Санкт-Петербургский государственный университет

**МАМАЕВ Иван Дмитриевич**

**Выпускная квалификационная работа**

**АВТОМАТИЧЕСКОЕ ВЫЯВЛЕНИЕ СКРЫТЫХ СООБЩЕСТВ В ТЕКСТАХ РУССКОЯЗЫЧНОГО КОРПУСА СОЦИАЛЬНЫХ СЕТЕЙ**

Уровень образования: магистратура

Направление 45.04.02 «Лингвистика»

Основная образовательная программа BM.5805.2019

«Компьютерная и прикладная лингвистика»

Профиль «Компьютерная лингвистика и интеллектуальные технологии»

Научный руководитель:

кандидат филол. наук,

доцент, кафедра математической лингвистики СПбГУ,

Митрофанова О.А.

Рецензент:

доктор техн. наук,

профессор, кафедра теоретической и прикладной лингвистики МГУ,

Лукашевич Н.В.

Санкт-Петербург

2021

**ОГЛАВЛЕНИЕ**

[ВВЕДЕНИЕ 4](#_Toc73346902)

[ГЛАВА 1. ИНТЕРНЕТ И СОЦИАЛЬНЫЕ СЕТИ 9](#_Toc73346903)

[1.1. Интернет-общение и его лингвистические особенности 9](#_Toc73346904)

[1.2. Определение социальных сетей и проблемы их анализа 15](#_Toc73346905)

[Выводы к первой главе 19](#_Toc73346906)

[ГЛАВА 2. СКРЫТЫЕ СООБЩЕСТВА 20](#_Toc73346907)

[2.1. Скрытые сообщества: определение и применение 20](#_Toc73346908)

[2.2. Определение скрытых сообществ в социальных сетях математическими методами 24](#_Toc73346909)

[2.3. Кластерный анализ как способ определения скрытых сообществ в социальных сетях 27](#_Toc73346910)

[Выводы ко второй главе 31](#_Toc73346911)

[ГЛАВА 3. МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ 32](#_Toc73346912)

[3.1. Определение и основные понятия тематического моделирования 32](#_Toc73346913)

[3.2. Автоматическое назначение меток тем 38](#_Toc73346914)

[Выводы к третьей главе 42](#_Toc73346915)

[ГЛАВА 4. ЭКСПЕРИМЕНТАЛЬНОЕ ВЫЯВЛЕНИЕ СКРЫТЫХ СООБЩЕСТВ В СОЦИАЛЬНЫХ СЕТЯХ 43](#_Toc73346916)

[4.1. Создание корпуса постов социальной сети ВКонтакте 43](#_Toc73346917)

[4.2. Построение автор-тематической модели 47](#_Toc73346918)

[4.3. Разработка алгоритма автоматического назначение меток тем 50](#_Toc73346919)

[4.4. Построение модели скрытых сообществ 55](#_Toc73346920)

[4.5. Сравнение различных методов выявления скрытых сообществ, оценка и анализ результатов 62](#_Toc73346921)

[Выводы к четвёртой главе 68](#_Toc73346922)

[ЗАКЛЮЧЕНИЕ 69](#_Toc73346923)

[СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ 71](#_Toc73346924)

[СПИСОК ЭЛЕКТРОННЫХ РЕСУРСОВ 80](#_Toc73346925)

[ПРИЛОЖЕНИЕ 1. ПРИМЕРЫ НЕКОТОРЫХ ВИЗУАЛИЗИРОВАННЫХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ 83](#_Toc73346926)

[ПРИЛОЖЕНИЕ 2. РЕЗУЛЬТАТЫ ЭКСПЕРТНЫХ ОЦЕНОК ДЛЯ КАНДИДАТОВ В МЕТКИ ТЕМ 87](#_Toc73346927)

**ВВЕДЕНИЕ**

В связи со стремительным ростом информационных технологий внимание исследователей социальных сетей приковано к текстам пользователей и выявлению связей между ними. Социальные сети – это зеркало общества, они позволяют оценить тенденции развития социума по формальным признакам. На основании общности этих тенденций выделяют пользовательские сегменты – скрытые сообщества. Для анализа используют графовые методы или кластерный анализ, однако эти методы чаще всего упускают из вида лингвистические признаки. Современные алгоритмы компьютерной лингвистики позволяют упростить задачу анализа дискурса, обработать и представить репрезентативные данные для больших текстовых коллекций, в том числе и для постов социальных сетей. Основываясь на тематической структуре текстов, можно выявить более устойчивые связи между пользователями.

**Актуальность работы** обусловлена необходимостью сочетания методов лингвистического и количественного анализа данных для построения семантической сети пользователей на основе тематической близости их постов.

**Новизна** исследования заключается в том, что в данной работе обсуждаются эксперименты по выявлению скрытых сообществ методами тематического моделирования и автоматического назначения меток тем, которые впервые выполнены на корпусе постов русскоязычных пользователей ВКонтакте[[1]](#footnote-1).

**Объект исследования** – скрытые сообщества, выявленные с помощью алгоритмов тематического моделирования среди пользователей русскоязычного сегмента социальной сети ВКонтакте.

**Предмет исследования** – количественные и качественные параметры скрытых сообществ, которые позволяют судить об актуальном состоянии групп и их тематике.

**Цель исследования** состоит в том, чтобы выявить скрытые связи между пользовательскими постами социальных сетей, основанные на общности лингвистических параметров текстов и их тематики.

Для достижения данной цели требуется решить следующие **задачи**:

1) исследовать специфику интернет-дискурса и выявить специфические лингвистические параметры текстов социальных сетей;

2) подготовить лингвистические данные для проведения экспериментов: произвести сборку и предобработку исследовательского корпуса;

3) разработать процедуру выявления скрытых связей между текстами социальных сетей;

4) произвести планирование и проведение экспериментов:

1. построение автор-тематической модели ATM для определения тематической специализации авторов;
2. автоматическое назначение меток тем из внешних и внутренних источников;
3. соотнесение данных о тематике текстов авторов с формулировками их интересов и с принадлежностью их к тем или иным сообществам;
4. выявление скрытых групп авторов с точки зрения общности их интересов;
5. сравнение трёх методов выделения скрытых сообществ: графового, кластерного и смешанного.

5) оценить результаты экспериментов.

**Материалом исследования** послужил корпус объёмом 25768 русскоязычных постов 714 пользователей социальной сети ВКонтакте объёмом 8679402 токенов.

**Гипотеза** исследования формулируется следующим образом: алгоритмы тематического моделирования позволяют выявить семантическое сходство анализируемых текстов, поэтому они могут применяться для построения моделей скрытых сообществ.

В данной работе применяются методы корпусной лингвистики, дистрибутивной семантики, вероятностного тематического моделирования и лингвистические приёмы анализа интернет-дискурса. Также были использованы следующие инструменты:

1. библиотека gensim[[2]](#footnote-2) для построения дистрибутивной и тематической моделей корпуса;
2. библиотека pymorphy2[[3]](#footnote-3) для морфологического анализа русскоязычных постов;
3. приложение Gephi[[4]](#footnote-4) для визуализации модели скрытых сообществ (на основе тематического моделирования и на основе графовых методов);
4. библиотеки scikit-learn[[5]](#footnote-5) и matplotlib[[6]](#footnote-6) для построения кластеров скрытых сообществ.

**Теоретическая значимость** исследования состоит в создании методики выявления скрытых сообществ на основе тематического сходства текстов корпуса.

**Практическая значимость** исследования заключается в том, что результаты исследования могут быть использованы для решения дальнейших задач не только в компьютерной лингвистике, но и в социологии: выявление онлайн-трендов, выявление потенциально опасных группировок и т.д.

**Объём и структура диссертации**. Работа состоит из четырёх глав, введения, заключения, списка использованной литературы, списка электронных ресурсов и двух приложений. В главах 1–3 отражены теоретические аспекты исследования. Глава 4 посвящена экспериментам с корпусом постов: сбор и обработка текстовых данных, построение тематической модели, автоматическое назначение меток тем, а также графическая реализация модели скрытых сообществ. В работе обсуждаются сложности, с которыми пришлось столкнуться при реализации экспериментов, а также делаются выводы по каждому этапу. В приложениях объёмом 10 стр. приведены примеры визуализированных тематических моделей и результаты экспертных оценок для автоматически извлечённых меток тем. Общий объём работы – 92 стр., основное содержание изложено на 82 стр. В тексте ВКР содержится 4 таблицы, 5 диаграмм и 16 рисунков. Список литературы включает 70 источников, из них – 30 на русском языке, 40 – на иностранных. Список электронных ресурсов состоит из 28 позиций.

Для ознакомления с корпусом пользовательских постов, скриптами программ по тематическому моделированию и автоматическому назначению меток тем, а также визуализированными результатами необходимо перейти в репозиторий GitHub[[7]](#footnote-7).

Основные положения данной ВКР были представлены в следующих публикациях.

1. Mamaev I. Automatic Detection of Hidden Communities in the Texts of Russian Social Network Corpus / I. Mamaev, O. Mitrofanova // Artificial Intelligence and Natural Language. AINL 2020. Communications in Computer and Information Science. — Vol. 1292. — Springer, Cham, 2020. — P. 17–33.
2. Mamaev I. Hidden Communities in the Russian Social Network Corpus: a Comparative Study of Detection Methods / I. Mamaev, O. Mitrofanova // Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020) co-located with 16th International Conference on Computational and Cognitive Linguistics (TEL 2020). — 2020. — P. 69–78.

**ГЛАВА 1. ИНТЕРНЕТ И СОЦИАЛЬНЫЕ СЕТИ**

**1.1. Интернет-общение и его лингвистические особенности**

В современном мире информационные технологии охватывают различные сферы повседневной жизни человека. Как следствие, в начале XXI века на смену традиционному живому общению людей пришло виртуальное общение в сети – интернет-общение. Люди очень много времени проводят в социальных сетях, переписываются друг с другом, оставляют комментарии к постам в интернет-группах, а также делают репосты – повторные публикации «какого-либо сообщения, то есть цитирование какого-либо поста методом пересылки с указанием первоисточника» [Калинина, 2014, с. 81]. В связи с освоением новых технологий современный дискурс начинает адаптироваться под интернет-коммуникацию.

Общение в интернете можно рассматривать с разных точек зрения: лингвистической, информационной, психологической и т.д. Такой подход позволяет сделать вывод, что само определение данного термина должно быть многогранным. Например, О.Н. Морозова определяет интернет-коммуникацию как «перцептивно-лингвистическую коммуникацию, сопряженную с актами познания и создания механизмов когерентности личности смыслообразующих систем», а также выделяет целую совокупность признаков, составляющих данное определение: коммуникативность и диалоговость, наблюдательность, межличностное взаимодействие, кросскультурный диалог и лингвистический компонент [Морозова, 2010, с. 152].

Из признаков, перечисленных выше, лингвистический компонент является наиболее важным для дальнейшего практического исследования. Качественные отличия интернет-общения от нормативных письменных постулатов можно проследить на многих языковых уровнях: графика, орфография, лексика и синтаксис.

**Графические особенности интернет-общения**

В письменной речи нельзя передать жесты и мимику, которые помогают выразить экспрессию в устной речи, именно поэтому графические средства являются неотъемлемой частью общения в интернете. Они помогают передать необходимую эмоцию.

Во-первых, нестандартными вариантами графики считаются повторение знаков препинания или их различная комбинация [Crystal, 2004, p. 89]. Так, например, в некоторых постах в социальных сетях используется многократное повторение вопросительных знаков для выражения полного непонимания: «*Кто в курсе что со светом на Английской 5????*»[[8]](#footnote-8).

Следующий момент, на который стоит обратить внимание, – это нестандартное использование прописных и строчных букв в пределах одного слова [Herring, 2012]: использование только прописных букв, чередование строчных и прописных букв и наоборот, а также иные комбинации: «*АУ БЕЗДЕЛЬНИКИ!!!!!!КАМЕРЫ В РУКИ И БЕГОМ СНИМАТЬ,КАК НАРУШАЮТСЯ ПРАВА ЖИВОТНЫХ!!!!КАК ИХ ИЗБИВАЮТ И ТАК ДАЛЕЕ!!!!!*»[[9]](#footnote-9)

В социальных сетях одним из способов выражения эмоций является использование эмотиконов (смайликов) – инновационные графические средства, которые расширяют возможности письма и компенсируют его недостающие средства выражения, присутствующие в устной речи [Нашхоева, 2011, с. 322]. «Исходя из словарей и библиотек смайлов, мы можем знать, что это за изображение, что оно обозначает, однако об истинном смысле использованного изображения знает только автор. Таким образом, только автор знает, какой смысл и значение он вкладывает в текст, сопровождая его тем или иным смайлом» [Смирнова, 2019, с. 78]: «*Многие из вас высказывали свое мнение о проблеме использования животных в цирках и шапито, многие выражали солидарность в вопросе отказа использования животных в развлекательных целях. 🤔*»[[10]](#footnote-10) В данном случае эмотикон передаёт вопросительную эмоцию.

**Орфографические особенности интернет-общения**

Орфография в интернет-коммуникации также подвергается некоторым изменениям. Явление, при котором сообщение представляет собой набор аббревиатур и сокращений, называется язык SMS [Бодулева, Зарипова, c. 77].

Один из самых частотных способов орфографического оформления текста – употребление сокращённой формы слова с целью сэкономить время при отправке сообщения или передать определённую эмоцию [Herring, 2012]: *«Когда появляются дешёвые туры на несезонные направления, постоянно кто-то да и напишет, мол, "хрен ли там делать щас". Спрашивали – отвечаем…»*[[11]](#footnote-11) В данном примере вместо нормативного слова *сейчас* используется редуцированная форма *щас*. В английском языке в качестве сокращённых форм используется числобуквенные комплексы: l8 – late, 2morrow – tomorrow, 2night – tonight и т.д.

Вместо сокращённой формы слова могут использоваться акронимы – сокращения, которые образованы из начальных букв или частей слов или словосочетаний: LOL – lots of fun, IMHO – in my humble opinion и т.д. [López-Rúa, 2007].

Стоит отметить, что сам интернет не является единственным результатом влияния на форму написания слова. Л. Сквайрс считает, что на орфографический вид слова влияет стиль текста и его тематика [Squires, 2010, p. 463].

