



КРИБРУМ
Мы слушаем сеть

Описание системы

Система поиска по социальным медиа «Крибрум. Публичный поиск»

Москва, 2020

АННОТАЦИЯ

Система «Крибрум. Публичный поиск» предназначена для поиска информации внутри текстов, написанных во всех видах социальных медиа и СМИ:

- социальные сети: Facebook, ВКонтакте, Twitter, Instagram, YouTube Одноклассники, Ответы Mail.ru;
- каналы Телеграм;
- блоги, форумы и тематические порталы;
- интернет-СМИ и информагентства.

Система «Крибрум. Публичный поиск» осуществляет поиск только в документах, содержащих тексты (тексты с картинками или только тексты).

Система «Крибрум. Публичный поиск» способна находить тексты на заданных платформах, написанные на любом языке.

ОГЛАВЛЕНИЕ

1	Общие сведения о системе	Ошибка! Закладка не определена.
1.1	Обозначение и наименование программы.....	5
1.2	Разработчик программы	5
1.3	Минимальные программные средства	5
1.4	Назначение программы	5
2	Структура интерфейса системы.....	6
2.1	Авторизация.....	6
2.2	Раздел Поиск.....	7
2.2.1	Область поисковых запросов	7
2.2.2	Область выдачи результатов поиска	8
2.2.3	Область подробного рассмотрения найденной публикации с комментариями	9
2.3	Раздел Рейтинг авторов	9
2.3.1	Область поисковых запросов	9
2.3.2	Область выдачи результатов рейтинга	9
2.4	Раздел Статистика фраз	10
2.4.1	Область поисковых запросов	10
2.4.2	Область выдачи результатов статистики	10
3	Техническая архитектура Системы	12
3.1	Подсистема сбора информации	12
3.2	Механизм обработки сообщения в Системе	12
4	Функциональное описание разделов Системы	14
4.1	Сбор текстовой информации	14
4.1.1	Строка поисковых запросов	14
4.1.2	Расширенная настройка поисковых запросов	14
4.1.3	Область отображения результатов поиска	15
4.1.4	Область подробного рассмотрения найденной публикации с комментариями	16
4.2	Обзор рейтинга авторов.....	17
4.2.1	Настройки показа рейтинга	18
4.2.2	Отображение результатов выдачи	18
4.3	Обзор статистики фраз	19
4.3.1	Настройки показа статистики	19
4.3.2	Отображение результатов выдачи	20

Перечень терминов и сокращений

Автор	— пользователь, от имени которого опубликовано сообщение на интернет ресурсе (площадке мониторинга).
Источник сообщения	— площадка мониторинга, на которой обнаружено релевантное сообщение по отношению к объекту мониторинга.
Настройка поиска	— расширенный набор параметров поиска, включающий в себя поиск по имени автора, по платформе и периоду, за который осуществляется поиск.
Оригинал	— сообщение, опубликованное первым в ряду одинаковых или частично повторяющихся.
Перепечатки	— ряд одинаковых или частично повторяющихся друг друга сообщений.
Платформы	— социальные сети, по которым осуществляет поиск система «Крибрум. Публичный поиск» (Facebook, ВКонтакте, Instagram, Twitter, Одноклассники и др.).
Поисковый запрос	— последовательность символов, которую пользователь вводит в поисковую строку, чтобы найти интересующую его информацию. Чаще всего, поисковый запрос задаётся в виде набора слов или фразы, иногда используя расширенные возможности языка запросов поисковой системы.
Пост	— любая статья или запись в социальной сети.
Рейтинг	— авторитетность автора, определяется как величина регулярной аудитории данного пользователя, представляющая собой совокупность всех уникальных читателей (подписчиков, фолловеров, друзей) всех аккаунтов автора на всех площадках; в расчетах также учитывается авторитетность читателей автора.
Репост	— функция в социальных сетях, которая позволяет скопировать информацию и опубликовать ее у пользователя на странице или разослать друзьям.
Система	— сервис автоматизированного мониторинга социальных медиа «Крибрум.Публичный поиск».
Сообщение	— это отдельная текстовая публикация в Интернете, содержащая осмысленное упоминание объекта мониторинга на одной из площадок мониторинга. Сообщение может иметь вид записи, комментария, поста, новости, заметки, статьи, рецензии в блоге, микроблоге, социальной сети, форуме, онлайн-СМИ, интернет-магазине или другом виде информационных интернет-ресурсов и социальных медиа.

1 ОБЩИЕ СВЕДЕНИЯ О СИСТЕМЕ

Обозначение и наименование программы

Наименование: «Крибрум. Публичный поиск».

Обозначение: Система поиска по социальным медиа «Крибрум. Публичный поиск».

Разработчик программы

Разработчиком системы является АО «Крибрум».

Минимальные программные средства

Система мониторинга и анализа поведения аккаунтов в социальных медиа «Крибрум. Публичный поиск» является мульти-платформенной. Для использования системы достаточно наличие любого современного браузера.

Назначение программы

Система «Крибрум. Публичный поиск» предназначена для быстрого и легкого поиска текстовых материалов по социальным, а также для определения рейтинга авторов и количества упоминаний объектов мониторинга.

Задачи системы:

- Быстрый поиск текстовых данных по сайтам и платформам социальных сетей в сети Интернет с помощью графического интерфейса поиска и выдачи результатов поиска;
- Определение рейтинга авторов;
- Сбор статистики по количеству упоминаний объектов мониторинга за определенный период.

Функции системы:

- сбор информации с сайтов и платформ социальных сетей в сети Интернет;
- индексирование, обеспечивающее быстрый поиск по накопленной информации;
- подсчет рейтинга авторов;
- подсчет количества упоминаний объектов мониторинга.

Данные функции системы позволяют пользователю легко находить нужную информацию в социальных медиа.

2 СТРУКТУРА ИНТЕРФЕЙСА СИСТЕМЫ

2.1 Авторизация

Система представляет собой онлайн-сервис, доступный через интернет-браузер, с авторизацией пользователей по логину и паролю или через сервисы популярных социальных сетей.

Установка дополнительного клиентского программного обеспечения для работы с Системой не требуется.

Вход в систему «Крибрум. Публичный поиск» осуществляется по ссылке <https://pubsearch.kribrum.ru>.

Для входа в систему требуется авторизация с помощью кнопки **Войти**, которая расположена в верхнем правом углу окна (рис. 1):

