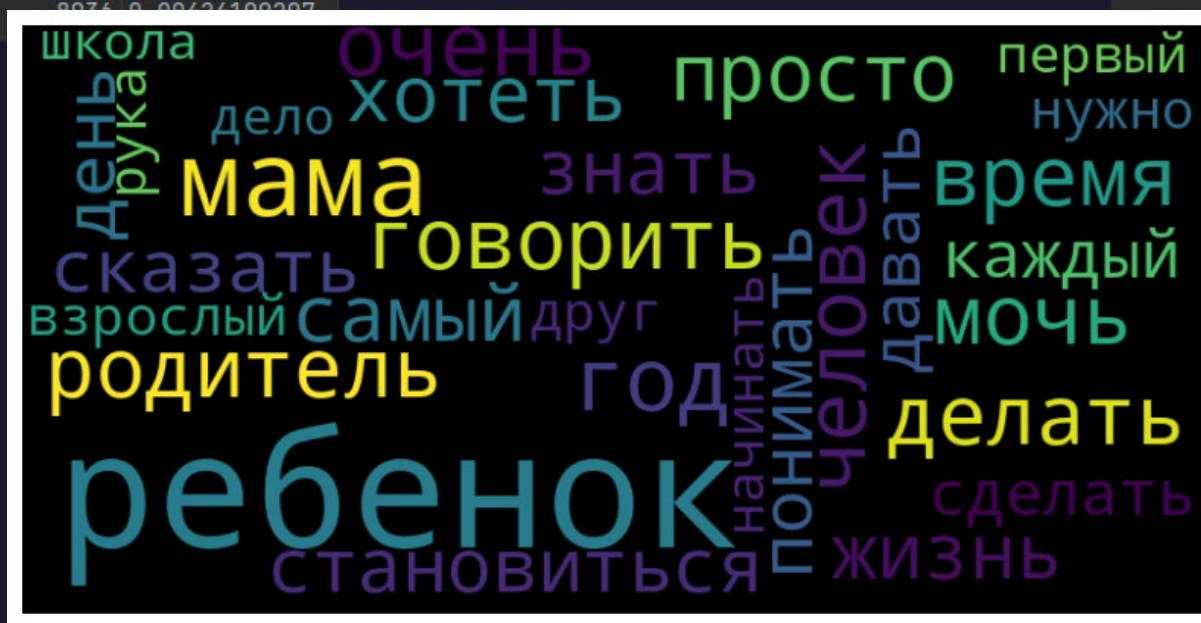
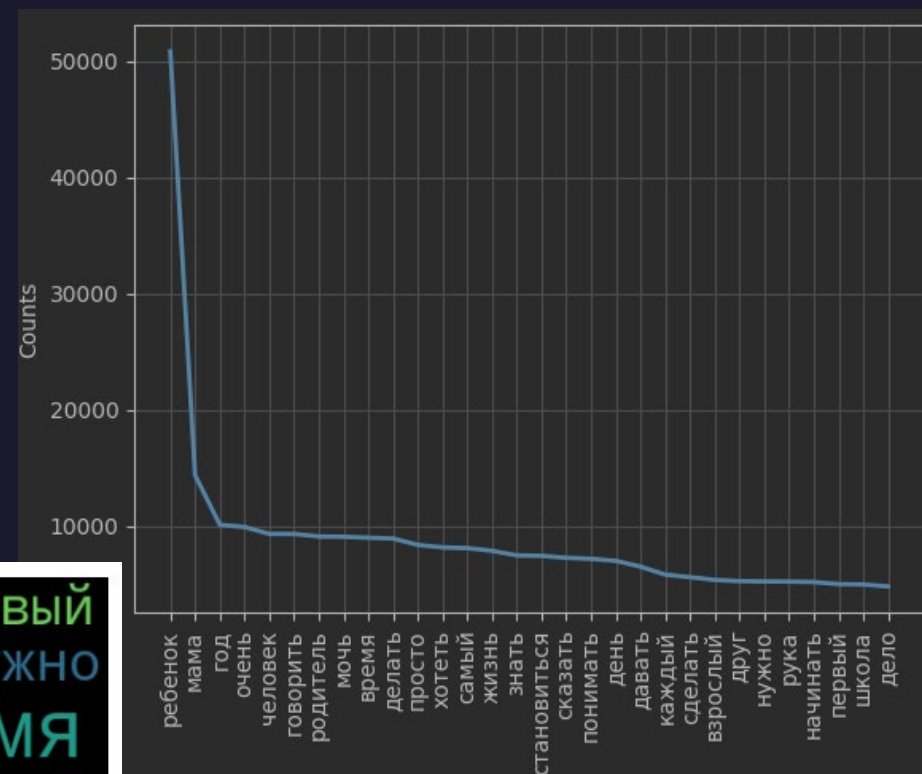