**Лексические особенности интернет-общения**

В текстах социальных сетей происходит формирование характерного для интернет-коммуникации пласта лексики, который в некоторых случаях может отклоняться от нормативных употреблений. Одним из самых частых процессов можно назвать заимствование – «элемент чужого языка (слово, морфема, синтаксическая конструкция и т.п.), перенесённый из одного языка в другой в результате контактов языковых, а также сам процесс перехода элементов одного языка в другой…» [Ярцева, 2008, с. 1419]. В предложении *«Как вам мой сегодняшний лук?»* последнему слову, которое традиционно определяется либо как овощ, либо как орудие для стрельбы, можно приписать ещё одно значение – внешний вид, наряд. Данное слово восходит к английскому глаголу *look*, который в русском языке имеет значение *«смотреть, выглядеть»*. Слово претерпело адаптацию на фонетическом и морфологическом уровнях [Мамаев, Зайцева, 2019, c. 111], в результате чего оформилось как омоним к уже известным словам.

В некоторых случаях может произойти и дальнейшая адаптация иноязычного слова, что приводит к созданию фонетического аналога в русском языке. Так, слово *e-mail* для краткости и быстроты написания сообщения приобрело форму *«мыло»: «Лучше сначала напишите на мыло, что хотите, чтобы я мог проконсультироваться, если это необходимо, так как сам не все умею…»*[[12]](#footnote-12)

**Синтаксические особенности интернет-общения**

Синтаксический уровень интернет-коммуникации обладает своими особенностями, однако стоит учесть, что данный аспект неполно освещён в работах отечественных и зарубежных лингвистов. Американская исследовательница С. Херринг утверждает, что синтаксис наравне с морфологией интернет-языка, в отличие от других языковых уровней, описан недостаточно [Herring, 2012].

Также С. Херринг указывает, что одной из характерных особенностей синтаксиса является использование высказываний, указывающих на действие или состояние человека: *\*смеётся\*, \*молчит\** [Herring, 2012]. Эти высказывания в современном мире вытесняются использованием соответствующих эмотиконов как встроенных в социальную сеть, так и напечатанных с помощью графических символов (точки, запятые, скобки, тире и т.д.).

Среди синтаксических приёмов, которые могут привлечь внимание читателей интернет-постов и сообщений, учёные выделяют парцелляцию – «экспрессивный стилистический приём, который делит предложение на самостоятельные отрезки, графически выделенные двоеточием, точкой, многоточием и т.п.» [Иноземцева, 2011, с. 114]. В одной из популярных групп по поиску дешёвых путешествий Vandrouki данный приём очень часто используется для акцентирования внимания на каких-либо особенностях тура: *«Прочтите этот пост и спите спокойно! Специально для вас подборка дешевых вариантов вернуться из Европы в Москву всего от 2800 рублей за полет. Без пересадок и с багажом»[[13]](#footnote-13)*.

Синтаксический уровень, как и другие уровни языка интернет-общения, также подвержен упрощению. Холодковская Е.В. отмечает, что для языка общения в интернете, особенно в комментариях, характерны «языковая краткость, меньшая нормированность, использование эмоционального синтаксиса, упрощение синтаксических структур и влияние синтаксиса разговорной речи» [Холодковская, 2014, с. 82]. Эллипсис, инверсии, использование простых предложений и неполных предложений – это приёмы, благодаря которым пользователи социальных сетей достигают краткости своих постов, комментариев или сообщений: *«Очень дешево! 6 ночей на тропическом острове Хайнань из Петербурга всего за 14600 рублей с человека! В цене прямые перелеты с багажом, 4\* отель, завтраки, трансферы и страховка»[[14]](#footnote-14)*.

Помимо описанных лингвистических факторов важно отметить, что в интернет-коммуникации пользователей часто присутствует экстралингвистический компонент – мемы. Мем можно определить как «использующийся в коммуникации знак, имеющий устойчивую форму, которая содержит изменяющийся концепт» [Смородина, 2019, с. 78]. Мемы обыгрывают какую-то жизненную ситуацию, они чаще всего представлены в виде изображений с текстовым сопровождением. Среди характеристик мемов выделяют вирусность, эмоциональность, серийность, минимализм формы, актуальность и юмористическую составляющую [Смородина, 2019, с. 79].

Мемы, в отличие от стандартных лингвистических особенностей интернет-общения, рассчитаны на определённую категорию лиц. Так как основой для мемов служат актуальные события, то только та категория пользователей, которые осведомлены об определённой новости, способна считать его культурную коннотацию [Щурина, 2012, с. 163]. Так, например, мем, представленный на рисунке 1, смогут понять те категории лиц, которые:

* владеют английским языком;
* разбираются в терминологии общей фонетики;
* способны выявить игру слов.

При этом сам текст мема может подвергаться тем же лингвистическим модификациям, что и тексты постов или сообщений. На рисунке 1 применяется стяжённая форма *wanna* вместо полной формы *want to*. Также в конце второго предложения опущена точка.



**Рис. 1.** Пример мема

Таким образом, интернет-коммуникация – особый вид взаимодействия между людьми. Они общаются неформально, стараясь сократить время для того, чтобы напечатать сообщение, что приводит к изменениям на многих языковых уровнях.

**1.2. Определение социальных сетей и проблемы их анализа**

Интернет-пространство – эта социальная среда, в которой человек на сегодняшний день проводит большое количество времени: это область профессионального и бытового общения, отдыха, развлечения и т.д. В связи с этим актуальным представляется рассмотрение проблем, которые связаны с интернет-общением, в частности, в социальных сетях. Анализом социальных сетей занимаются как и специалисты в области прикладной математики, так и лингвисты, изучающие интернет-дискурс и выявляющие языковую специфику в виртуальной среде.

В настоящее время интернет-дискурс анализирует такие сферы интернет-пространства, как электронная почта, чаты, гостевые комнаты, а также непосредственно общение пользователей. В.И. Карасик противопоставляет два типа традиционного дискурса, основанных на устных и письменных текстах [Карасик, 2000, с. 5]. Интернет-дискурс имеет те же самые оппозиции: к устным текстам относится общение посредством видеоконференции в Skype или Zoom, а к письменным текстам – переписка в социальной сети.

Социальные сети – это «платформы, онлайн-сервисы или веб-сайты, предназначенные для построения, отражения и организации социальных взаимоотношений, визуализацией которых являются социальные графы. Характерной особенностью социальных сетей является то, что их контент наполняется самими участниками сети, и наличие явно установленных связей между участниками» [Вебер, Пименова, 2014, с. 634]. С течением времени, как результат, появился целый спектр возможностей проанализировать и изучить поведение людей в виртуальном пространстве.

Первой проблемой, с которой исследователи сталкиваются при изучении социальных сетей, – проблема поиска и извлечения информации. Интернет обладает следующими характеристиками: быстрота, доступность и оперативность. Они позволяют использовать интернет как средство познания, однако фильтрация получаемых знаний остаётся одним из важных элементов работы в социальных сетях, в которых часто формируются сообщества – объединённые группы людей в онлайновом режиме вокруг общих ценностей и интересов [Вебер, Пименова, 2014, с. 634].

В разделе 1.1 рассматривалась проблема влияния интернета и социальных сетей на письменную форму языка человека. Однако с недавнего времени преобразования стали происходить и в устной речи, что связано с появлением такой функции в социальных сетях, как голосовые сообщения. Н.В. Орлова приводит результаты исследования данного нововведения. Характерными преимуществами являются:

1) существенная экономия времени: несмотря на тот факт, что письменный язык в социальных сетях упрощается, пользователям проще будет записать этот же текст в аудиофайл, так как в среднем скорость произнесения текста больше, чем скорость его набора;

2) доступность к аудиофайлу в любое время: преимущество голосового сообщения над обычным телефонным звонком заключается в том, что пользователь в любое удобное для него время может прослушать файл и узнать какую-либо информацию;

3) больший спектр экстралингвистических факторов: на письме эмоции передаются с помощью средств графики и пунктуации (см. раздел 1.1.), однако многие люди предпочитают передавать эмоции голосом, а не письменными средствами.

В то же время существует и ряд недостатков, которые могут помешать адекватной коммуникации пользователей.

1. Помехи технического и внешнего уровня: шум ветра, стук открывающегося окна, посторонняя музыка, проблемы артикуляционного аппарата собеседника, случайно приложенный к динамикам палец собеседника – это те факторы, которые могут повлиять на восприятие информации и исказить её при использовании функции голосового сообщения. Данную проблему можно встретить не только в социальных сетях или мессенджерах, но и во всём интернете. Например, схожие проблемы присутствуют в системах машинного перевода, когда человек начитывает текст для перевода в микрофон. Неправильно распознанный звуковой сигнал будет преобразован в неверное слово, которое впоследствии при переводе будет распознана реципиентом как ошибка.

2. Проблема личного пространства: человек, прослушивающий голосовое сообщение не в наушниках, может подвергнуться критике со стороны окружающих, так как им не нравится данный способ передачи личной информации. Рамки понятия «личная информация» размываются [Орлова, 2018, с. 58–64].

Проблема распространения голосовых сообщений, таким образом, имеет как положительные, так и отрицательные стороны, которые необходимо учитывать при исследовании.

Ещё одной немаловажной стороной изучения социальных сетей является само взаимодействие между пользователями. А.А. Матусевич и Н.К. Рябцева в своих исследованиях социальных сетей и интернет-коммуникации среди русскоязычных пользователей утверждают, что для успешной коммуникации собеседников важны следующие критерии:

1) использование речевых единиц разных уровней;

2) владение особым стилем языка – сетевым сленгом;

3) использование экстралингвистических компонентов (например, визуализация сообщения или комментария) [Матусевич, 2016; Рябцева, 2016].

С точки зрения лингвистики перспективным направлением анализа также представляется изучение самих пользователей социальных сетей, анализ их постов и комментариев, составление корпуса интернет-текстов и выявление скрытых сообществ. Данный аспект будет рассмотрен в следующем разделе.

Итак, процесс общения в интернет-среде можно рассматривать как с точки зрения технических параметров, так с точки зрения различных языковых приёмов, которые помогают достичь успешной коммуникации.

**Выводы к первой главе**

Интернет-пространство на сегодняшний день является огромным полем для лингвистических исследований. Большое количество статей посвящено как общим вопросам функционирования языка в интернет-среде, так и частным, которые требуют тщательного рассмотрения. Инновации в интернет-технологиях и сервисах позволяют расширить круг средств, благодаря которым можно сделать текст более ярким и экспрессивным, что, несомненно, является материалом для последующего анализа. При этом следует учитывать следующие факторы:

1) язык в интернете функционирует по своим правилам, он претерпевает изменения, которые можно выявить на многих уровнях (лексический, синтаксический и т.д.);

2) при рассмотрении постов в социальных сетях и комментариев к ним важно обращать внимание на стиль сообщения, а также способ достижения успешной коммуникации;

3) сочетание лингвистических и экстралингвистических компонентов в постах социальных сетей и комментариях к ним поможет достичь наибольшего эффекта и побудить читателя к ответной реакции.

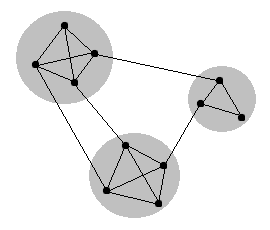
**ГЛАВА 2. СКРЫТЫЕ СООБЩЕСТВА**

**2.1. Скрытые сообщества: определение и применение**

В последнее время проблема определения сообществ превратилась в задачу сетевого анализа, что позволяет обеспечить понимание структуры социальных сетей. Несмотря на значительный прогресс в сфере сетевого анализа, существует такой вид сообществ, который называют скрытыми сообществами. Скрытые сообщества – те группы людей, которые имеют общие интересы, но связи между которыми либо размыты, либо отсутствуют. Реальный мир сети содержит неявную структуру сообществ: в жизни встречаются как секретные организации, так и временные группы. Эти структуры имеют неочевидные связи, они наименее сплочённые, чем обычные сообщества: семьи, коллеги или лучшие друзья [He, Li, Soundarajan, 2018, p. 92].

Аналогичная тенденция проявляется и в глобальной сети интернет. Так, например, если в социальных сетях существует несколько групп, посвящённых теме здоровья, то эти люди могут делать репосты на свою страницу, чтобы сохранить важную для них информацию. В то же самое время люди могут оставлять и свои собственные мысли, которые не привязаны ни к одному из сообществ. В результате личную страницу каждого пользователя можно разделить по тематическим блокам в зависимости от его записей. Можно взять несвязанных между собой пользователей и объединить их в скрытые сообщества по определённым тематикам.

В общем виде скрытые сообщества представляют совокупность нескольких семантических сетей (рис. 2). Каждая вершина в семантической сети имеет связи, однако в данном случае скрытые сообщества можно определить как связи между вершинами, которые находятся в разных семантических сетях.



**Рис. 2.** Графическое представление скрытых сообществ

Выявление скрытых сообществ применяется во многих сферах жизни человека. В работе [Митягин, Якушев, Бухановский, 2012] описывается процесс анализа текстов социальных сетей и определения лиц, склонных к наркомании. Авторы рассматривают две задачи: выявление пользователей, которые с некоторой вероятностью занимаются незаконным оборотом наркотиков, а также выявление тематик постов пользователей, которые склонны к употреблению наркотиков. Их алгоритм разделён на три основных этапа и обобщён в виде следующей схемы.

1. Определение страниц пользователей социальных сетей, посты которых связаны с искомой тематикой – потреблением наркотиков.

2. Переход по ссылкам на страницы данных пользователей.

3. С помощью алгоритмов извлечения мнений и взглядов оцениваются посты пользователей. Если искомые мнения не найдены, то поиск прекращается.

В данном случае использование данных социальных сетей в моделировании помогает определить взаимосвязь распространения наркокультуры как в социальных сетях, так и в обществе.



**Рис. 3.** Алгоритм поиска скрытых сообществ людей, склонных к наркомании [Митягин, Якушев, Бухановский, 2012]

В работе исследователей из ВШЭ П.А. Мейлахса и Ю.Г. Рыкова рассматриваются проблемы взаимодействия СПИД-отрицателей в социальных сетях: проводится анализ их постов, процента оставленных комментарием и лайков, процесса общения между участниками данных сообществ [Мейлахс, Рыков, 2015]. Было установлено, что диссиденты применяют различные стратегии для убеждения людей, которые не определись в вопросе о существовании вируса ВИЧ: описание личного опыта, идеологические аргументы, «научные» аргументы и т. д. Таким образом, люди, которые образовывали скрытые сообщества, читая лишь отдельные посты и статьи на эту темы, были подвержены влиянию пользователей данной группы и вступали в них, в результате чего статус скрытых сообществ менялся на статус явных.