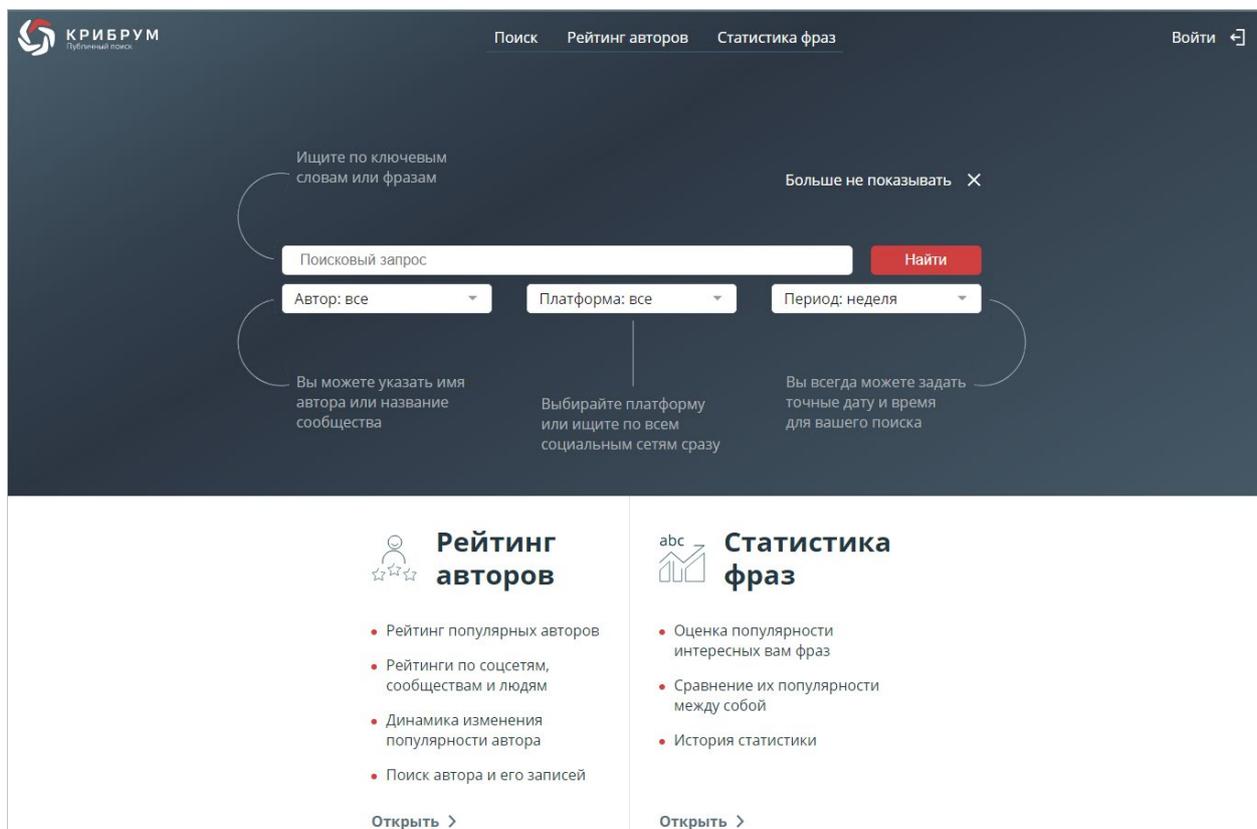


Рисунок 1 – Стартовая страница в системе «Крибрум. Публичный поиск»

Войти можно через популярные сервисы социальных сетей или электронную почту (рис. 2):

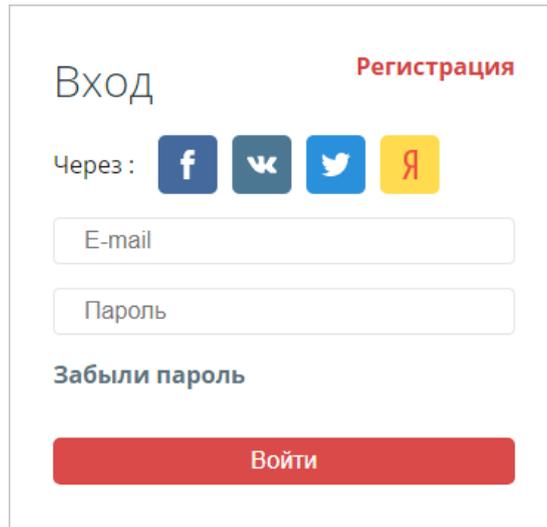


Рисунок 2 – Страница авторизации

2.2 Раздел Поиск

2.2.1 Область поисковых запросов

В разделе **Поиск** имеется две версии параметров поиска (рис. 3):

- сокращенная (использование только поисковой строки в формате Sphinx*), в строку можно вводить ключевые слова и фразы;
- расширенная (поиск по авторам, платформам и периоду).

* Sphinx (англ. SQL Phrase Index) — система полнотекстового поиска, распространяемая по лицензии GNU GPL. Отличительной особенностью является высокая скорость индексации и поиска, а также интеграция с существующими СУБД (MySQL, PostgreSQL) и API для распространённых языков веб-программирования (официально поддерживаются PHP, Python, Java; существуют реализованные сообществом API для Perl, Ruby, .NET[1] и C++). Описание формата поисковых запросов Sphinx по ссылке: <http://sphinxsearch.com/docs/manual-2.3.2.html#extended-syntax>

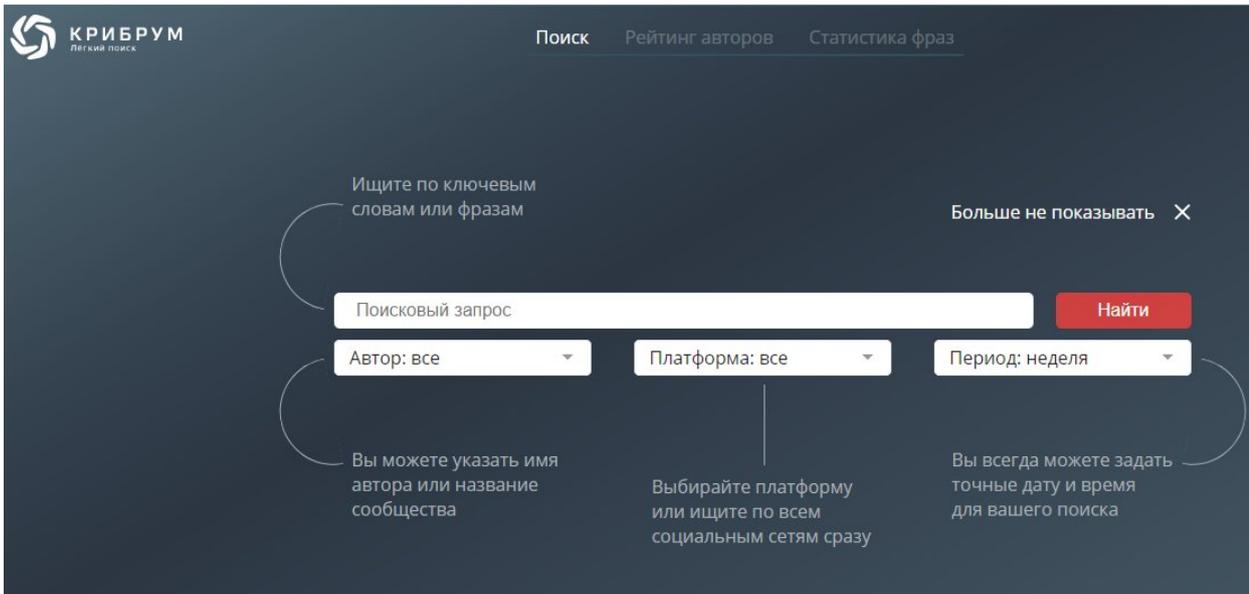


Рисунок 3 – Поисковая область в разделе Поиск

2.2.2 Область выдачи результатов поиска

Результаты поиска выводятся в области отображений результатов поиска в виде списка публикаций с дополнительными статистическими параметрами (рис. 4).