# Топ-30 самых частотных слов

obj_for_freq	absolute...	relative...
0 ребенок	50799	0.02411470470...
1 <b>мама</b>	14388	0.00683010238...
2 год	10126	0.00480689580...
3 очень	9924	0.00471100473...
4 человек	9335	0.00443140157...
5 говорить	9326	0.00442712919...
6 родитель	9114	0.00432649104...
7 мочь	9094	0.00431699688...
8 время	9004	0.00427427313...
9 делать	8976	0.00426100007...

39 <b>папа</b>	4231
----------------	------

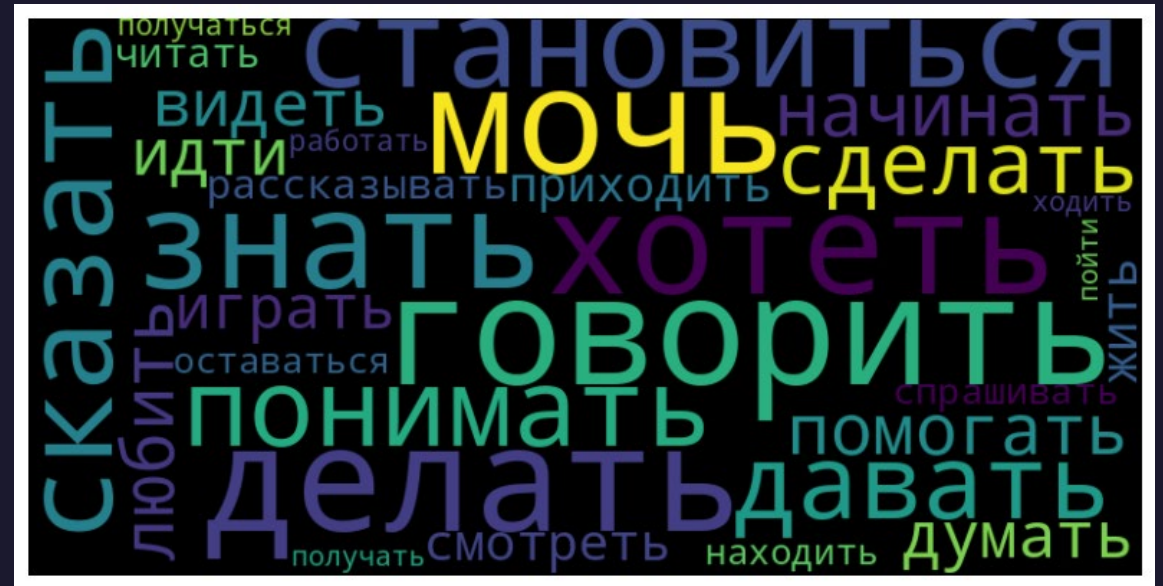


# Частота по разным частям речи

топ-30 прилагательных



топ-30 глаголов





топ-30  
существительных





# Биграммы с метриками



÷  bigram ÷	$\frac{123}{123}$ freq...	$\frac{123}{123}$ pmi ÷	$\frac{123}{123}$ chi_sq ÷	$\frac{123}{123}$ likelihood...	$\frac{123}{123}$ student_t ÷	÷  bigram ÷	$\frac{123}{123}$ freq...	$\frac{123}{123}$ pmi ÷	$\frac{123}{123}$ chi_sq ÷	$\frac{123}{123}$ likelihood...	$\frac{123}{123}$ student_t ÷
0 ('учительница', 'история')	6	3.76417...	70.03733943...	20.265688965760...	2.269209994268...	1... ('идти', 'изгородь')	5	8.85254...	2304.860627...	56.362955929428...	2.231230687467...
1 ('история', 'техас')	5	9.57795...	3814.527128...	60.521783739378...	2.233142250539...	1... ('изгородь', 'никогда')	5	9.48385...	3573.691028...	60.740322539255...	2.232945072404...
2 ('техас', 'лесли')	5	18.2160...	1522439.374...	130.37889199383...	2.236060633807...	1... ('никогда', 'позволять')	17	3.83395...	209.9331998...	59.052487761068...	3.833978264068...
3 ('лесли', 'раш')	5	18.8940...	2435906.0	140.96390380436...	2.236063387692...	1... ('позволять', 'ребенок')	119	2.25553...	362.9026630...	191.88927338127...	8.624223160386...
4 ('раш', 'написать')	5	10.6963...	8291.699581...	74.158770177369...	2.234720409986...	1... ('ребенок', 'заходить')	8	0.50136...	0.996821117...	0.8842847330304...	0.830319231974...
5 ('написать', 'образ')	5	2.87238...	27.32575319...	11.304677213277...	1.930709178993...	1... ('заходить', 'поиграть')	7	6.66743...	697.8594718...	51.095723800308...	2.619722735913...
6 ('образ', 'семья')	8	2.07468...	19.64124805...	10.849263397102...	2.156992441524...	1... ('поиграть', 'мяч')	7	6.77230...	751.5159153...	52.117711925830...	2.621547617565...
7 ('семья', 'поколение')	9	3.60917...	92.73778829...	28.678648935369...	2.754161285369...	1... ('мяч', 'фрисби')	5	12.3907...	26843.69283...	79.499363979701...	2.235651590153...