Скрытые коллективы выявляются могут выявляться не только среди социально опасных групп, но и в научных сообществах. В работе [Хорошевский, Ефименко, 2017] приводятся результаты работы по выделению скрытых сообществ на основе анализа электронных копий трудов конференции «Диалог» за 2000–2009 гг. Были выявлены тематики, которые являются популярными среди докладчиков и на которые будет нацелено внимание исследователей в ближайшем будущем: искусственный интеллект, синтез речи и т.д.

В недавнем исследовании О.С. Смирновой было рассмотрено выявление личностей, склонных к суицидальному поведению, посредством автоматического анализа детских и подростковых аккаунтов в социальной сети ВКонтакте, размещённых в открытом доступе. Автор выделяет значения признаков суицидального поведения, классифицирует их и проводит экспертную оценку. Подход к автоматизированному выявлению групп риска аккаунтов социальной сети позволяет в дальнейшем оградить детей и подростков от суицидального контента [Смирнова, 2017].

Из вышесказанного следует, что методики выявления скрытых сообществ применяются для многих сфер сетевого анализа, однако возникает вопрос, какой из ныне существующих алгоритмов является наиболее оптимальным. Некоторые учёные проводят сравнительный анализ, описывая достоинства и недостатки различных подходов. Так, например, в статье [Fortunato, 2016] отмечается, что одни из часто встречающихся методов – это алгоритмы кластеризации и графовые алгоритмы. Среди них нельзя выделить универсальный. Для того чтобы разрешить этот вопрос, необходимо сделать акцент на проверке экспериментальных данных, а также собственных характеристиках метода, чтобы определить возможные сферы его применения. Для этого авторы предлагают краткое освещение методов, приложений с их реализацией и рекомендаций по их использованию.

Наравне с вышеописанными вариантами реализации поиска скрытых сообществ существует другой не менее эффективный способ – машинное обучение. В [Liu, Xue, Wu, 2020] авторы заявляют, что классические методы для работы с сообществами, такие как кластеризация, статистический вывод и графовый поиск, не эффективны, поскольку методы глубокого обучения демонстрируют возможность обрабатывать многомерные графические данные с впечатляющей производительностью. Применяются три основных метода из этой области – глубокие нейронные сети, графы с эмбеддингами и нейронные графы. По результатам серии экспериментов было установлено, что новый способ обнаружения скрытых сообществ действительно может работать с большими данными, но некоторые проблемы остались нерешёнными. Среди них можно выделить улучшение аннотирования узлов полученного графа, варьирование числа скрытых сообществ и пр.

Учитывая вышесказанное, мы можем сказать, что задача выявления скрытых сообществ позволяет решать различные социальные проблемы и предсказывать поведение групп как в реальной жизни, так и в интернет-среде.

**2.2. Определение скрытых сообществ в социальных сетях математическими методами**

В разделе 2.1 данной работы были описаны приёмы, которые применялись для выявления скрытых сообществ, однако в них в явном виде не указывались математические алгоритмы для анализа. Существует несколько математических подходов, благодаря которым можно выделить скрытые сообщества. В частности, в работе [Коломейченко, Чеповский, Чеповский, 2014] предлагается следующий алгоритм.

1. На первом этапе каждая вершина считается отдельным сообществом. Вычисляется и запоминается целевая функция L(M):

где

вероятность перехода между сообществами на каждом шаге случайного блуждания, *qi* – вероятность покинуть сообщество *i*,

вероятность остаться в сообществе *i*, *pα –* вероятность посетить вершину *α, Q –* количество сообществ, *Pi –* количество вершин в сообществе *i*. *H* – энтропия соответствующей величины.

2. Случайное блуждание формирует последовательность вершин.

3. С учётом частоты встречаемости вершин в полученной последовательности выбираются подмножества вершин, которые объединяются в сообщества.

4. Для заданного разбиения вычисляется метрика *L(M)*. Если её значение стало меньше, то разбиение *M* сохраняется и продолжается работа алгоритма (возврат к пункту 2). В другом случае, если значение целевой функции *L(M)* не уменьшилось, переходим к пункту 5.

5. Полученное разбиение M считать результатом выделения сообществ в сети [Коломейченко, Чеповский, Чеповский, 2014, с. 28].

В работе [Baumes, Goldberg, Magdon-Ismail, 2004] предлагается следующий метод: вводится функция

в которой *X(t) –* размер самого большого постоянного компонента, который можно выделить во время коммуникационных циклов (от *1* до *t*); *h –* размер скрытого сообщества, которое необходимо выявить; *ε –* уровень доверия, *t –*

время работы циклов до того момента, когда возникнет *X(t)*. При достижении уровня значений, при которых *X(t) ≥ H(t)*, можно утверждать, что выявлено скрытое сообщество при *1- ε* [Baumes, Goldberg, Magdon-Ismail, 2004, p. 6–7].

В статье [He, Li, Soundarajan, 2018, p. 92] в алгоритме *HICODE (Hidden Community Detection)* выделяется два этапа обнаружения скрытых сообществ. Первый этап – это нахождение количества слоёв сообществ, т.е. уровень того, насколько скрытной является исследуемая группа. Первый слой – это группа с наиболее сильными и явными связами, каждый следующий слой обладает меньшей степенью связи.

На втором этапе – этапе уточнения – происходит «улучшение» качества слоёв. Данный этап возникает из-за того, что более сильная структура сообщества может исказить сведения о структуре более слабых сообществ [He, Li, Soundarajan, 2018, p. 96].

Алгоритм *HICODE* также применялся и американским учёным Д. Зальцем и его коллегами для выявления скрытых групп в онлайн-форумах, в частности – в группах сайта Reddit. Алгоритм был улучшен методом, прогнозирующим веса рёбер и добавляющим прогнозируемые границы сообществ [Salz, Benavides, Li, 2019]. В результате было установлено, что структура сайта не упорядочена, границы между сообществами становятся более размытыми, что связано с расширением исследуемого сайта и более тесным взаимодействием между форумами.

В [Gmati, Mouakher, Gonzalez-Pardo, 2018] представлен метод выявления скрытых сообществ с применением двудольного графа, состоящего из двух наборов узлов и рёбер, соединяющих эти узлы. Новый подход был назван Fast-Bi Community Detection (FBCD), его основная идея заключается в том, чтобы исследовать множество максимального соответствия в двудольном графе, при этом уменьшив сложность алгоритма. Эксперименты проводились на разных социальных сетях, в том числе и Facebook. Среди выделенных сообществ оказались сообщества Southern Women, American Revolution и др.

Также стандартные математические модели, основанные на графах, подвергались модификации. В работе [Chobe, Zhan, 2019] авторы делают акцент на том, что современные методы выявления онлайн-сообществ не учитывают ключевую информацию, которая содержится в узле графа, поэтому он предлагает усовершенствовать алгоритм, используя поиск по ключевым словам – уникальному идентификатору узла. После построения массива ключевых слов применяется поиск по пересекающимся идентификаторам. Делается вывод, что данный метод позволяет расширить и персонализировать уже существующие модели сообществ в режиме онлайн.

Таким образом, огромное разнообразие математических методов выделения скрытых сообществ позволяет выбрать наиболее оптимальный подход для анализа пользовательской текстовой коллекции.

**2.3. Кластерный анализ как способ определения скрытых сообществ в социальных сетях**

Алгоритмы машинного обучения начали активно развиваться в последнее десятилетие, что позволило применить новые методы при выявлении скрытых сообществ. Одна из первых работ появилась в 2000-х гг. В [Mishra, Schreiber, Stanton, 2007] на основе показателей плотности и разреженности реальных сообществ проводится кластеризация скрытых. В статье показано, что сообщества могут приобретать приблизительно то же количество скрытых связей, что и реальные группы.

Социальные сети могут использоваться террористическими организациями для распространения сообщений с целью оказания влияния на людей и их дальнейшей вербовки в экстремистские группировки. Кластерный анализ, как и графовые методы, также применяется для выявления подобного рода сообществ. В [Krzemień, Sánchez-Rebollo, Puente, 2019] анализируются посты в сети Twitter, которые могут выявить потенциальных лидеров террористических организаций и их последователей. Исследователи прибегают к методам Big Data для анализа постов в режиме реального времени с целью классификации пользователей по различным параметрам: уровень активности, влияние на других пользователей, содержание сообщений и пр. Для классификации пользователей использовались методы нечёткой кластеризации. Однако авторы подчёркивают, что данный способ подойдёт не только для обнаружения террористов: подобным образом сотрудники страховых компаний могут выявлять нежелательных клиентов, а сотрудники миграционных служб могут извлечь дополнительную информацию о кандидатах, с которыми они общаются в данный момент.

В [Singh, Shakya, Biswas, 2016] был представлен алгоритм, основанный на кластеризации и семантическом сходстве анализируемых постов англоязычного сегмента Twitter. Семантическая близость выявлялась за счёт использования онлайн-тезауруса WordNet. В результате удалось выявить три устойчивых группы в исследуемом наборе данных.

Рейханиан А. также проводит аналогичное исследование, в котором комбинируется кластерный подход и подход, основанный на тематическом сходстве исследуемых текстов [Reihanian, 2016]. Материалом для исследования послужили открытые англоязычные датасеты, посвящённые отзывам на книги и фильмы: Movielens 100k, Book-Crossing, CIAO, MovieTweetings и Movielens Latest. Каждый объект подвергался предобработке, ему присваивалась тема. Подобная процедура проводилась с каждым последующим элементом, после чего к группе аннотированных данных применялся кластерный анализ. В результате сравнения созданного подхода с классическими алгоритмами было установлено, что качество сформированных моделей улучшилось в среднем на 1–2%.

В [Jia, Zhang, Zhang, 2019] выдвигается предположение, что типовые алгоритмы кластеризации наподобие k-средних или смешанные гауссовские модели не подходят для разрешения проблемы пересекающихся сообществ: в реальной жизни группы имеют более тесные связи, нежели те, которые обнаружены автоматическими методами. Авторы предлагают использовать алгоритм CommunityGAN (Community Detection with Generative Adversarial Nets) – выявление сообществ с применением генеративно-состязательных сетей. Генеративно-состязательная сеть является одним из алгоритмов машинного обучения без учителя, она состоит из двух нейронных сетей. Первая «порождает» образцы (сообщества), а вторая отличает истинные сообщества от ложных, тем самым позволяя уменьшить «зашумление» полученных моделей и представить более чёткую картину о пересекающихся сообществ. В результате анализа каждая вершина получает вес – коэффициент связи с другими сообществами. Эксперименты были проведены на пяти реальных датасетах, результаты позволили утверждать, что CommunityGAN существенно превосходит базовые алгоритмы.

В статье [Acharya, Zhang, 2020] кластеризация сети сообществ производится с помощью распределения Гумбеля на различных наборах данных: собрание политических произведений, роман «Отверженные» и пр. Как и во многих других статьях, сообщества-кластеры заранее не известны, поэтому применить какие-либо оценочные метрики представлялось затруднительным. Вместо этого авторы прибегли к метрике модулярности, которая оперирует только полученной сетью сообществ. Эксперименты с большим числом датасетов показали высокие результаты, что позволило сделать выводы о применимости данного алгоритма кластеризации к невзвешенным и неориентированным графам. Дальнейшие исследования предполагают работу со взвешенными и ориентированными графами.

Алгоритмы кластеризации также успешно применяются и на русскоязычном материале. В [Rytsarev, Kupriyanov, Kirsh, 2018] освещается проблема классификации аккаунтов Twitter на основе загруженных пользователями медиафайлов (фото, видео и пр.). Неоднородный характер контента, а также колоссальные объёмы анализируемой информации приводят к чрезмерной вычислительной сложности при обработке. Авторам удалось выделить 13 основных кластеров пользователей, объединённых такими тематиками, как животные, спорт, автомобили и пр. В [Хлопотов, Старцева, Макаренко, 2019] кластеризация проводится на основе мнений о фильмах. Перед исследователями стояла задача проанализировать комментарии, оставленные под фильмами на сайте Кинопоиск, и выявить их взаимосвязь с тематическими сообществами ВКонтакте. В результате анализа групп и фильмов были выявлены тенденции тематик каждого кластера, а сами пользователи с помощью алгоритма k-средних были разбиты на три группы – фанаты комиксов, юмористы, а также интеллектуалы.

Стоит отметить, что, несмотря на различные подходы к выявлению скрытых сообществ, экспериментальные данные показывают, что каждый из вышеописанных методов является достаточно эффективным способом, который впоследствии можно подвергнуть модификации. В нашей работе мы проведём анализ социальных сетей и выявление скрытых сообществ на основании не только математических методов, но и лингвистических.

**Выводы ко второй главе**

Одной из задач анализа данных является выявление скрытых сообществ и их последующее описание. Решение этого вопроса требует комплексного подхода, поскольку существуют следующие особенности проблемы:

1) в русско- и англоязычных статьях описано огромное количество алгоритмов выявления скрытых сообществ, которые, к сожалению, не создают общей картины, так как отсутствует единый набор данных для проверки и единая совокупность полученных результатов;

2) из пункта 1 также следует, что выбор пути реализации алгоритма из существующего многообразия всегда зависит от конкретной исследовательской задачи;

3) современные методы глубокого обучения, в отличие от графовых и кластерных, обрабатывают больший объём входных данных, с их помощью можно получить более качественные результаты, но при этом у них тоже есть свои недостатки.

**ГЛАВА 3. МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ**

**3.1. Определение и основные понятия тематического моделирования**

Тематическое моделирование – это способ построения модели коллекции текстов, который описывает переход от совокупности документов и слов в документах к набору тем, характеризующих содержание документов. Тематические модели – модели со скрытыми переменными, для выявления которых применяются алгоритмы нечеткой кластеризации, которая относит любое слово или документ с некоторой вероятностью к нескольким темам, т.е. алгоритм определяет семантическую близость.

Процедура тематического моделирования имеет широкое применение при работе как с корпусами художественных текстов, так и с корпусами, построенными на основе веба. В частности, данная процедура используется при анализе текстов социальных сетей (новости и посты), расширении тематик документов, информационном поиске и пр.

Первые работы, связанные с тематическим моделированием, появились в конце XX века. В [Li, Jain, 1998] был описан подход с применением векторной модели документа: каждому слову присваивался некоторый вес, который мог рассчитываться, например, с помощью TF-IDF или другими методами расчёта весов.

На сегодняшний день для практического исследования применяются многочисленные алгоритмы вероятностного тематического моделирования и их разновидности. Среди них можно выделить следующие алгоритмы.