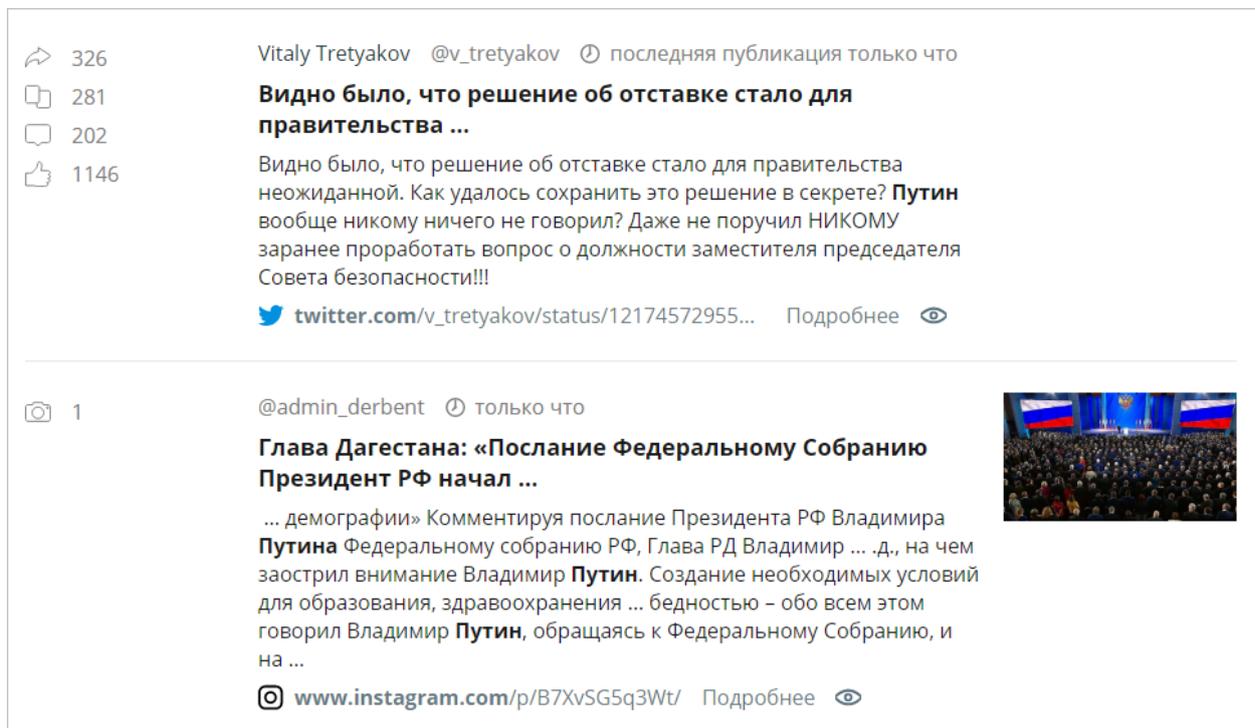


Рисунок 4 – Структура поля публикации

Поле публикации содержит кнопку **Подробнее** для перехода в область подробного рассмотрения найденной публикации с комментариями (см. раздел 4.2.4).

2.2.3 Область подробного рассмотрения найденной публикации с комментариями

В области подробного рассмотрения найденной публикации с комментариями отображается полный текст публикации с дополнительными и статистическими параметрами (рис. 5).

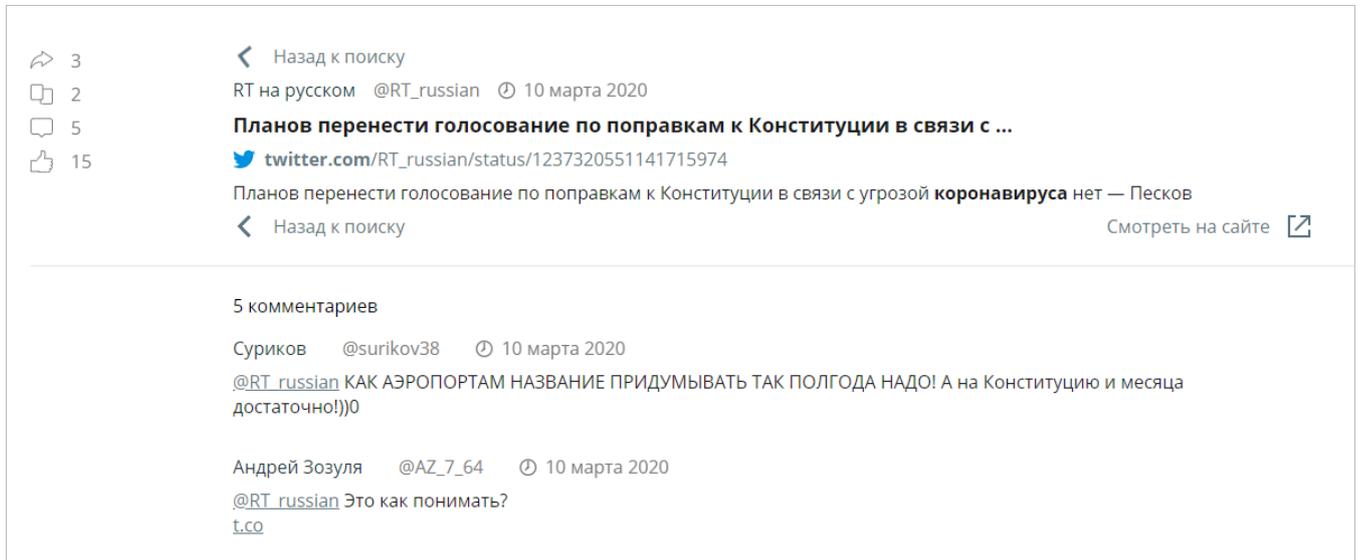


Рисунок 5 – Структура области подробного рассмотрения найденной публикации с комментариями

2.3 Раздел Рейтинг авторов

2.3.1 Область поисковых запросов

Область поисковых запросов в разделе **Рейтинг автора** содержит (рис. 6):

- значки социальных сетей (при выборе которых можно отфильтровывать определенные платформы для просмотра рейтинга).
- строку поиска автора (для просмотра рейтинга и количества подписчиков) определенного автора.

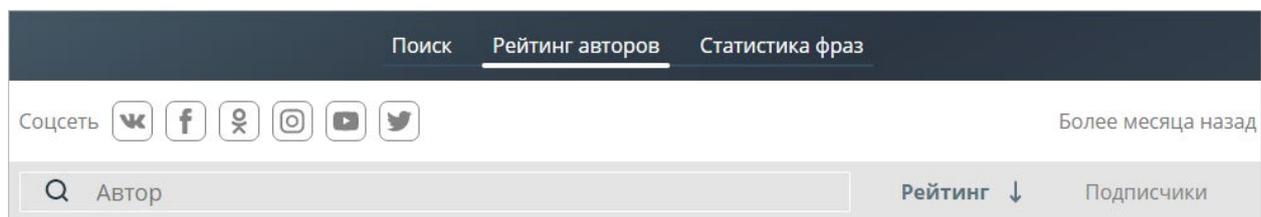


Рисунок 6 – Строка поисковых запросов рейтинга

2.3.2 Область выдачи результатов рейтинга

Область выдачи результатов рейтинга отображает (рис. 7):

- название аккаунта автора

- количество откликов на сообщение автора
- рейтинг автора
- количество подписчиков (с указанием, на сколько увеличилось или уменьшилось количество подписчиков)

сбербанк	Рейтинг	Подписчики ↓
54 704 ↑	85	684 тыс. ↑ 1 758
9 849 ↑	383	435 тыс. ↑ 5 798
62 895 ↑	69	34 тыс. ↑

Рисунок 7 – Область выдачи результатов рейтинга

2.4 Раздел Статистика фраз

2.4.1 Область поисковых запросов

Область поисковых запросов в разделе Статистика фраз содержит (рис. 8):

- поисковую строку для ввода слова или фразы, для которой нужно найти количество упоминаний этого слова/фразы в социальных сетях;
- значки платформ (для поиска по определенной платформе – Вконтакте, Facebook, Одноклассники, Instagram, YouTube, Twitter);
- знак «+» для добавления еще одной платформы в результаты выдачи;
- кнопку **Показать** для показа результата выдачи.