8 ('поколение', 'рассказыват...	5	3.10330...	33.60149157...	12.724723223928...	1.975873620210...	1... ('часто', 'кричать')	10	2.40083...	34.77182403...	17.137322537613...	2.563483510126...
9 ('рассказывать', 'правда')	9	2.28557...	27.77884353...	14.260897194812...	2.384689447512...	1... ('кричать', 'ребенок')	109	1.92753...	230.2224215...	135.59620823548...	7.695775628240...
40 ('видеть', 'сердце')	7	2.44455...	25.43952022...	12.342606032059...	2.159719584045...						
41 ('сердце', 'становиться')	8	1.67803...	12.14063010...	7.6514319294308...	1.944521441466...						
42 ('приводить', 'пример')	67	7.60940...	12959.96407...	583.57672871638...	8.143436804072...						
43 ('пример', 'сострадание')	5	7.88019...	1168.556295...	45.222332668766...	2.226576989738...						
44 ('сострадание', 'внимание')	10	7.93757...	2433.753092...	92.396107450276...	3.149378808229...						
45 ('внимание', 'чувство')	10	2.08218...	24.75576316...	13.637705013131...	2.415487878114...						
46 ('чувство', 'человек')	30	1.49413...	35.33384176...	23.614804887663...	3.532843017253...						
47 ('человек', 'хороший')	47	1.34946...	44.46648211...	31.072842590281...	4.165230591805...						
48 ('хороший', 'поступок')	12	4.25085...	205.5265652...	48.413748071662...	3.282150217228...						
49 ('поступок', 'ребенок')	14	1.07345...	8.289009009...	6.3140895810812...	1.963700182718...						

# Триграммы с метриками

÷	trigram	÷	frequency	÷	pmi	÷	chi_sq	÷	likelihood_ratio	÷	student_t	÷
0	(учительница, история, тexas)		5		21.738974		1.750061e+07		145.403740		2.236067	
1	(история, тexas, лесли)		5		28.493862		1.890067e+09		201.636824		2.236068	
2	(тexas, лесли, раш)		5		37.131936		7.530026e+11		271.493970		2.236068	
3	(лесли, раш, написать)		5		29.612299		4.103556e+09		215.273848		2.236068	
4	(раш, написать, образ)		5		21.788295		1.810925e+07		152.409541		2.236067	
...	...		...		...		...		...		...	
930108	(обучалок, рисовалок, французский)		1		31.938164		4.114768e+09		44.370749		1.000000	
930109	(рисовалок, французский, художница)		1		26.616236		1.028712e+08		58.948455		1.000000	
930110	(учиться, считать, рисовалок)		1		16.816763		1.155854e+05		72.729030		0.999991	
930111	(книга, адель, фабер)		1		22.663375		7.119397e+06		200.704857		1.000000	
930112	(элейн, мазлиш, говорить)		1		20.573334		2.280323e+06		272.934228		0.999999	