1. Вероятностный латентно-семантический анализ (Probabilistic Latent Sematic Analysis, pLSA). В основе метода лежит факторный анализ: выявление скрытых связей изучаемых явлений и объектов. При классификации и кластеризации документов данный метод позволяет извлечь контекстно-зависимые значения лексических единиц [Глушков, 2018].
2. Скрытое размещение Дирихле (Latent Dirichlet Allocation, LDA). Модель, разработанная Девидом Блеем в 2003 году [Blei, Ng, Jordan, 2003], позволяет выделить неявные группы, в результате чего исследователи могут сделать выводы о том, почему некоторые элементы текстовых документов пересекаются.
3. Робастная тематическая модель SWB (special words with background) – это тематическая модель, в которой добавлены два компонента: фон и шум. Под фоном понимаются общеупотребительные слова, снижающие релевантность поиска тем, под шумом – специфичные для определённого документа термины либо редкие термины, отражённые в нехарактерных для корпуса темах [Воронцов, URL].
4. Динамическая тематическая модель – разновидность тематических моделей, которая позволяет отследить развитие тематик в определённый период времени [Воронцов, URL].

Вышеописанные методы имеют как свои достоинства, так и недостатки. Например, к положительным сторонам pLSA и LDA стоит отнести возможность подсчёта вероятности отношения документа к выявленным темам с их последующей группировкой. К недостаткам относят постоянное переобучение при загрузке новых текстовых данных в модель, выявление и удаление дубликатов, а также неавтоматизированную семантическую интерпретацию полученных тем [Воронцов, URL].

Для облегчения работы с интерпретацией полученных тем алгоритмы тематического моделирования не раз подвергались адаптации. Например, в [Ramage, Hall, Nallapati, 2009] авторы представили улучшенную версию скрытого размещения Дирихле под названием Labeled LDA, который множеству тематических слов приписывал вероятную тему. Темы приписывались на основе тегов, присвоенных документам в сети интернет. F-мера при работе с интернет-текстами достигла 52% по сравнению со стандартной моделью, чья F-мера равнялась 39%. Однако стоит отметить, что для работы с корпусами, которые основаны не на вебе, рассматриваемый алгоритм из-за отсутствия внутренних меток текстов может выдавать результаты низкого качества.

Алгоритмы тематического моделирования активно применяются и сегодня. В работе [Митрофанова, Шиморина, 2014] алгоритм LDA применялся для построения тематической модели корпуса Живого Журнала, определения тем пользователей и их интерпретации. Полученные результаты позволили сделать вывод о динамике языка в русскоязычной интернет-среде. В других работах [Митрофанова, 2015; Митрофанова, 2019] алгоритмы тематического моделирования уже применялись для выявления тем в художественных текстах: в сказках А.Н. Афанасьева и в романе М.А. Булгакова «Мастер и Маргарита».

Активная работа ведётся и в моделировании тематик социальных сетей: например, в лаборатории интернет-исследований ВШЭ. В работе О. Кольцовой [Koltsova, Nagornyy, 2019] на корпусах региональных российских новостей в социальных сетях и комментариев к ним исследовалось восприятие читателями социальных проблем. С помощью LDA формируются основные проблемы, поднимаемые в новостях, а также на основе комментариев пользователей оценивается их важность. Несмотря на различную степень настроений, выражаемую пользователями, авторы приходят к выводу, что большая степень поляризации мнений приводит к актуализации проблем.

Одной из модификаций LDA является алгоритм тематическое моделирование битермов (biterm topic modeling, BTM), который, в отличие от латентного размещения Дирихле, используется для построения модели коротких текстов (посты в Twitter или ответы на вопросы). В статье [Bodrunova, Blekanov, Kukarkin, 2019] оценивается работа BTM на примере постов в Twitter на русском и английском языках. Основная тематика анализируемых текстов – этнические конфликты. На основе интерпретируемости, надёжности и значимости тем была выявлена оценка процедуры моделирования. Результаты эксперимента показали, что выявление тематики поста происходит лучше, если дискуссия является многоязычной, но при этом интерпретируемость самой темы не зависит от много- или одноязычия и объёма данных.

Алгоритм BTM применяется и для выявления причинно-следственных связей между оффлайновой и сетевой активностью граждан. В работе [Smoliarova, Bodrunova, Yakunin, 2018] был исследован корпус постов Twitter, посвящённый острым конфликтным ситуациям. Для оценки взаимосвязи между онлайн-активностью и офлайн-активностью пользователей, авторы использовали тематическое моделирование, которое позволило установить дальнейшую корреляцию.

Другая обновлённая модель LDA – это автор-тематическая модель (author-topic modeling, ATM), которая сочетает в себе тематическую модель со связями между словами, документами и темами, и авторскую модель, в которой отражена связь авторов с документами. На рисунке 4 представлена схема ATM.



**Рис. 4.** Автор-тематическая модель

На схеме переменные и указывают на параметры скрытого размещения Дирихле, – это тематическое распределение документов, указывает на тему каждого слова в документе, – автор слова в документе, содержит информацию о распределении слов в каждой теме.

В последнее десятилетие учёные при тематическом анализу текстовых данных всё чаще прибегают к использованию ATM. В одной из недавних работ по анализ социальных сетей В.В. Сампетова создала и описала внутреннюю организацию корпуса русскоязычной социальной сети Пикабу и выявила схожие темы между пользователями [Mitrofanova, Sampetova, Mamaev, 2020]. В ходе построения автор-тематической модели авторам удалось исследовать соотношение тем авторских постов в корпусе и установить тематическое сходство постов пользователей данного сообщества.

В работе [Manika, Margam, 2019] был представлен метод для анализа текстовых данных в области библиотечных и информационных наук. Автор строит автор-тематическую модель для статей, опубликованных в журнале *DESIDOC Journal of Library and Information Technology* за 2007-2018 гг. Полученные результаты позволяют отслеживать современное состояние рассматриваемых областей, а также предсказывать актуальные проблемы библиотечного и информационного дела в Индии.

В статье [Panicheva, Mirzagitova, Ledovaya, 2017] авторы поставили перед собой цель оценить методы агрегирования семантических признаков при гендерной классификации текстов русскоязычных пользователей в социальных сетях, для этого был собран корпус постов Facebook[[15]](#footnote-15). Они использовали три модели: латентное размещение Дирихле, автор-тематическое моделирование и дистрибутивная семантическая кластеризация (DSC), при этом алгоритм ATM показал наилучшие результаты. Было установлено, что политические темы превалируют у пользователей мужского пола.

Для реализации процедур тематического моделирования были созданы различные приложения и библиотеки для языков программирования.

1. Stanford NLP Topic Modeling Tool[[16]](#footnote-16) – это модуль тематического моделирования, который включён в лингвистический процессор Стэнфордского университета. Этот инструмент разработан на языках программирования Java и Scala, включает в себя алгоритмы классического варианта LDA и LDA с метками тем.
2. MALLET[[17]](#footnote-17) представляет собой одну из ранних библиотек для тематического моделирования. Эта библиотека создана на Java, она включает в себя алгоритм LDA с сэмплированием Гиббса, доступна как в виде консольного приложения, так и в виде интерфейсного.
3. Gensim[[18]](#footnote-18) – это библиотека для Python, включающий LDA, а также модули для автор-тематического и динамического моделирования. Помимо тематического моделирования пакет gensim помогает реализовать процедуры, связанные с дистрибутивной семантикой.
4. Orange[[19]](#footnote-19) – это приложение для машинного обучения, использующая библиотеки Python для работы с текстовыми данными. Достоинствами данного приложения заключаются в построении «конвейера» целого ряда лингвистических процедур (токенизация, обработка стоп-словарём и пр.), пользователи также могут реализовать тематического моделирование корпуса с помощью классического LDA.

Таким образом, выделение тем в текстовых корпусах является одной из актуальных задач, при этом алгоритмы тематического моделирования позволяют упростить задачу.

**3.2. Автоматическое назначение меток тем**

Тематическое моделирование является одним из способов выявления скрытых тем в текстах, однако данный метод имеет существенный недостаток: с его помощью нельзя автоматически обобщить содержание темы документа, так как алгоритм в исходном виде не производит подбор меток тем и лишь упорядочивает слова внутри темы по вероятностям и присваивает каждой теме номер. Сам пользователь имеет возможность вручную обобщить темы, однако исследователи в области тематического моделирования не раз предпринимали попытки разработать процедуру автоматического назначения меток тем. Автоматический метод помогает сэкономить время на ручной ввод и облегчить интерпретацию данных.

Одна из первых работ по автоматическому назначению меток тем появилась в 2007 г. [Mei, Shen, Zhai, 2007]. Проблема определения тем формулируется как присваивание семантических помет данному множеству слов. Хорошая семантическая метка должна полностью охватывать тему.

Исследователи предложили два метода: извлечение фраз и получение n-грамм. Преимущество первого метода заключается в том, что полученные фразы уже содержат в себе грамматическую и осмысленную цельнооформленность, а преимуществом второго метода является независимость от обучаемой области, однако с точки зрения лингвистики они не всегда могут быть семантически целостными, а их грамматическая составляющая может быть нарушена (например, в результате лемматизации элементы n-грамм теряют исходные согласовательные связи).

Ранжирование кандидатов происходило с помощью релевантности нулевого порядка, которая предполагает выбор метки с наибольшим весом, определяемым по формуле:

В этой формуле – кандидат в метку темы, – множество слов, образующих тему, – i слово в теме.

Через два года была предпринята попытка улучшить качество получаемых меток. В статье [Magatti, Calegari, Ciucci, 2009] возможные кандидаты отражались в дереве тем (Topics Tree). Для работы с деревом был использован сервис *Google Directory*, каждый узел дерева – это метка-концепт. Последующее ранжирование производилось с помощью пяти различных метрик схожести: косинус, коэффициент Шимкевича-Симпсона («показатель родового сходства»), мера Жаккара, коэффициент Танимото и матрица взаимного сходства. Если большинство метрик выбирали одну метку как истинную тему, то она просто присваивается. При отрицательном результате подключались семантические анализаторы.

Одна из основополагающих работ появилась в 2011 г. [Lau, Grieser, Newman, 2011]. Информация о кандидатах в метки была взята из внешних источников – электронного справочного ресурса *Wikipedia* на английском языке и с сайтов, которые были выданы информационно-поисковой системой (ИПС) *Google.* Для проверки взаимосвязанности кандидатов и совокупности тематических слов и дальнейшего ранжирования использовались такие статистические метрики, как критерий Стьюдента, критерий 𝝌² Пирсона, функция правдоподобия и т.д.

В [Aletras, Stevenson, 2013] для решения задачи автоматического именования авторы пытаются извлечь метки из изображений, отвлекаясь от субъективных текстовых материалов. По мнению исследователей, данный метод имеет преимущество по сравнению со статистическими методами, так как изображения не зависят от языка. Из пяти тематических слов составляется запрос в ИПС *Google* и извлекаются первые 20 изображений. Из них извлекается текстовая информация и визуальная информация, в результате чего формируется визуальный мешок слов (bag-of-visual words, BOVW).

Процесс ранжирования меток также заключается в виде построения графа, узлы которого – изображения. Алгоритм PageRank ранжирует изображения с точки зрения близости к множеству тематических слов. Для взвешивания рёбер графа авторы прибегают к методу точечной взаимной информации, семантическому анализу, а также к визуальным особенностям изображения.

В последнее десятилетие учёные стали чаще прибегать к нейросетевым алгоритмам. Компания *Google* представила нейронный машинный перевод. Практически в это же время появились и первые попытки адаптировать нейронный сети под процесс автоматического наименования тем. В [Bhatia, Lau, Baldwin, 2016] авторы модифицируют метод, описанный в [Lau, Grieser, Newman, 2011], и создают векторные представления (word2vec, doc2vec), чтобы извлечь из статей *Wikipedia* возможных кандидатов. Для ранжирования и выбора метки с помощью заранее размеченных вручную текстов и кандидатов была обучена машина опорных векторов, которая извлекает наилучший вариант.

В [Mehdi, Seyedamin, Kochut, 2017] был предложен способ назначения меток в тем с использованием онтологий, а процесс ранжирования опирался на иерархию меток-концептов в самой онтологии.

Среди отечественных лингвистов также проводились эксперименты по автоматическому наименованию тем на русскоязычном материале. В [Ерофеева, Митрофанова, 2019] приводятся этапы работы по созданию выбора метки для множества слов и оценки результатов. Сама процедура включает извлечение кандидатов в метки тем с использованием морфологических фильтров и алгоритма PageRank и ранжированием полученных кандидатов. В [Kriukova, Erofeeva, Mitrofanova, 2018; Mirzagitova, Mitrofanova, 2016] были представлены два алгоритма, названных *Labels-ESA* и *Labels-Y*. *Labels-Y* назначает метки в результате выдачи запросов ИПС Yandex, а *Labels-ESA* отбирает метки тем из заголовков статей русскоязычного сегмента *«Википедии»* с помощью алгоритма ESA (Explicit semantic analysis).

В работе [Mitrofanova, Sampetova, Mamaev, 2020] предлагается извлекать метки из тем с помощью дистрибутивных моделей, а также с помощью внешних источников для извлечения родо-видовых отношений. Поскольку опыты проводились на русскоязычной сети Пикабу, то эталоном для сравнения меток являлись предварительно обработанные пользовательские хэштеги.

Извлечение родо-видовых связей было также описано в [Tikhomirov, Loukachevitch, Parkhomenko, 2020; Yadrintsev, Ryzhova, Sochenkov, 2020; Arefyev, Fedoseev, Kabanov, 2020]: для этого привлекались алгоритмы word2vec, нейронная сеть Bert, а также русскоязычный онлайн-ресурс Wiktionary[[20]](#footnote-20). Авторы отмечают, что применение этих методов позволяет улучшить интерпретируемость неразмеченного корпуса текстов

Опираясь на вышеизложенные теоретические описания, можно сделать вывод, что все существующие алгоритмы делятся на два типа: алгоритмы, работающие с внешними источниками (информационно-поисковые системы, справочные ресурсы и пр.), и алгоритмы, работающие с внутренними источниками (сам корпус). Комбинация полученных меток из разных источников позволит улучшить интерпретируемость темы.

**Выводы к третьей главе**

Тематическое моделирование – это процесс автоматического выделения тематических множеств в коллекции документов. Одним из самых популярных алгоритмов является Латентное размещение Дирихле и вероятностный латентно-семантический анализ: они показали хорошие результаты на различных корпусных данных. Автор-тематические модели как подвид тематических моделей позволяют учитывать авторство текстов. Этот алгоритм на сегодняшний день очень востребован в прикладной лингвистике.