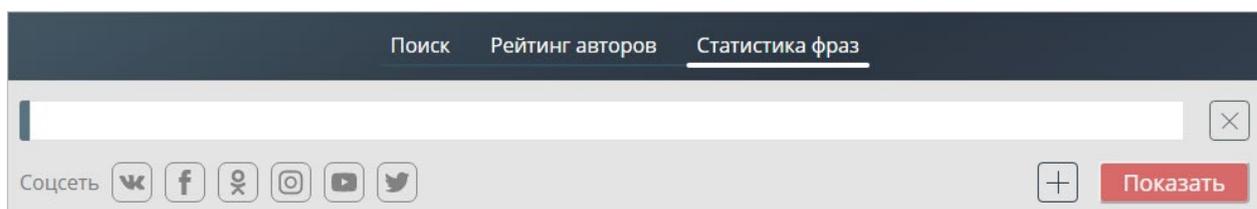


Рисунок 8 – Область поисковых запросов в разделе Статистика фраз

2.4.2 Область выдачи результатов статистики

В области выдачи раздела **Статистика фраз** отображается график или несколько графиков (в зависимости от количества вводимых слов или фраз). По вертикали указывается количество упоминаний слова или фразы в социальных сетях, по горизонтали – временной промежуток за последние несколько месяцев (рис. 9).

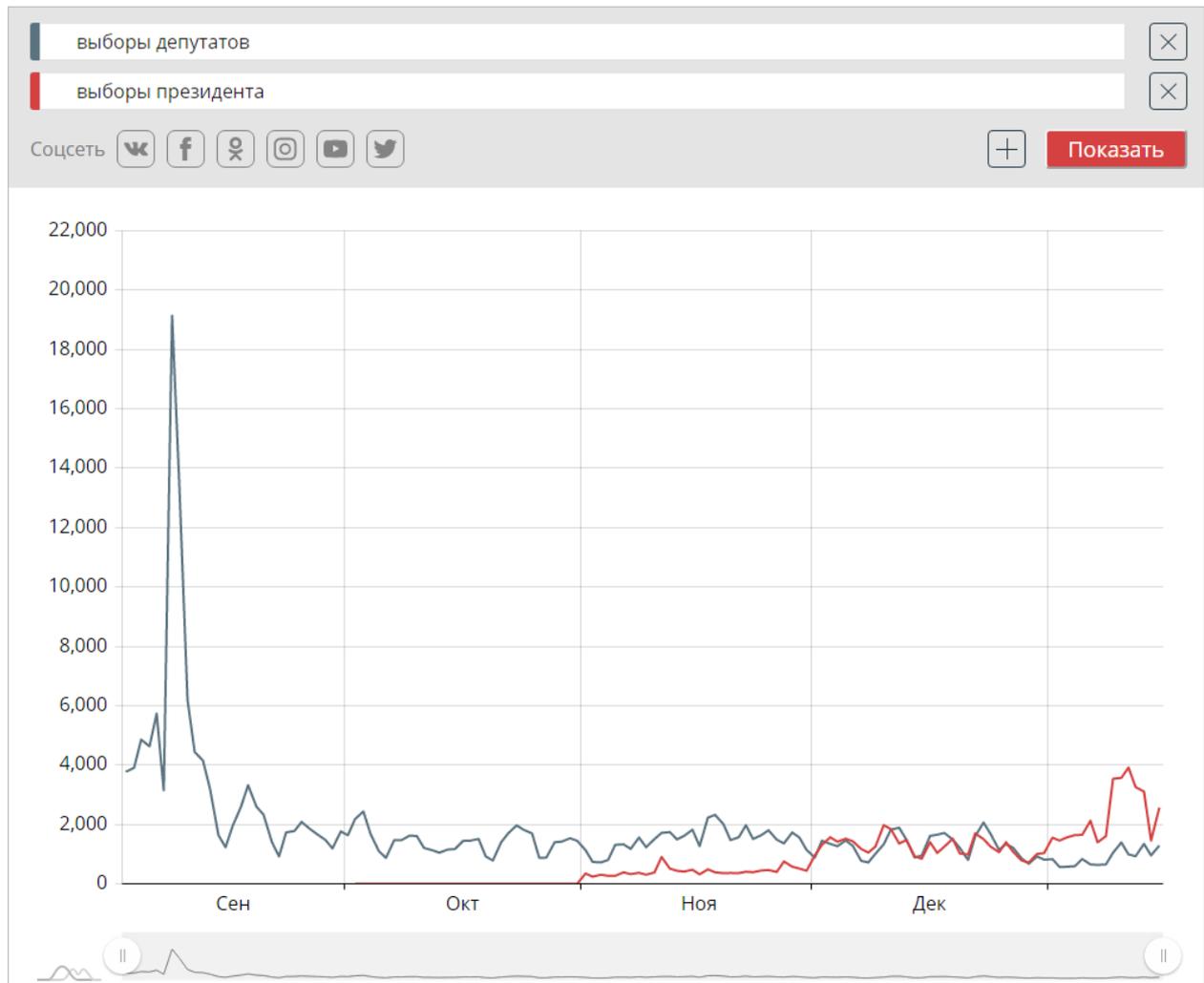


Рисунок 9 – Область выдачи результатов в разделе Статистика фраз

3 ТЕХНИЧЕСКАЯ АРХИТЕКТУРА СИСТЕМЫ

3.1 Подсистема сбора информации

Подсистема сбора информации обеспечивает постоянный мониторинг русскоязычных текстовых публикаций, представленных в форме пользовательского и редакционного контента, в том числе в виде новостей, статей, комментариев, постов, твитов, записей и т.п., в открытых Интернет-источниках, относящихся ко всем типам социальных медиа, включая, но не ограничиваясь социальные сети (не менее 220 миллионов аккаунтов):

- ВКонтакте;
- Facebook;
- Одноклассники;
- Instagram;
- YouTube;
- Twitter;
- Livejournal (доступна только в разделе **Поиск**);
- Ответы Mail.ru (доступна только в разделе **Поиск**);
- Телеграм (доступна только в разделе **Поиск**).

Подсистема собирает и обрабатывает около 95 000 000 (девяносто миллионов) информационных сообщений из социальных медиа в сутки.

Основная часть информации извлекается в период времени от 5 (пяти) секунд до 30 (тридцати) минут с момента публикации.

Добавление новых источников осуществляется по запросу Пользователя в рамках сопровождения Системы.

Система позволяет сортировать данные по источнику данных и выделять из общего массива данные, поступившие из определенных источников.

Мониторинг информационного поля (интернет ресурсов) осуществляется с помощью поисковых роботов - спайдеров.