# Поиск словосочетаний с ключами «мама» и «папа»

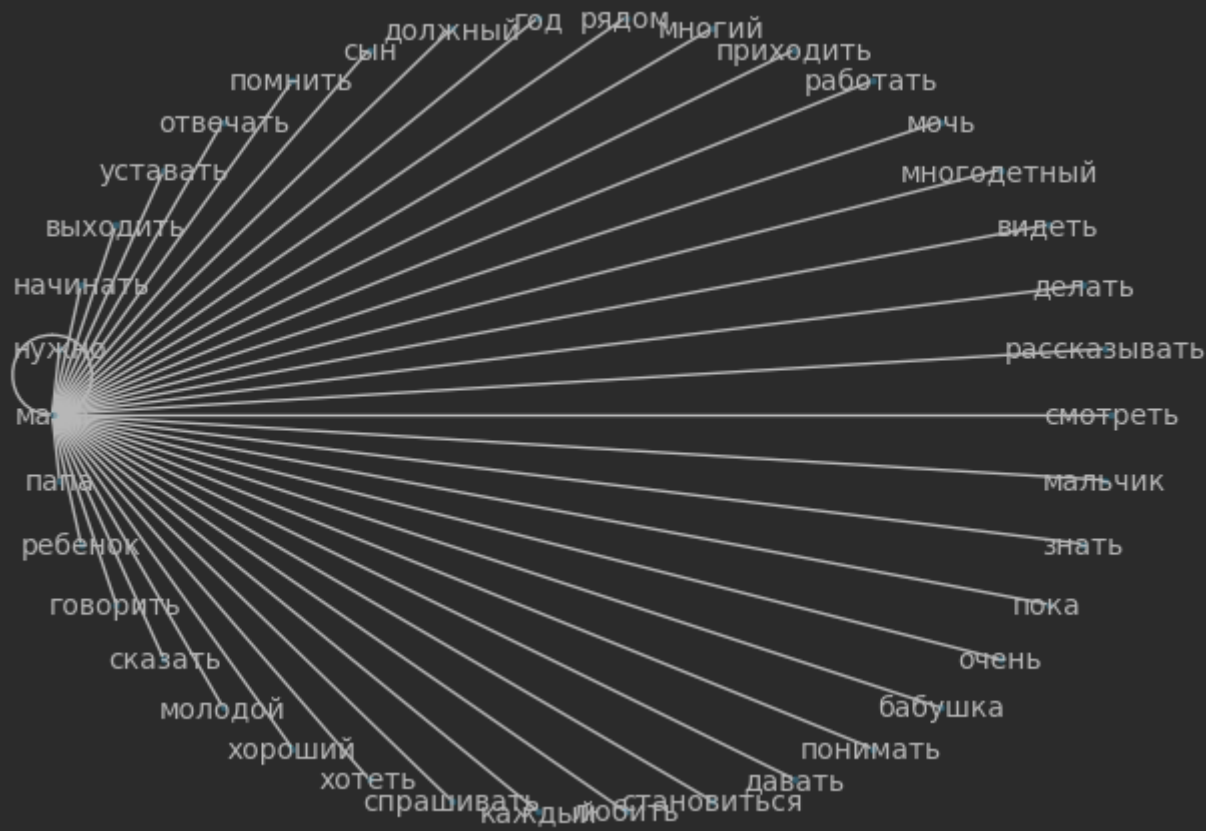
÷	bigram	÷	frequency	÷	pmi	÷	chi_sq	÷	likelihood...	÷	student_t	÷
0	(обратно, мама)		16		2.79...		81.651645		35.058010		3.422627	
1	(приготовливать, мама)		13		2.23...		38.259137		20.059540		2.840510	
2	(мама, салат)		7		2.93...		40.679411		16.555438		2.299714	
3	(мама, успевать)		12		0.47...		1.341701		1.194834		0.978069	
4	(мама, просто)		40		-0.3...		1.839547		1.968763		-1.501698	
...	...		...		...		...		...		...	
4744	(мама, немецкий)		1		0.57...		0.159403		0.138624		0.326644	
4745	(мама, суетиться)		1		1.70...		1.572258		0.984523		0.692855	
4746	(доводить, мама)		1		-0.3...		0.069426		0.075438		-0.299459	
4747	(мама, дита)		1		2.35...		3.345062		1.680148		0.805081	
4748	(мама, напыхиться)		1		7.40...		168.3012...		10.263428		0.994093	