Стоит отметить, что сам по себе процесс тематического моделирования скорее автоматизированный, чем автоматический, ведь такая проблема, как интерпретация полученных тематических множеств, в большинстве случаев решается с привлечением экспертов. Для автоматизации этой процедуры прорабатывается такое дополнение к тематическим моделям, как автоматическое назначение меток тем. На материале англоязычных текстов подобные алгоритмы уже нашли своё применение, однако для русскоязычных данных существует лишь небольшой набор экспериментальных алгоритмов. Вопрос об автоматическом назначении меток тем является актуальным, исследователи создают новые пути решения этой задачи.

**ГЛАВА 4. ЭКСПЕРИМЕНТАЛЬНОЕ ВЫЯВЛЕНИЕ СКРЫТЫХ СООБЩЕСТВ В СОЦИАЛЬНЫХ СЕТЯХ**

**4.1. Создание корпуса постов социальной сети ВКонтакте**

Для проведения экспериментов был создан корпус постов социальной сети ВКонтакте. Весь процесс можно разделить на шесть этапов: отбор пользователей вручную, написание парсера для сбора информации о текстовых постов, написание токенизатора для выделения в полученных текстах токенов, создание стоп-словаря и удаление слов, создающих «шумы», добавление би- и триграмм в корпус текстов, ручное постредактирование.

Отбор пользователей производился по таким параметрам, чтобы исключить их возможное пересечение: были выбраны как известные личности, так и простые люди, живущие в разных городах и увлекающиеся сразу несколькими темами.

Следующий этап заключался в написании парсера для извлечения текстов постов. В интернете доступны различные модули по сбору информации пользователей ВКонтакте (например, VKMiner[[21]](#footnote-21), созданный в ВШЭ), однако, на наш взгляд, они не позволяют скачать полную информацию о пользователях.

Для написания парсера использовался язык программирования Python 3.7[[22]](#footnote-22), библиотека beautifulsoup4[[23]](#footnote-23) для работы с веб-данными, vk-api[[24]](#footnote-24) для доступа к пользовательским страницам, а также библиотека запросов request[[25]](#footnote-25). Так как на одной странице стены пользователя размещено фиксированное количество постов, равное 20, было принято решение проанализировать первые десять страниц и выбрать текстовые посты, датированные не раньше 2018 года. К текстам самих постов также предъявлялись некоторые требования: материалом для последующего анализа могли послужить тексты, чья длина была больше 200 символов, так как в данном случае они являются репрезентативными и несут минимальную смысловую информацию[[26]](#footnote-26). Отдельно стоит отметить, что возникла необходимость создать конвертер эмотиконов в кодовое представление, так как среда IDLE не может распознать многие из этих символов, в результате чего выдавалось сообщение об ошибке.



**Рис. 5.** Пример кода для парсера ленты пользователя ВКонтакте

Наиболее сложным с технической точки зрения являлся этап автоматической обработки текстов постов, поскольку для верной выдачи слов в тематических моделях нужно учесть особенности интернет-дискурса. Для обработки корпуса постов необходимо реализовать такие процессы, как токенизация на основе регулярных выражений, лемматизация (нормализация), обработка полученных токенов с помощью стоп-словаря, добавление би- и триграмм в корпус. Тексты пользователей в некоторых случаях представляют собой сочетание символов разных алфавитов. Например, большинство постов, посвящённых описанию фильмов, включали в себя символы греческого алфавита: русская буква *М* с графической точки зрения очень похожа на греческую букву *Μ* (мю). Такая замена обусловлена тем, что пользователи социальной сети во избежание удаления поста в связи с нарушением авторских прав прибегают к подмене символов, чтобы обмануть систему. В результате токенизации слово *«Μашина»* распознавалась как два различных токена: *«Μ»* и *«ашина»*. Такая ситуация наблюдалась у большинства слов, представленных числобуквенным комплексом или комплексом из букв различных алфавитов, поэтому для корректной обработки данные слова редактировались вручную перед обработкой, либо вносились в стоп-словарь, если это были служебные части речи.

Среди наиболее частых проблем при токенизации текстов постов можно выделить следующие:

* сочетание букв разных алфавитов в одном слове;
* написание слов на арабском или японском языке, такие слова добавлялись в исключения и не были использованы для последующего анализа;
* использование диакритических символов, которые также были удалены из текстов постов;
* использование других нерепрезентативных символов Unicode (галочки, солнышки, сердечки и другие).

Для добавления токена в корпус его необходимо подвергнуть лемматизации. Для этого была использована библиотека pymorphy2[[27]](#footnote-27), в котором есть функция нормализации слова *morph.parse(token)[0].normal\_form*. Сам модуль при нормализации слова обращается к словарю OpenCorpora[[28]](#footnote-28) и приводит нормальную форму слова. В случаях, если слово не было найдено, модуль по аналогии со множеством других похожих лексем анализирует его и приводит возможную нормальную форму. Во время обработки модуль предсказывал леммы для слов, связанных с названиями социальных сетей, неологизмами, окказионализмами, неизвестными фамилиями и именами собственными, а также морфологически и графически адаптированными иностранными словами. В большинстве случаев приводилась верная начальная форма, однако некоторые варианты оказывались ошибочными и требовали дальнейшего постредактирования. Среди них можно выделить следующие случаи:

* Социальная сеть *Instagram (Инстаграм)* получило лемму «инстагра», т.е. модуль изначально распознал заимствованное слово как форму женского рода дательного падежа множественного числа (ср. *мамам – мама, играм – игра* и т.д.).
* Слово *«тревел»* было распознано как форма прошедшего времени от глагола *«треветь»*.
* Имя существительное *«море»* в именительном падеже было распознано программой как форма предложного падежа единственного числа от слова *«мор»*.
* Имя существительное *«фейк»* приобрело форму *«фейка»* в значении *«маленькая фея»*.

Во время лемматизации каждая лексема проверялась на наличие в стоп-словаре, созданного на основе Национального корпуса русского языка[[29]](#footnote-29) и «Нового частотного словаря русской лексики» О.Н. Ляшевской и С.А. Шарова[[30]](#footnote-30) куда вошли слова, не несущие смысловой нагрузки для текста и создающие «шумы»: предлоги, союзы, частицы, междометия, символы различных алфавитов, обсценная лексика, сокращения и т.д. Итоговое количество стоп-слов равняется 1423.

Итоговый этап создания корпуса – добавление би- и триграмм. В [Нокель, Лукашевич, 2015] указано, что с точки зрения семантической организации некоторые семантические единицы (устойчивые словосочетания, идиомы и т.д.) не могут быть разбиты на униграммы. Было принято решение добавить в корпус биграммы и триграммы, которые встречаются в тексте более пяти раз. Создание биграмм и триграмм было организовано с помощью пакета gensim для тематического моделирования[[31]](#footnote-31), в частности – библиотек Phrases и Phraser. Символ *«\_»* соединил униграммы и превратил их в лексическую конструкцию: *«гераниевый\_масло»*, *«белый\_ворона»* и т.д.



**Рис. 6.** Код для добавления биграмм и триграмм в корпус

Итоговая версия корпуса представляет из себя список файлов расширения .txt в формате *USER\_ID\_\**, где *\** – номер страницы пользователя. Такое название оправдано удобством для дальнейшего вызова файлов и построения автор-тематической модели для пользователей. Каждый файл содержит список списков, каждый подсписок содержит в себе леммы, полученные в результате автоматической обработки текста. Общий объём корпуса составляет до обработки составляет 8679402 токенов, после – 2352979 токенов (714 пользователей, 25768 постов).

**4.2. Построение автор-тематической модели**

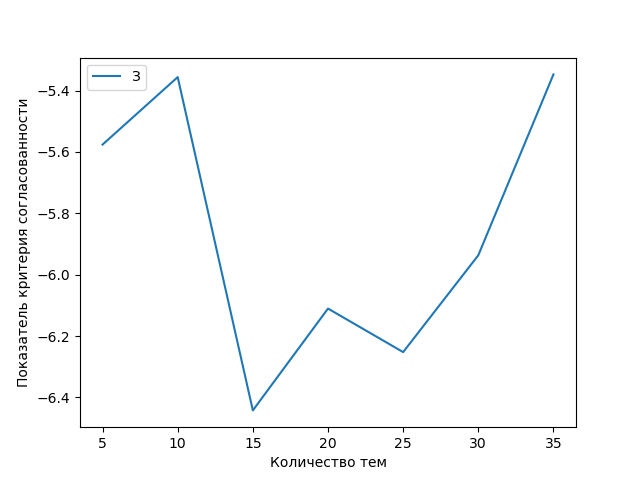
После построения корпуса текстов социальных сетей необходимо создать и обучить модель ATM – автор-тематическую модель. Для реализации необходимо обратиться к библиотеке gensim и адаптировать программный код, созданный для выделения тем в научных текстах. Весь процесс построения ATM состоит из трёх этапов.

1. Поиск наиболее оптимального количества тем в каждом отдельно взятом пользовательском подкорпусе с использованием меры U-Mass. Данная процедура заключается в том, чтобы построить график зависимости количества тем от репрезентативности самих тем. В теории формула имеет вид:

*,* (6)

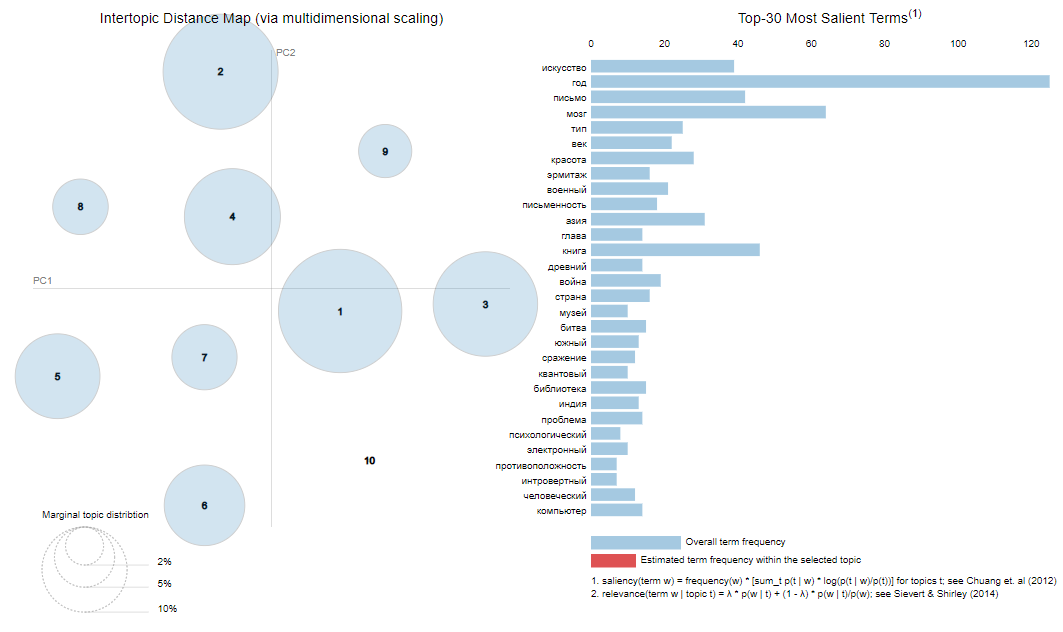
где – количество документов, в которых находятся слова и , а – величина, указывающая, сколько документов содержит слово [Stevens, Kegelmeyer, Andrzejewski, 2012]. Наиболее высокое значение U-Mass показывает, что при данном количестве тем модель будет работать наиболее эффективно. Стоит отметить, что на практике расчёт оптимальных тем данным способом имеет недостаток – создание тем-дублетов, в результате чего иногда приходилось изменять количество тем в меньшую сторону.

Была проведена серия экспериментов для выявления оптимального количества тем со следующими параметрами: минимальное количество тем -5, максимальное – 35, шаг – 5. В результате были получены данные для дальнейшего построения автор-тематической модели.



**Рис. 7.** Оптимальное количество тем для пользователя с ID 385450

2. Построение автор-тематической модели с выводом определённого количества тем для каждого пользователя и последующая визуализация тематической модели с помощью библиотеки pyLDAvis[[32]](#footnote-32) (см. рис. 8). В приложении 1 можно ознакомиться с другими визуализированными моделями.



**Рис. 8.** Визуализация тематической модели для пользователя с ID 385450

В качестве метода многомерного шкалирования (mds, multidimensional scaling) был выбран tsne (t-distributed Stochastic Neighbor Embedding), суть которого заключается в том, что близкие по смыслу темы не накладываются друг на друга, а находятся в непосредственной близости, а разные темы находятся на определённом удалении. Параметр mmds (metric multidimensional scaling, метрическое многомерное шкалирование) также можно применить, однако близкие темы будут накладываться друг на друга в центре схемы, что создаёт определённые неудобства для процентного соотношения тем в подкорпусе.

3. Сохранение результатов в txt-файле, который будет впоследствии использоваться для процедуры автоматического выявления кандидатов в метки тем.

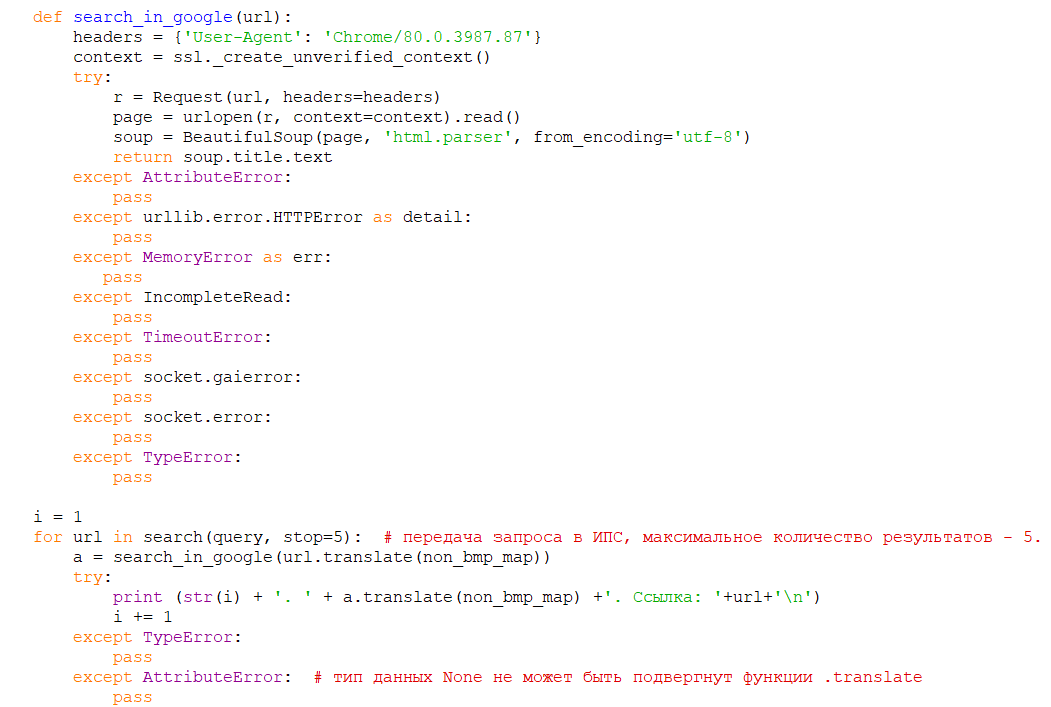
**4.3. Разработка алгоритма автоматического назначение меток тем**

Перед построением модели скрытых сообществ в социальной сети ВКонтакте необходимо определить тематики постов пользователей. В англоязычной литературе используется термин *automatic topic labelling* – автоматическое назначение меток тем.