Спайдер - программа, которая выполняет задачу просмотра страниц Интернета и собирает информацию о них в виде документов.

Каждый спайдер работает по своей области интернета (пр. социальные сети, сайты, форумы, блоги и проч.). Система индексирует только публичные сообщения и открытые данные пользователей.

После сбора сообщения спайдерами ему присваивается идентификатор автора. В случае, если автора нет в базе - присваивается новый идентификатор и добавляется в базу.

Затем сообщение отправляется в базу данных.

3.2 Механизм обработки сообщения в Системе

Система «Крибрум. Публичный поиск» осуществляет поиск по сообщениям социальных сетей, собранных подсистемой сбора информации Крибрум.

В качестве СУБД во всех системах Крибрум используется NoSQL-система Cassandra.

СУБД Cassandra 1 представляет собой хранилище всех текстовых данных подсистемы сбора информации, которые скачивает поисковая система.

Все публикации, найденные подсистемой сбора информации, сохраняются в базе данных системы для последующего автоматического анализа содержания, составления отчетов, ретроспективного анализа.

Ограничения по объему собираемых и хранимых данных отсутствуют. Также нет никаких ограничений по срокам хранения данных в системе.

СУБД Cassandra 2 системы «Крибрум. Публичный поиск» представляет собой хранилище текстовых данных, которые являются результатом поискового запроса, сформулированного пользователем Системы.

Далее сообщения добавляются в очередь (query). Система использует для этого платформу обмена сообщениями RabbitMQ.

Задача платформ обмена сообщениями RabbitMQ – принимать и отдавать сообщения. После того, как сообщения оказались в базе данных, они отправляются в очередь RabbitMQ. Очередь не имеет ограничений на количество сообщений. Любое количество поставщиков может отправлять сообщения в одну очередь, также любое количество подписчиков может получать сообщения из одной очереди.

Очередь обмена сообщениями RabbitMQ 1 – очередь подсистемы сбора информации.

Очередь обмена сообщениями RabbitMQ 2 – очередь системы «Крибрум. Публичный поиск».

В системе «Крибрум. Публичный поиск» все найденные документы проходят проверку идентификатором дублей. Дублями считаются два схожих документа.

Дубли получают один и тот же идентификатор, в интерфейсе в поле сообщения отображается счетчик "Количество похожих сообщений" (рис. 10).

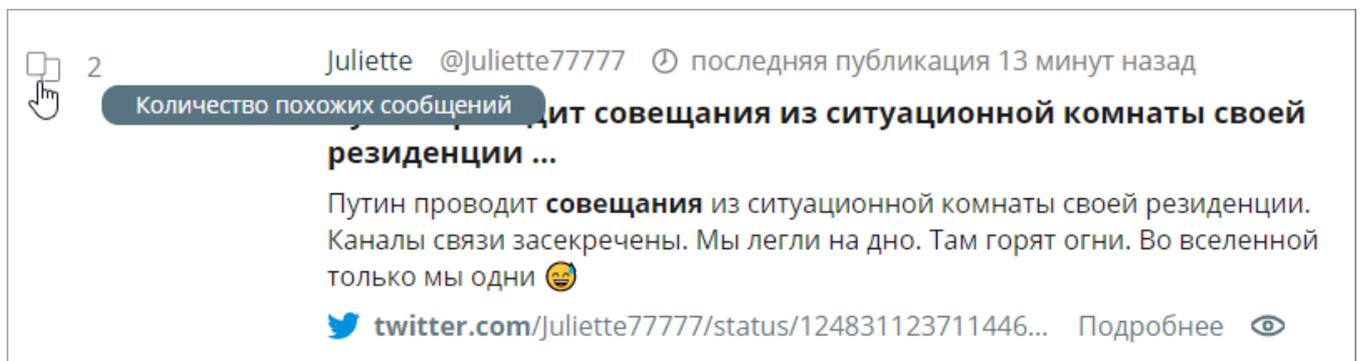


Рисунок 10 – Счетчик "Количество похожих сообщений"

При создании пользователем поискового запроса проводится поиск по документам, запрашиваемым через очередь обмена сообщениями RabbitMQ 2.

Результаты поискового запроса сохраняются в СУБД Cassandra 2 системы «Крибрум. Публичный поиск» и отображаются в интерфейсе пользователя Системы.

4 ФУНКЦИОНАЛЬНОЕ ОПИСАНИЕ РАЗДЕЛОВ СИСТЕМЫ

4.1 Сбор текстовой информации

Раздел **Поиск** предназначен для поиска текстовых данных по сайтам и платформам социальных сетей в сети Интернет с помощью графического интерфейса.

4.1.1 Строка поисковых запросов

Предмет или тема поиска задаются с помощью поисковых параметров.

В поисковой системе «Крибрум. Публичный поиск» имеется две версии параметров поиска – сокращенная и расширенная версии параметров (рис. 11).

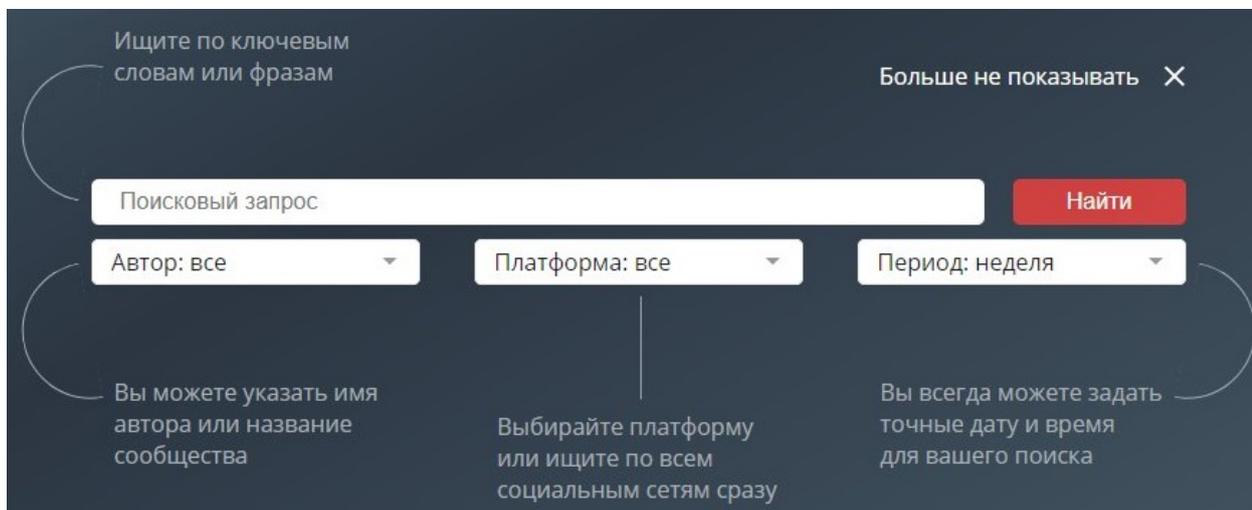


Рисунок 11 – Строка поисковых запросов

Введите поисковый запрос в виде набора ключевых слов или фраз в поисковую строку и нажмите кнопку **Найти**. Результаты поиска отображаются в виде ленты публикаций.

4.1.2 Расширенная настройка поисковых запросов

В меню расширенных настроек поиска реализованы следующие функции:

- выбор автора публикаций (если он заранее известен);
- выбор платформы социальных медиа (рис. 12);
- выбор временного периода поиска (рис. 13).