÷	bigram	÷	frequency	÷	pmi	÷	chi_sq	÷	likelihood...	÷	student_t	÷
0	(настаивать, папа)		4		3.547...		39.173029		12.425791		1.828912	
1	(папа, заставать)		3		4.487...		61.535474		13.040321		1.654834	
2	(папа, везти)		4		2.751...		19.565409		8.481855		1.702985	
3	(мандарин, папа)		3		4.754...		75.205541		14.129131		1.667870	
4	(папа, заменять)		3		2.776...		15.027081		6.452069		1.479341	
...	...		...		...		...		...		...	
2022	(особенность, папа)		1		0.860...		0.367470		0.295368		0.449393	
2023	(папа, благословлять)		1		5.999...		62.093513		6.460338		0.984368	
2024	(папа, второй)		1		-1.89...		1.984805		2.803242		-2.710084	
2025	(папа, север)		1		4.039...		14.535625		3.747846		0.939207	
2026	(папа, колдовать)		1		4.525...		21.109433		4.397726		0.956577	

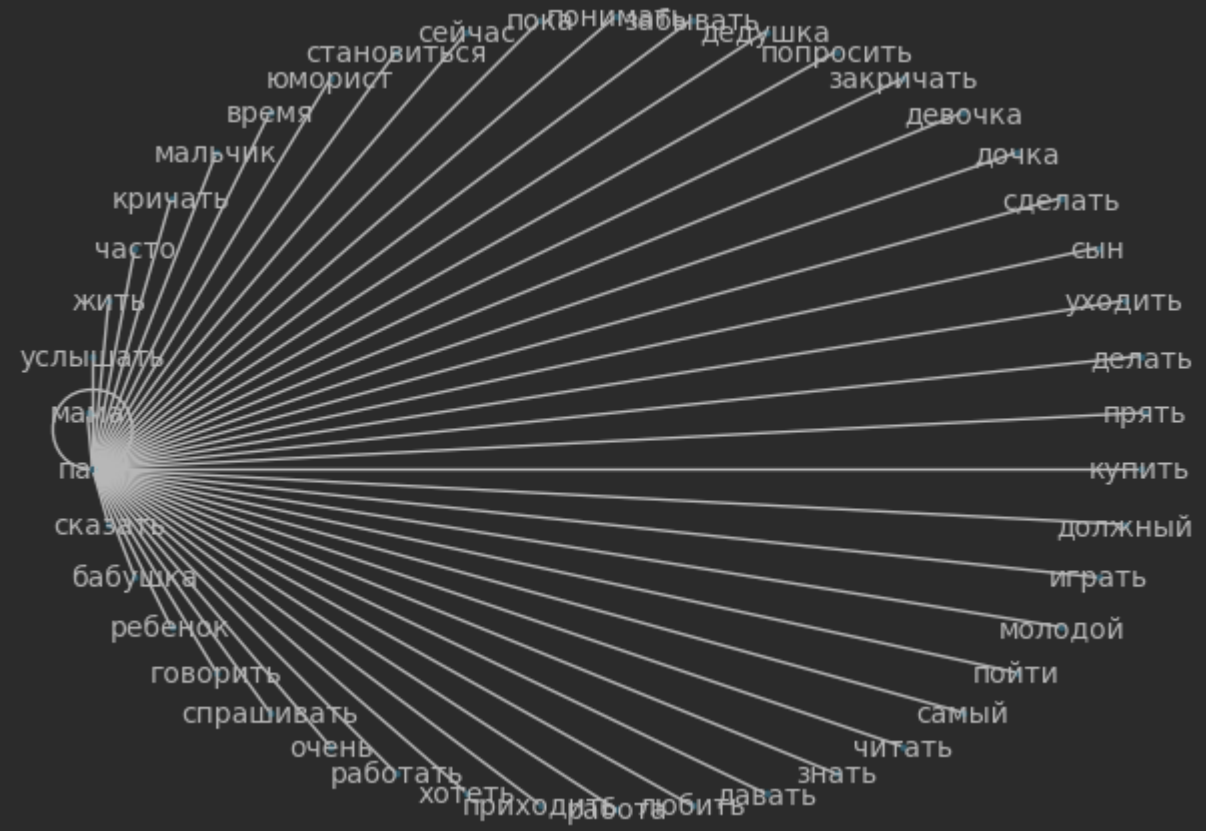


# Визуализация биграмм с ключами «мама» и «папа» с помощью графа

ключ «мама»



ключ «папа»



# 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]
[счастье, детство, удивительно, надежда, сумерки]
[сколько, совместный, анастасия, красильникова...]
[санта, готовый, ребенок, сын, человек]
[мультфильм, новогодний, серия, сборник, уолт]
[иллюстратор, зимний, волшебство]
[ребенок, гипертонный, сын, характер, интересный]
[рисовалок, учиться, считать, рисовать, учить]
[ребенок, говорить, мочь, родитель, сейчас]
...
```

# Задача 5

## 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](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 и приписать словам из комментариев теги “positive”, “negative”, “neutral” по совпадению в словаре 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
```