В [Ерофеева, Митрофанова, 2019] описывается алгоритм поиска кандидатов в метки тем с помощью информационно-поисковой системы Яндекс и статей интернет-ресурса Википедия. Такой комбинированный подход по результатам работы оказался успешным, однако эксперимент проводился на основе корпуса лингвистических текстов. В случае анализа корпуса социальных сетей обращение к Википедии кажется нецелесообразным, так как данный ресурс в большей степени ориентирован на научные тексты, а не на тексты общей тематики. Сложность анализа также заключается в том, что посты одного и того же пользователя могут относиться к совершенно разным и даже несвязанным между собой темам: например, человек может интересоваться как готовкой, так и антропологическими исследованиями.

В нашем исследовании процесс автоматического назначения меток тем происходил в несколько этапов.

1. Написание парсера на Python для автоматического извлечения ключевых слов из тем, создания поискового запроса и его ввода в информационно-поисковую систему Google (внешний источник). Выбор данного сервиса обусловлен тем, что результаты запроса уже ранжированы с помощью PageRank. Для обхода блокировок в результате многократных запросов был использован модуль *User-Agent*, который позволяет сочетать различные версии операционной системы, спецификаций и версий браузера.



**Рис. 9.** Часть кода для парсинга результатов запросов в Google

2. Написание парсера на Python для создания автоматического извлечения ключевых слов из тем, создания запроса и обращения к RuWordNet[[33]](#footnote-33) для поиска гиперонимов. К сожалению, не все слова в исследуемом тезаурусе обладают гиперонимическими отношениями: имена собственные, неологизмы, заимствованные слова, поэтому результат в данном случае поиска был нулевым. Сочетание гиперонимов помогает создать целостное представление о теме.

3. Вычленение сочетаний из заголовков сайтов, полученных в результате парсинга запросов в ИПС Google с помощью pymorphy2[[34]](#footnote-34) на основе следующих шаблонов: ADJF+NOUN, NOUN+NOUN, NOUN+NOUN+NOUN, PREP+NOUN+NOUN, NOUN+PREP+NOUN, NOUN+COMP, ADVB+ADJF+NOUN, NOUN+CONJ+NOUN, ADJF+ADJF+NOUN, VERB+ADVB, VERB+INFN, PRTF+NOUN, PRTS+NOUN, INFN+ADVB, VERB+PREP+NOUN, INFN+ADVB, PREP+ADJF+NOUN. Ранжирование полученных групп производилось по частоте встречаемости в Национальном корпусе русского языка.

4. Для извлечения меток из корпуса необходимо построить дистрибутивную модель корпуса. Под меткой в данном случае мы понимаем средний вектор тематических слов, извлечённый из дистрибутивной модели корпуса.

5. Создание опросника на сервисе Google Формы[[35]](#footnote-35), результаты которого помогут определить корректность автоматически сформированных кандидатов в метки тем. На данный момент не существует «золотого стандарта», который служил бы опорой для выявления меток тем. Десять независимых экспертов оценили кандидаты по шкале от 0 (не соответствует) до 1 (соответствует).

6. Сравнение ответов, ручное постредактирование и финальное назначение меток тем, которые позволили создать в дальнейшем сеть скрытых сообществ.

**Таблица 1.** Сравнение полученных меток

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Тема** | **Метки из внешних источников (Google и RuWordNet)** | **Метки из внутренних источников (корпус)** |
| *834* | *адрес, внимание,* ***страница,*** *надёжно, ссылка, взламывать, скоро, перевод,* ***фотография****, страничка* | *Россия Москва, информационная безопасность,* ***рядовой пользователь****, победитель,* ***изображение****,* ***фрагмент*** | ***ссылка****,* ***видео****, просить,* ***фотография****, появиться* |
| *57296* | *конкурс,* ***университет****, государственный, молодёжный, проект научный, российский, год, имя, студенческий* | *новость науки,* ***Российская ассоциация,***  ***учебных заведений****, соревнование, опыт, молодежный* | *октябрь, июнь, белгород, миллиард,* ***институт*** |
| *94855* | ***мышца****,* ***рост****, минимальный, адаптационный, нагрузка, максимум, выяснить, спешить, гипертрофия, учёный* | ***адаптация мышц****, путь тренера, магазин спорт.питания, орган, биология, груз* | ***рост****,* ***движение****, высокий, отдельный, выступать* |
| *5019159* | ***съёмка****, ребёнок, заранее, родитель, фото, воспитатель, одежда,* ***снимка****,* ***коллаж****, группа* | *в детском саду, личный блог, детская* ***фотосъемка****,* ***изображение****, снять, состав* | *прекрасный,* ***образ****, радость, здоровье, сад* |

**Тема пользователя 41003889:** *овсянка, банка, ложка, йогурт, молоко, крышка, стакан, закрывать, добавлять, мёд.*

**Таблица 2.** Результаты экспертной оценки кандидатов в метки тем для пользователя 41003889

|  |  |  |  |
| --- | --- | --- | --- |
| **Bi- and trigrams** | ***рецепты приготовления*** | ***запаренная овсянка*** | ***великолепные завтраки*** |
| **Assessing results** | 10 | 6 | 9 |
| **Unigrams** | ***крупа*** | ***кулинария*** | ***искусство*** |
| **Assessing results** | 3 | 6 | 0 |

**Тема пользователя 820468:** *человек, джокер, оказаться, обстоятельство, убить, язык, несколько, изучение, выбрать, жертва.*

**Таблица 3.** Результаты экспертной оценки кандидатов в метки тем для пользователя 820468

|  |  |  |  |
| --- | --- | --- | --- |
| **Bi- and trigrams** | ***лучший фильм*** | ***фильм года*** | ***российские фильмы*** |
| **Assessing results** | 9 | 7 | 0 |
| **Unigrams** | ***существо*** | ***событие*** | ***отклонение*** |
| **Assessing results** | 2 | 4 | 4 |

С более подробными результатами экспертных оценок можно ознакомиться в приложении 2.

Самый простой способ подобрать метку для темы – взять первое тематическое слово из полученного множества, однако по результатам, приведённым в таблице 1, первое слово не всегда описывает тему.

Несмотря на то, что униграммные кандидаты оценивались ниже, чем би- и триграммы, стоит отметить, что униграммы подходят для более общих тем, в то время как би- и триграммы следует использовать для узконаправленных тематик. Сочетание экспертных и автоматических процедур помогает лучше интерпретировать полученные тематические модели.

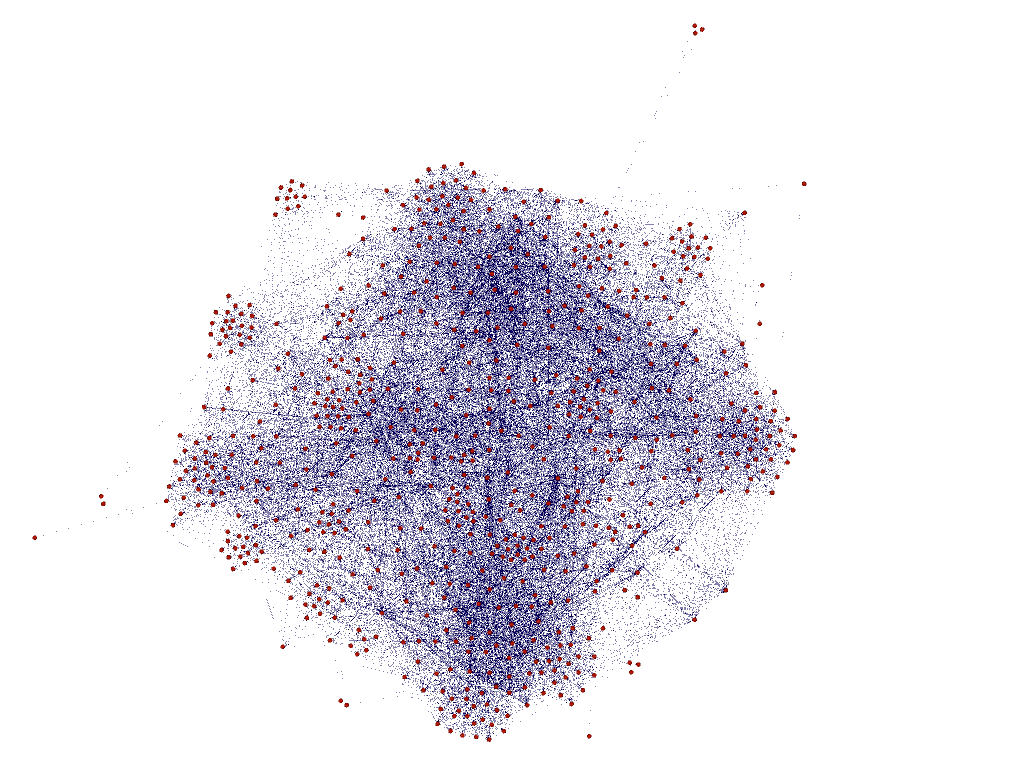
**4.4. Построение модели скрытых сообществ**

По результатам работы алгоритма автоматического выявления кандидатов меток тем была составлена сводная таблица пользователей и основных тематик их постов, которую позже была преобразована в таблицу связи тем с пользователями. Так как алгоритм мог выявить различные способы наименования темы, то необходимо было найти для них единое понятие. При работе с темами были выявлены следующие способы их объединения:

1. при выявлении согласованных и несогласованных определений доминирующим выбиралось сочетание с согласованным определением: *«организации студентов», «студенческие организации» — «студенческие организации»*;
2. антонимичные темы объединялись в одну, так как в их основе лежит единый семантический компонент: *«болезни», «здоровье» — «здоровая жизнь»*;
3. синонимы находятся в одном семантическом поле, поэтому они были объединены: *«фильмы и сериалы», «кино», «телевидение», «кинопроизводство» — «кино»; «религия», «православные советы» — «религия»; «фотосъёмка», «профессиональные фотосессии» — «фотографирование»*.

В результате для дальнейшей работы с семантическими сетями была построена таблица взаимосвязи тем, которые можно охарактеризовать как рёбра сети, пользователей, которые являются вершинами сети.

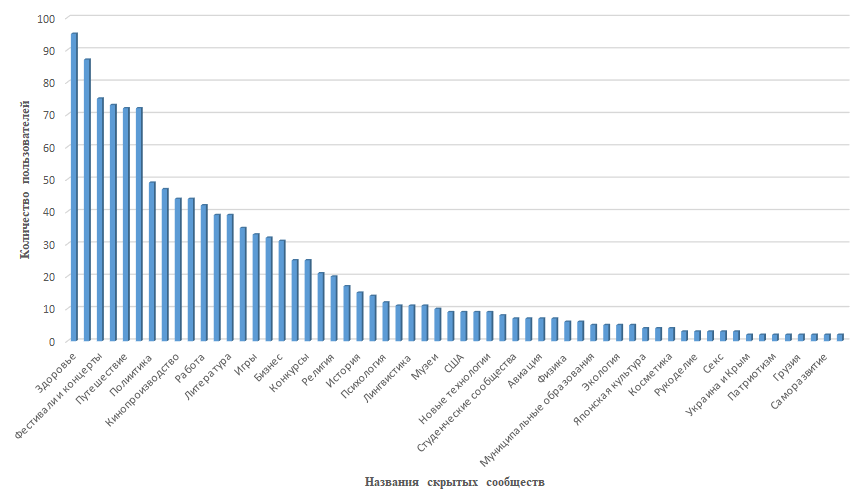
Для построения модели скрытых сообществ было использовано приложение Gephi[[36]](#footnote-36).



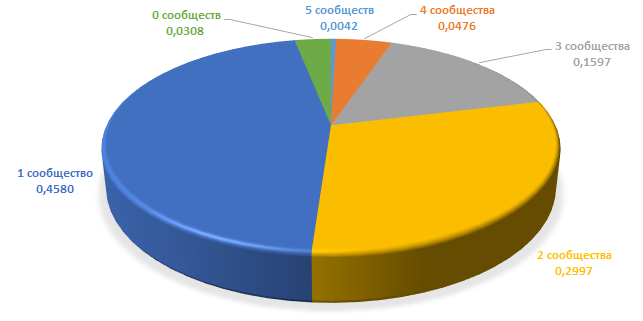
**Рис. 10.** Модель скрытых сообществ на основе общих тем постов

Скрытые сообщества в данной модели представлены в виде группы вершин, связанных рёбрами одних цветов. Среди 714 пользователей в итоге было выявлено 60 скрытых сообществ (диаграмма 1), при этом к ним относятся области широкого профиля (путешествие, здоровье и т.д.), так и узконаправленные темы (Украина и Крым, Грузия, патриотизм и т.д.).

На диаграмме 2 представлено соотношение пользователей социальной сети ВКонтакте по количеству скрытых сообществ, в которых они состоят. Почти половина (около 46%) проанализированных пользователей состоит в одном скрытом сообществе, меньше одного процента пользователей не состоят в скрытых сообществах, оставшаяся часть состоит в более чем одном сообществе.