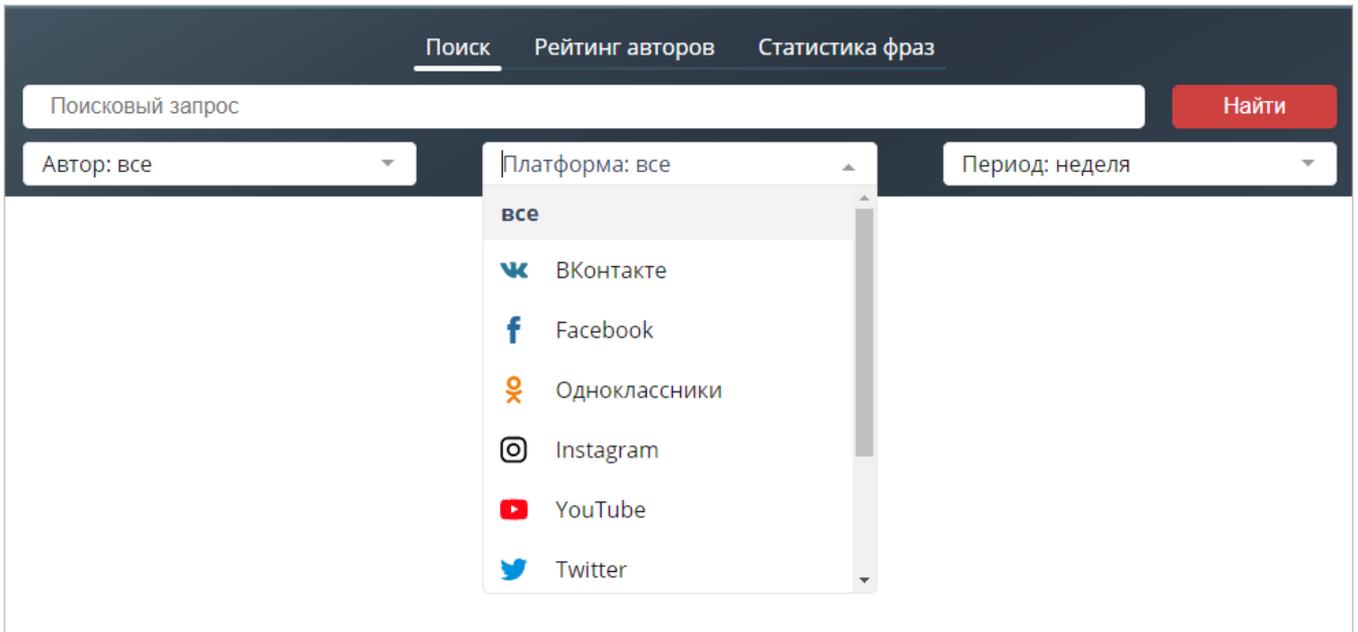


Рисунок 12 – Меню расширенных настроек поиска. Выбор платформы

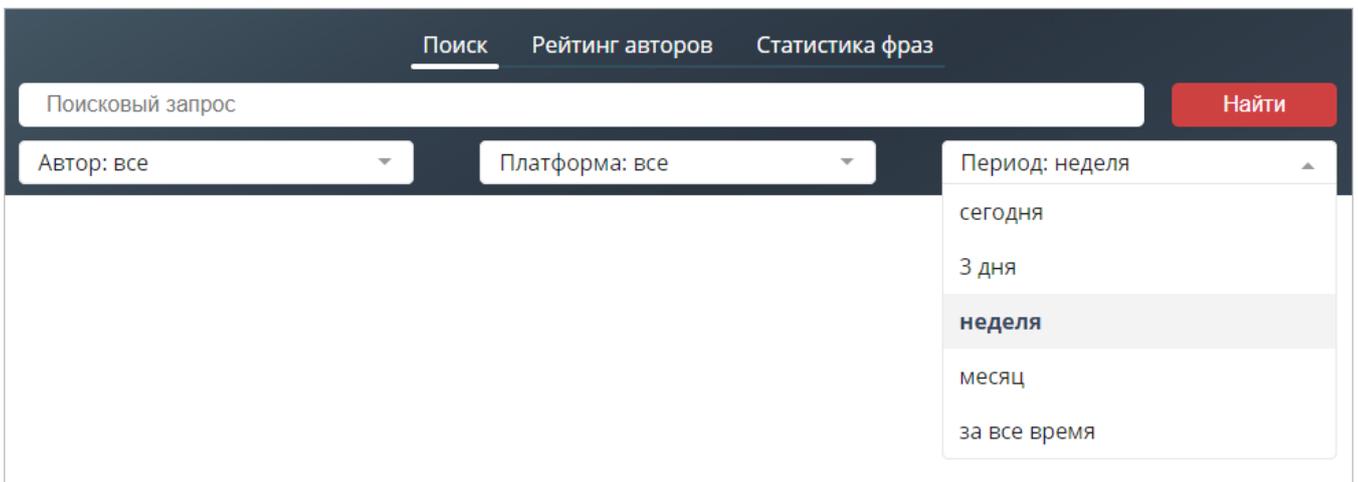


Рисунок 13 – Меню расширенных настроек поиска. Выбор периода поиска

4.1.3 Область отображения результатов поиска

Каждая найденная публикация содержит следующую информацию (рис. 14):

- название публикации;
- краткий текст публикации;
- изображение в публикации (если есть);
- счетчик количества изображений в публикации;
- счетчик количества похожих сообщений;
- счетчик количества репостов;
- имя автора публикации;
- время публикации;
- ссылка на оригинал публикации.

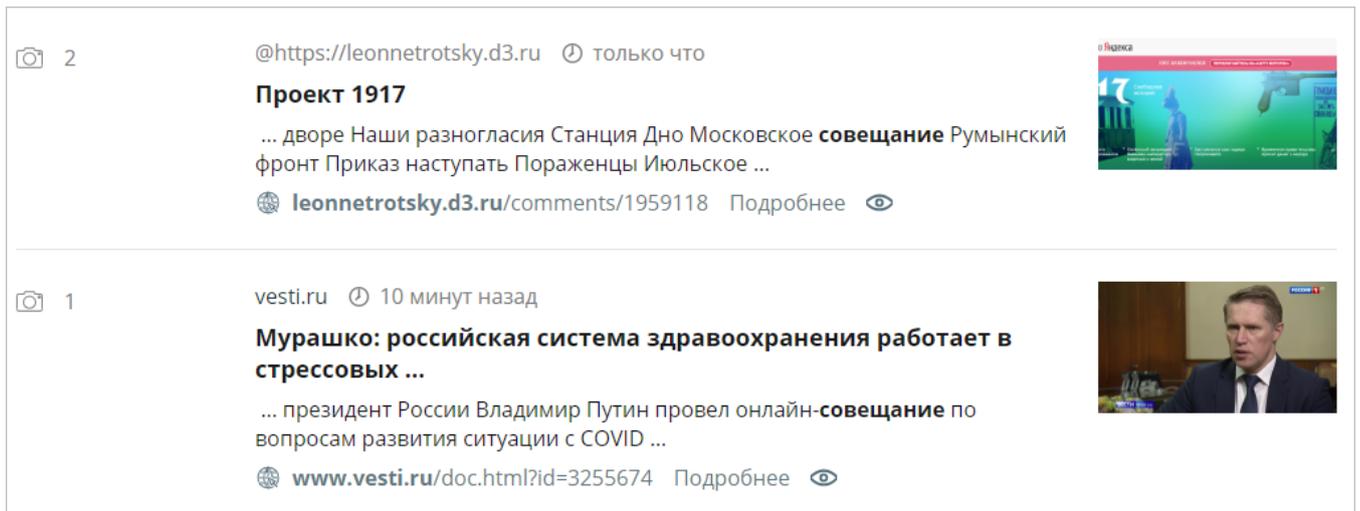


Рисунок 14 – Отображение результатов поиска

Поле публикации содержит кнопку **Подробнее** для перехода в область подробного рассмотрения найденной публикации с комментариями.