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


```
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)]
```

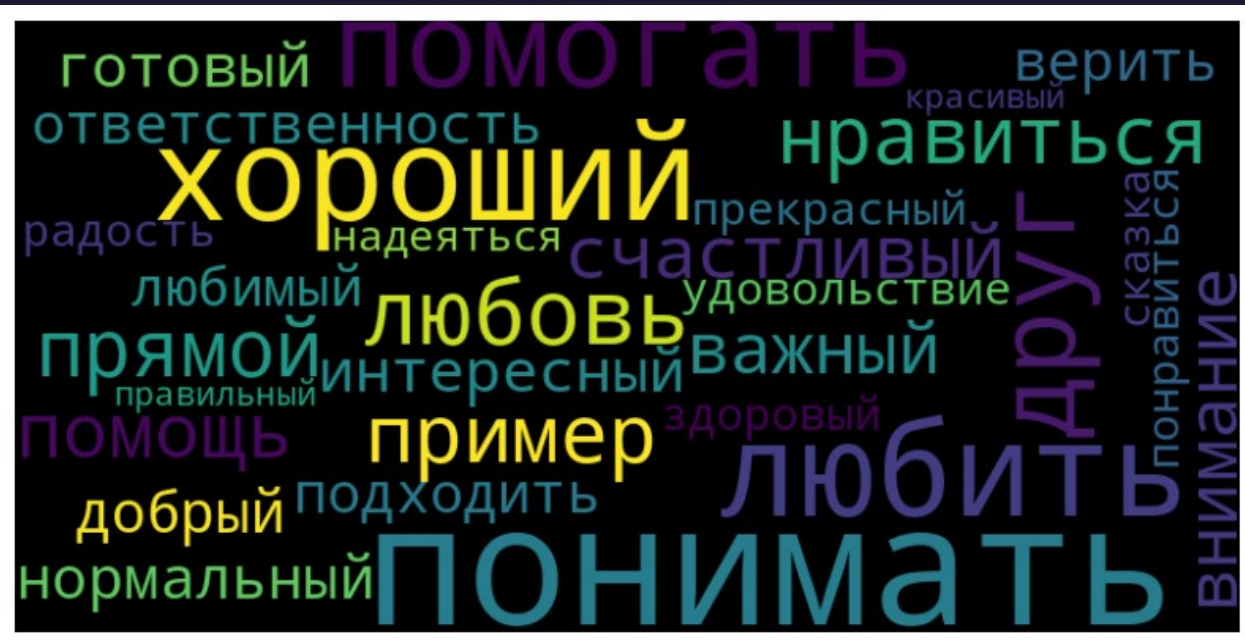
÷  lemma ÷	<u>123</u> absolute_freq ÷	<u>123</u> relative_freq ÷
0 понимать	18019	0.07176254121995125
1 хороший	9808	0.039061379892629
2 любить	9798	0.0390215538527711
3 помогать	8490	0.03381230783935769
4 друг	6573	0.0261776559985981...
5 любовь	4488	0.01787392668822583
6 пример	3610	0.0143772003887021...
7 нравиться	3499	0.0139351313462794...
8 прямой	3480	0.01385946187054944
9 помощь	3150	0.0125452025552387...

слова с оценкой “positive”



слова с оценкой “negative”

÷  lem... ÷	<u>123</u> absolute_freq ÷	<u>123</u> relative_freq ÷
0 проблема	7066	0.028636387584144...
1 чужой	3415	0.013839975035359...
2 игрушка	3304	0.013390125187944...
3 плохой	3199	0.012964591548496...
4 орать	2992	0.012125682373586...
5 бояться	2771	0.011230035380082...
6 плакать	2748	0.011136823249536...
7 уставать	2637	0.010686973402121...
8 сожаление	2096	0.008494461983635...
9 простой	1605	0.006504585631552...

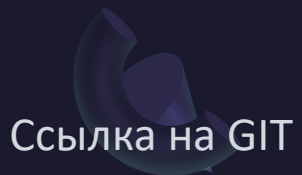


слова с оценкой “positive”





**Спасибо  
за  
внимание!**



Ссылка на GIT

[https://github.com/lavrentyukann/KL\\_25\\_Project](https://github.com/lavrentyukann/KL_25_Project)