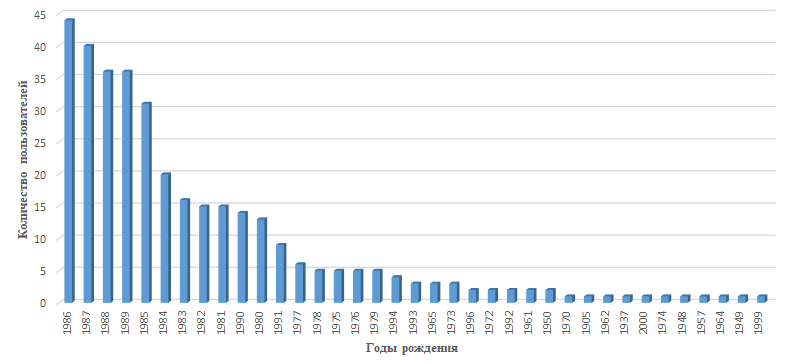


**Диаграмма 1.** Количество пользователей в скрытых сообществах

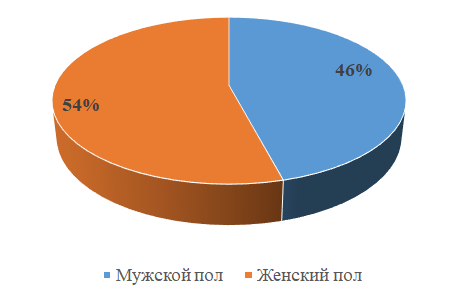
****

**Диаграмма 2.** Соотношение количества пользователей по количеству скрытых сообществ

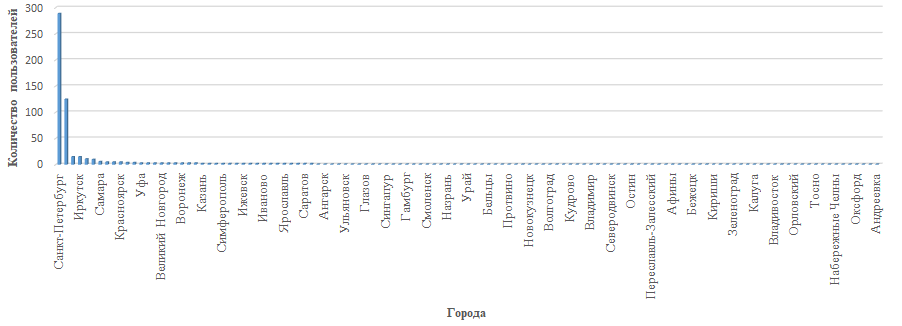
Диаграммы 3–5 показывают количественный анализ пользователей. После загрузки дополнительных данных пользователей с помощью vk\_api (имя, возраст, место жительства, группы и подписки) выяснилось, что многие пользователи проживают в крупных городах, но небольшая часть живёт и в других частях России: например, пользователь с ID 820468 живет в Туле, а пользователь с ID 70978256 живет в Санкт-Петербурге, они являются членами скрытого сообщества, посвящённого фестивалям. Небольшое количество пользователей проживает за территорией Российской Федерации. Средний возраст пользователей – 40 лет, при этом 365 пользователей скрывают год своего рождения.



**Диаграмма 3.** Распределение пользователей по годам рождения

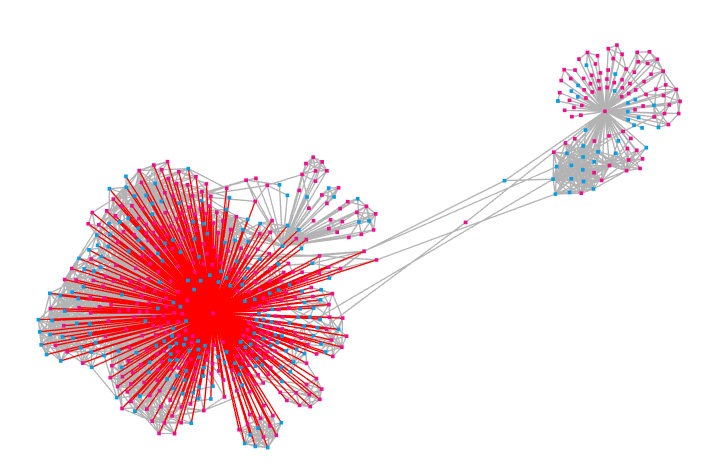
****

**Диаграмма 4.** Соотношение пользователей мужского и женского полов

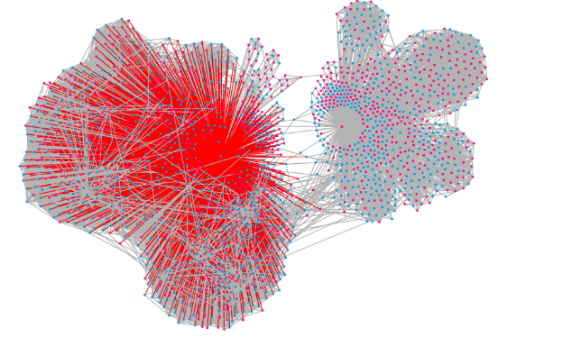
****

**Диаграмма 5.** Распределение пользователей по городам проживания

Методы, основанные на графах, позволяют определить расстояние между членами скрытых сообществ, таки образом, можно проверить теорию шести рукопожатий, выдвинутую американскими психологами С. Милгрэмом и Д. Тревэрсом. Суть идеи заключается в том, что все люди на Земле разделены уровнями общих знакомых, значение которых не превышает шести [Buchanan, 2002; Ипатов, Калагин, Кревецкий, 2018]. В 2018 году социальная сеть ВКонтакте предоставила доступ к сервису, который помогает проверить данную теорию[[37]](#footnote-37). В нашем исследовании мы выбрали онлайн-сервис[[38]](#footnote-38), который помогает визуализировать расстояние между пользователями. Пользователи 3854113 и 4817853, состоящие в скрытом сообществе «литература» находятся на расстоянии двух друзей в социальной сети (центры графов – целевые пользователи). Пользователи 56782587 и 18379994 из сообщества, посвящённого музыке, также находятся на расстоянии двух друзей. Если два графа не связаны, то между пользователями не установлены явные отношения, поэтому требуется добавить третьего пользователя для определения расстояния.



**Рис. 11.** Расстояние между пользователями 3854113 и 4817853



**Рис. 12.** Расстояние между пользователями 56782587 и 18379994

При анализе пересекающихся групп скрытых сообществ было установлено, что среди пользователей популярны новостные и образовательные группы (рис. 13): *Мир Discovery*, *E-squire* и др.



**Рис. 13.** Примеры пересекающихся групп пользователей скрытых сообществ

Члены скрытых сообществ редко являются членами одного и того же реального сообщества. Пользователь 878 скрытого сообщества «цифровая среда» – член сообщества *ВКонтакте для сообществ и SMMщики*. Пользователь 190868 из этого же скрытого сообщества входит в другую группу – *ЮТУБЕР*. Они оба являются членами группы *Стартапы и бизнес*, однако они не вошли в скрытое сообщество, посвящённое бизнесу. Возможно, что пользователи в социальных сетях делают меньше репостов из тех групп, в которых они не так сильно заинтересованы.

# **4.5. Сравнение различных методов выявления скрытых сообществ, оценка и анализ результатов**

Для оценки созданного подхода необходимо провести сравнение с уже существующими алгоритмами, поэтому мы прибегли к следующим методам: графовый подход, основанный на мере Жаккара, кластерный метод с привлечением doc2vec. Ниже представлено краткое описание построения моделей скрытых сообществ.

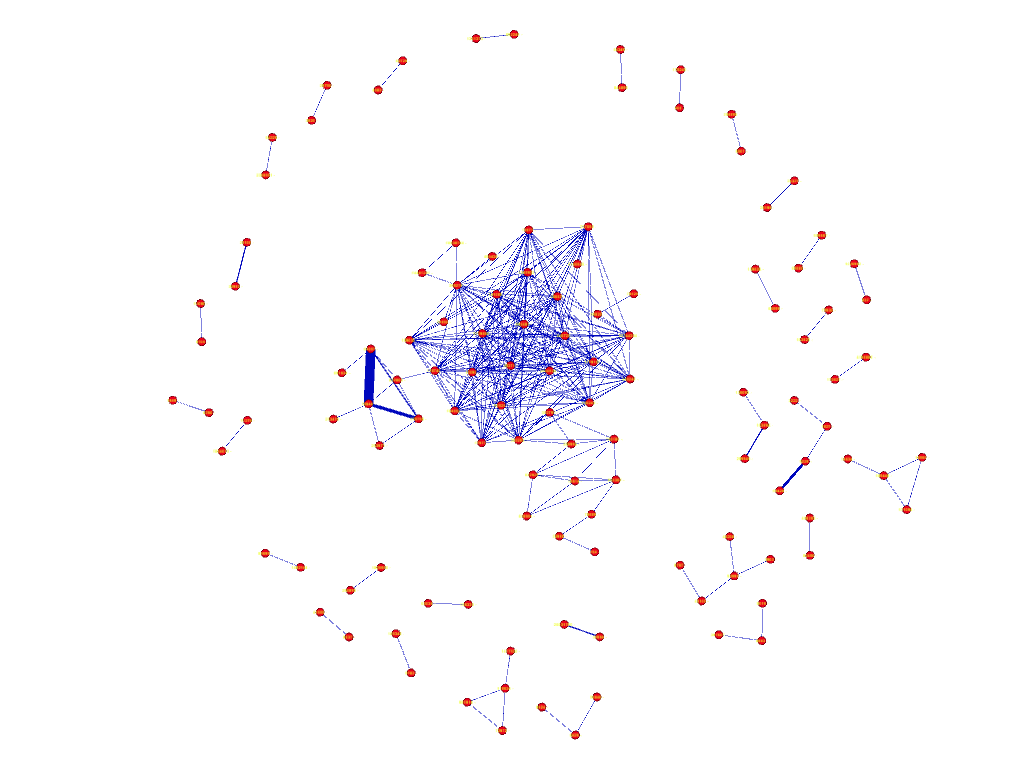
**Графовый метод**

Корпус постов состоит из более чем 25000 пользовательских текстов. Для удобства анализа мы представили их в виде *VK\_USER\_X\_POST\_Y*, где *X* – id пользователя, *Y* – порядковый номер поста. С помощью Python 3.7 был создан скрипт для расчёта значений меры Жаккара. Мы посчитали необходимым установить значение 0,75 минимальный порог сходства, т.е. ¾ содержания постов должно совпадать. Пары пользователей, которые удовлетворяли данным требованиям, были записаны в файл с расширением .csv. С помощью приложения Gephi был построен результирующий граф, отображающий состояние скрытых сообществ.

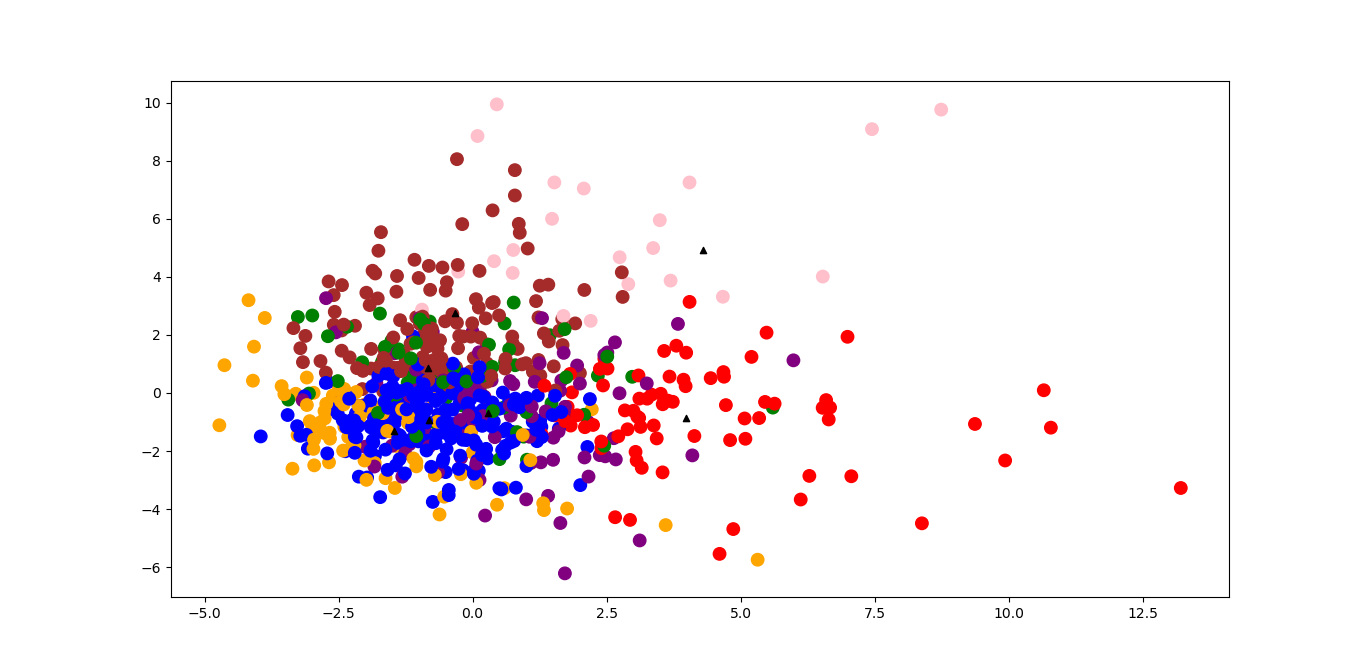
**Кластерный анализ**

Выделают огромное количество методов кластерного анализа: k-средних, DBSCAN, иерархический метод и т.д. В нашем исследовании для работы с большим количеством данных мы возьмём k-средних из библиотеки scikit-learn[[39]](#footnote-39) благодаря достаточно высокой степени работы и наглядности реализации. Для улучшения качества полученных моделей мы предварительно обучили модель doc2vec корпуса со следующими параметрами: размер контекстного окна – 5, размерность вектора – 100.

На рисунках 14–15 представлены две модели скрытых сообществ, выделенные методом Жаккара и кластерным методом.



**Рис. 14.** Скрытые сообщества, выделенные графовым методом

****

**Рис. 15.** Кластеры скрытых сообществ.

С помощью меры Жаккара было выявлено 344 пар семантически близких постов, которые сгруппировались в 34 сообщества. На рисунке 14 представлена фрагментарная структура взаимосвязей, существующих в корпусе. Первый очевидный результат графического анализа – это возможность выделения нескольких подмножеств: гигантский компонент (большая часть центральных узлов графа, которые связаны между собой) и отдельно стоящие узлы. Наличие большого количества пересечений свидетельствует о тесной взаимосвязи в рамках данного семантического поля.