4.1.4 Область подробного рассмотрения найденной публикации с комментариями

В области подробного рассмотрения (рис. 15) найденной публикации с комментариями отображается следующая информация:

- имя автора публикации;
- время публикации;
- название публикации;
- полный текст публикации;
- изображение в публикации (если есть);
- ссылка на оригинал публикации;
- счетчик количества изображений в публикации;
- счетчик количества похожих сообщений;
- счетчик количества репостов;
- счетчик количества комментариев;
- перечень комментариев.

Поле публикации содержит кнопку **Назад к поиску** для возврата на страницу поиска.

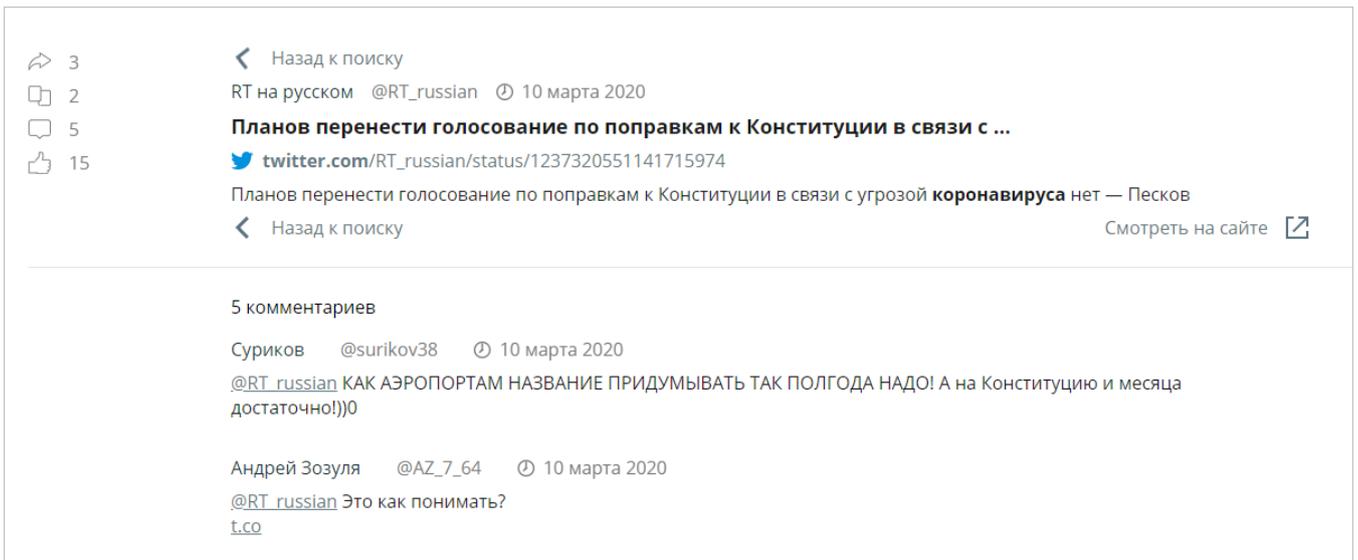


Рисунок 15 – Область подробного рассмотрения публикации

При наличии комментариев к публикации, они содержатся в виде перечня под найденной публикацией.

В поле комментария отображается следующая информация:

- имя автора комментария;
- время комментария;
- текст комментария;
- имена других комментаторов, с которыми идет переписка.

4.2 Обзор рейтинга авторов

Система «Крибрум. Публичный поиск» рассчитывает рейтинг, то есть авторитетность автора. Рейтинг определяется как величина регулярной аудитории данного пользователя, представляющая собой совокупность всех уникальных читателей (подписчиков, фолловеров, друзей) данного аккаунта; в расчетах также учитывается авторитетность читателей автора.

По умолчанию рейтинг авторов рассчитывается для всех следующих социальных сетей и отображается от автора с наибольшим рейтингом к автору с наименьшим (рис. 16):

- ВКонтакте;
- Facebook;
- Одноклассники;
- Instagram;
- YouTube;
- Twitter.

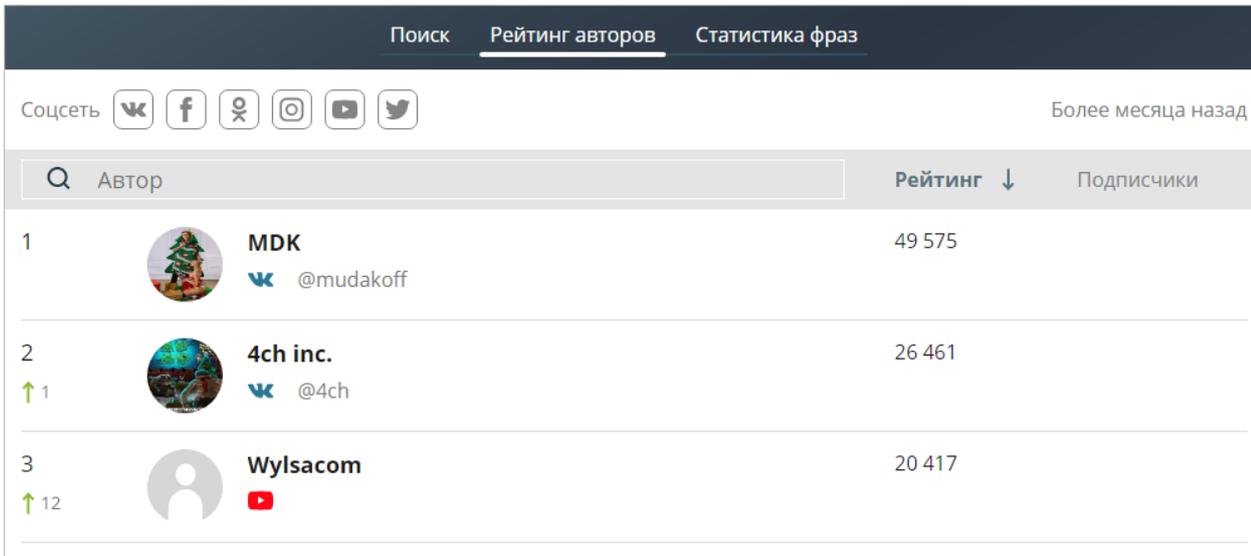


Рисунок 16 – Интерфейс раздела Рейтинг авторов

4.2.1 Настройки показа рейтинга

Система «Крибрум. Публичный поиск» позволяет также отфильтровывать авторов по социальным сетям или названию аккаунта (рис. 17).



Рисунок 17 – Фильтр в разделе Рейтинг авторов

Для просмотра рейтинга конкретного автора введите название аккаунта в поисковой строке раздела **Рейтинг авторов**.

Для обзора рейтинга авторов по какой-либо одной социальной сети нажмите на соответствующий значок соцсети.

При выборе соцсети Вконтакте или Одноклассники дополнительно можно отфильтровать аккаунты по людям или сообществам (рис. 18).



Рисунок 18 – Фильтр по людям или сообществам

4.2.2 Отображение результатов выдачи

По умолчанию список авторов отображается по убыванию рейтинга. Список также можно отфильтровать по количеству подписчиков, нажав на слово **Подписчики** рядом с поисковой строкой (рис. 19).

сбербанк		Рейтинг	Подписчики ↓
54 704 ↑	 Сбербанк  @sberbank	85	684 тыс. ↑ 1 758
9 849 ↑	 Сбербанк  @sberbank	383	435 тыс. ↑ 5 798
62 895 ↑	 Сбербанк для бизнеса  @sberbusiness	69	34 тыс. ↑

Рисунок 19 – Отображение рейтинга в поисковой выдаче

В поисковой выдаче указывается количество откликов на сообщения (стрелками показано увеличение или уменьшение количества), название аккаунта, рейтинг данного аккаунта и количество подписчиков (стрелками показывается увеличение или уменьшение подписчиков).

Нажмите на нужный аккаунт, чтобы показать все сообщения автора. Более подробная информация по сообщениям описана в разделе 4.1.3.

4.3 Обзор статистики фраз

Система «Крибрум. Публичный поиск» позволяет собирать статистику по количеству интересующих слов или фраз, упоминавшихся в социальных сетях.

4.3.1 Настройки показа статистики

В поисковой строке раздела **Статистика фраз** введите интересующее слово или фразу. По умолчанию фраза ищется во всех следующих соцсетях (рис. 20):

- ВКонтакте;
- Facebook;
- Одноклассники;
- Instagram;
- YouTube;
- Twitter.

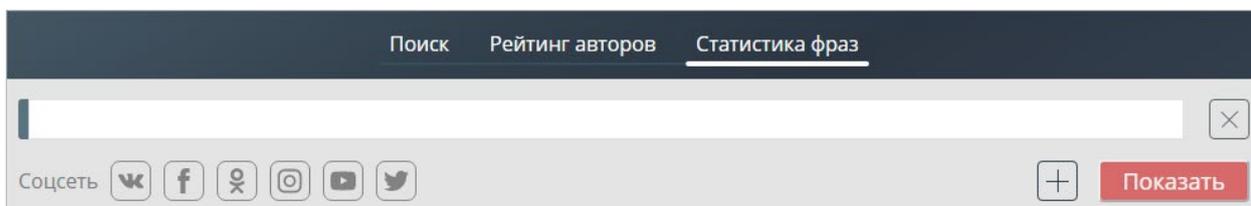


Рисунок 20 – Фильтр поиска в разделе Статистика фраз

Для выбора одной из соцсетей нажмите на соответствующий значок соцсети под поисковой строкой.

Для добавления еще одного слова или фразы нажмите на знак «+».

Для построения графика/ов нажмите кнопку **Показать**.

4.3.2 Отображение результатов выдачи

В зависимости от выбранных настроек может отображаться один или несколько графиков.

По вертикали указывается количество упоминаний слова или фразы в социальных сетях, по горизонтали – временной промежуток за последние несколько месяцев (рис. 21). Для более подробной визуализации необходимо передвинуть бегунок под графиком.

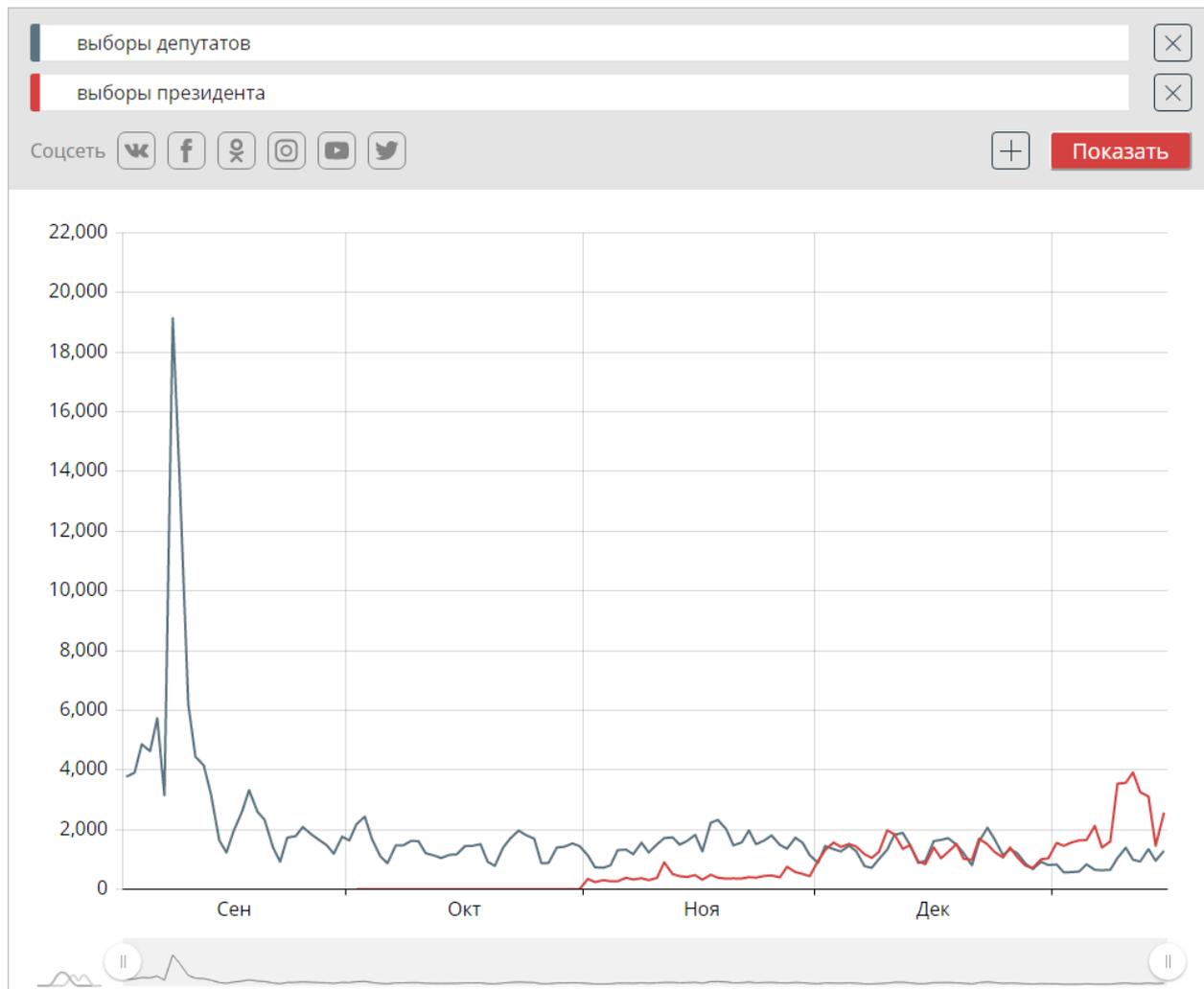


Рисунок 21 – График отображения количества упоминаний слова или фразы