Стоит отметить, что названия тематических скрытых сообществ не отображаются при их формировании, в результате чего дальнейший анализ осуществляется вручную. Так, например, пользователи 696481, 39371 и 2500528 входят в центральное скрытое сообщество. После анализа постов было установлено, что, несмотря на различие в тематиках, у всех есть общая черта – посты ссылались на конференции Zoom. При дальнейшем анализе у пользователей данного сообщества также появлялась информация про данную платформу.

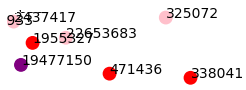
Если проанализировать фрагментарные сообщества, то они относятся к более стабильным общественным темам. Пользователей 981916, 118634, 72972 и 1417120 объединяет концертная тематика.

Аналогичная ситуация и у пользователей 62729 и 4583: их объединяет тема, посвящённая здоровью человека. При этом пользователи 1361373 и 892245 также интересуются здравоохранением, но они не были объединены в единую группу с вышеуказанными пользователями.

Вышеуказанные факторы могут указывать на следующие особенности при работе с графовыми методами.

1. Эти методы частично позволяют отслеживать современные тенденции в обществе. В конкретном случае это касается онлайн-встреч, вызванных пандемией коронавируса: пользователь 696481 проводил дистанционные встречи, посвящённые религиозным темам, а пользователь 2500528 комментировал дистанционное образование.
2. При более равномерном течении жизни в обществе итоговая картина могла бы выглядеть более раздробленной, при этом пользователи одного сообщества не имели бы связей. Это свидетельствует о невозможности учитывать семантические вариации постов.

Метод k-средних создаёт из набора рассматриваемых объектов k наиболее оптимальных групп, однако природа этих групп заранее неизвестна. В результате процесс кластеризации необходимо повторять некоторое количество раз с разными параметрами, чтобы найти наиболее устойчивый вариант скрытых сообществ. Было выявлено 7 групп скрытых сообществ. Для более детального анализа графическим данным были присвоены метки, соответствующие id пользователя (рисунок 16).



**Рис. 16.** Пример меток данных

Как мы уже отмечали, метод k-средних не требует больших временных затрат, он подходит для предварительных процедур, после которых требуется более мощные алгоритмы. В то же самое время данный алгоритм имеет ряд недочётов при анализе скрытых сообществ. Во-первых, алгоритм чувствителен к выбросам: т.е. теоретически те пользователи, темы которых не совпадают с другими, так или иначе будут отнесены к какому-либо классу. Пользователь 1415502 активно интересуется субкультурами, однако другие пользователи из корпуса не пишут посты на эту же тему. При этом методом k-средних было определено, что этот пользователь состоит в одном тематическом сообществе с пользователем 167175, который увлекается играми.

Также необходимо отметить, что алгоритм не справляется с задачей, при которой объекты могут принадлежать к разным тематическим группам. В отличие от графового метода, все пользователи могут входить только в одно скрытое сообщество. Так, например, посты пользователя 1955327 посвящены книгам и здоровью, хотя на рисунке 16 он отнесён только к одной тематической группе. Возможным решением проблемы может быть комбинирование различных методов кластеризации. В частности, алгоритм DBSCAN будет учитывать «шумы» при построении моделей, в то время как алгоритм c-средних – метод нечёткой кластеризации, улучшение метода k-средних – вместо однозначного ответа приписывает объектам вероятность вхождения в кластеры, что эквивалентно некоторым элементам модели, основанной на графах, когда одна вершина может входить в несколько подграфов.

При работе с гибридным подходом также необходимо отметить достоинства и недостатки. Многие алгоритмы тематического моделирования и их вариации уже успешно адаптированы для работы с русскоязычными коллекциями текстов социальных сетей. Также необходимо отметить, что при улучшении интерпретации моделей применялся алгоритм двойного ранжирования (Google PageRank и ipm в «Новом частотном словаре русской лексики» О.Н. Ляшевской и С.А. Шарова[[40]](#footnote-40)), чтобы избежать низкочастотных меток тем. В итоговой модели сообществ пользователи могут относиться к разным тематическим группам, так и обладать уникальными интересами, которые не пересекаются с интересами других людей.

В то же самое время мы столкнулись с определёнными сложностями. Так, например, в разделе 4.1 приведено описание разработки токенизации и лемматизации текстов социальных сетей. Особенности написания слов или их графического представления заставляло нас прибегать к постоянному улучшению алгоритмов или ручному редактированию текстов (удаление диакритических символов, приведение слова к словарному написанию и т.д.).

Алгоритмы автоматического назначения меток тем в русском и английских языках до сих пор не обладают стандартом, который позволил бы оценить качество работы разработанных методов. В разделе 4.3 мы указывали, что для оценки меток использовали сервис Google Формы и 10 независимых экспертов. Дальнейшие разработки могут быть связаны с созданием золотого стандарта, что позволит существенно сэкономить время при построении моделей скрытых сообществ.

Все вышеописанные достоинства и недостатки трёх подходов можно представить в едином формате. Ниже мы приводим сводную таблицу методов по определённым параметрам.

|  |  |  |  |
| --- | --- | --- | --- |
| **Параметр** | **Графовый метод** | **Кластерный анализ** | **Гибридный метод** |
| *Экономия времени* | *+* | *+* | *-* |
| *Возможность принадлежности пользователей к нескольким сообществам* | *+* | *В зависимости от алгоритма* | *+* |
| *Учёт лексико-семантических свойств* | *-* | *-* | *+* |
| *Возможность узнать наименование скрытого сообщества* | *-* | *-* | *+* |

**Таблица 4.** Сравнение методов выявления скрытых сообществ

Таким образом, разработанный нами алгоритм имеет большие преимущества перед уже существующими, однако процесс обработки данных требует большого количества времени, что в дальнейшем можно оптимизировать.

**Выводы к четвёртой главе**

После построение корпуса постов социальной сети ВКонтакте был проведён ряд семантических экспериментов, на основании которых можно сделать следующие выводы:

1. Отличительной особенностью текстов постов является использование ненормативных правил написания, что приводит к трудностям при автоматической обработке текстов, поэтому алгоритмы токенизации и нормализации приходится постоянно дорабатывать.

2. В результате построения тематической модели постов и процедуры автоматического назначения меток тем было установлено, что би- и триграммные кандидаты лучше объясняют узконаправленные темы, в то время как униграммные кандидаты подходят для общих тем.

3. Некоторые пользователи социальных сетей могут присутствовать в нескольких скрытых сообществах, что объясняется желанием людей оставлять записи, посвящённым разным тематикам.

4. Алгоритмы кластерного анализа и графовые методы выделяют рассредоточенные скрытые сообщества, что объясняется неполноценным учётом семантики исследуемых текстовых коллекций.

**ЗАКЛЮЧЕНИЕ**

На сегодняшний день люди активно пользуются социальными сетями: они обсуждают бытовые или рабочие дела, играют или ищут различный контент. Наше знание о способах интернет-общения растёт быстро, мы начинаем замечать, что между всеми вещами в интернет-пространстве существует гораздо большая взаимосвязь, чем мы предполагали ранее. Такая же взаимосвязь существует и между пользователями. Обнаружение скрытых сообществ и присущих им особенностей помогает нам всесторонне понять окружающее нас общество и объяснить сложные социальные явления.

Многие исследователи при построении алгоритмов поиска скрытых сообществ часто используют графовые методы. Объединение тематического моделирования и автоматического назначения меток тем позволяет восполнить пробелы в теории выявления скрытых сообществ.

В данной выпускной квалификационной работе был предложен гибридный алгоритм обнаружения сообществ русскоязычной социальной сети, были описаны преимущества и недостатки, а также проведён сравнительный анализ с уже существующими методами. Результаты позволяют получить новую информацию о возможных социальных группировках в интернет-пространстве. Качественная оценка процедур показала, что, несмотря на трудности при работе с постами социальных сетей, они сопоставимы с методами, основанными на математических моделях. Гипотеза, выдвинутая в работе, подтвердилась.

Использование данных из социальных сетей позволяет решить ряд смежных задач: например, отслеживание тенденций среди пользователей и, как следствие, создание соответствующего контента. Дальнейшие исследования будут связаны:

* с увеличением корпуса, включая посты других российских онлайн-платформ, что позволит обнаружить скрытые сообщества пользователей в различных социальных сетях;
* с привлечением других алгоритмов тематического моделирования: например, динамическое тематическое моделирование позволит сформировать динамические структуры скрытых сообществ и оценить изменение интересов пользователей в разные временные периоды;
* с улучшением алгоритма автоматического назначения меток тем и созданием «золотого стандарта» для оценки извлечённых кандидатов.

# **СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ**

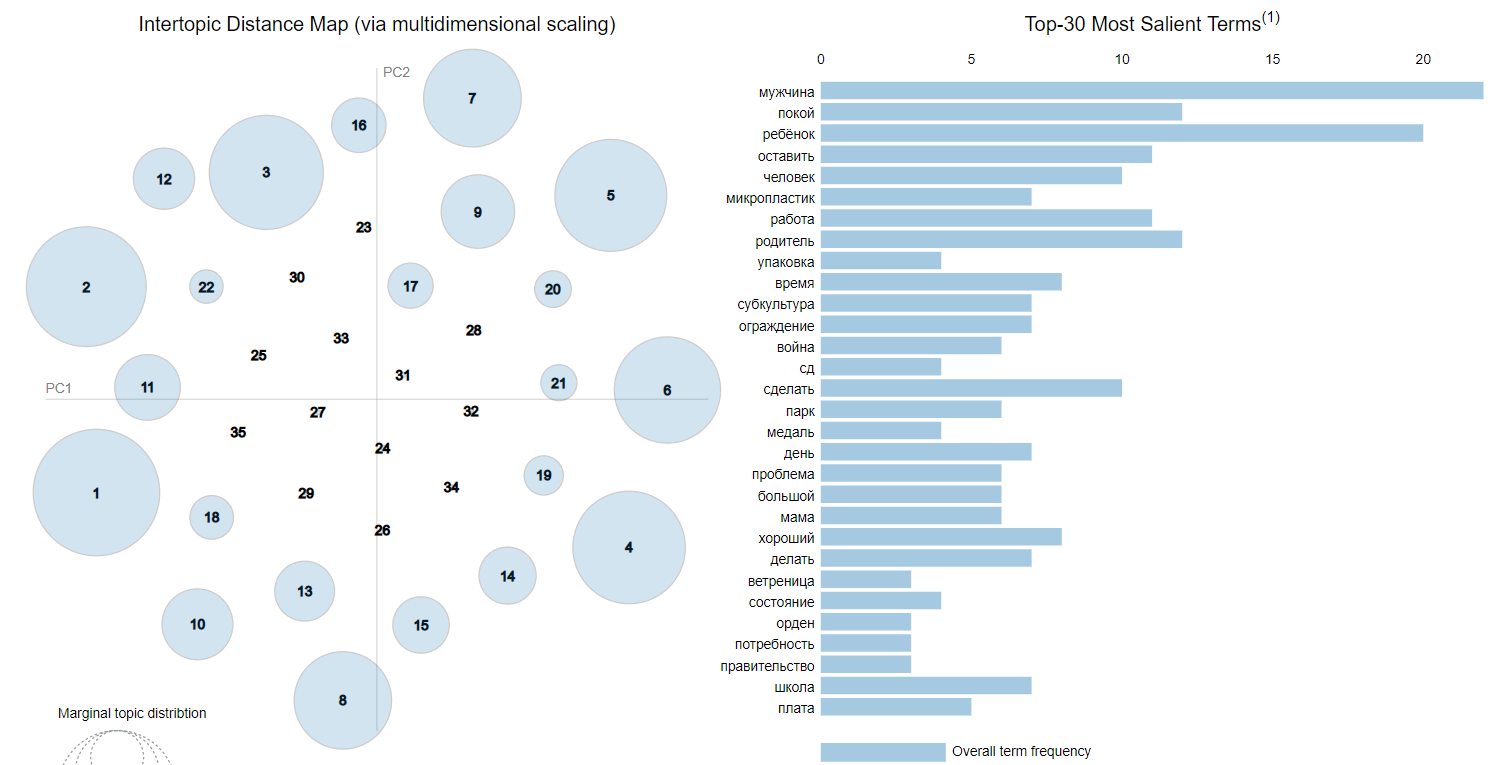
1. Бодулева А.Р. Лингвистические особенности СМС-сообщений английского языка / А.Р. Бодулева, А.З. Зарипова // Инновационная наука. — 2016. — №2 (5). — С. 76–78.
2. Вебер К.С. Сравнительный анализ социальных сетей / К.С. Вебер, А.А. Пименова // Вестник Тамбовского университета. Серия «Естественные и технические науки». — 2014. — Т. 19. Вып. 2. — С. 634–636.
3. Воронцов К.В. Вероятностное тематическое моделирование. [Электронный ресурс] URL: <http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf> (дата обращения: 21.01.2021).
4. Глушков Н.А. Анализ методов тематического моделирования текстов на естественном языке / Н.А. Глушков // Молодой учёный. — 2018. — №19. — С. 101–103.
5. Ерофеева А.Р. Автоматическое назначение меток тем в тематических моделях русскоязычных корпусов текстов / А.Р. Ерофеева, О.А. Митрофанова // Структурная и прикладная лингвистика: межвузовский сборник. Выпуск 12. К 60-летию отделения прикладной, компьютерной и математической лингвистики СПбГУ. — 2019. — №12. — С. 122–147.
6. Иноземцева Н.В. Парцелляция как основная синтаксическая модель заголовков англоязычных статей по методической проблематике / Н.В. Иноземцева // Вестник Оренбургского государственного университета. — 2011. — №11 (130). — С. 114–118.
7. Ипатов Ю.А. Анализ динамических характеристик сложных графовых структур / Ю.А. Ипатов, И.В. Калагин, А.В. Кревецкий [и др.] // Известия высших учебных заведений. Приборостроение. — 2018. — №6 (2). — С. 511–516.
8. Калинина Л.В. Вербальная конкуренция в пространстве Интернета: реклама, рерайт, репост / Л.В. Калинина // Вестник Вятского государственного гуманитарного университета. — Киров, 2014. — №3. — С. 80–88.
9. Карасик В.И. О типах дискурса / В.И. Карасик // Языковая личность: институциональный и персональный дискурс: сб. науч. тр. — Волгоград, 2000. — С. 5–20.
10. Коломейченко А.А. Алгоритм выделения сообществ в социальных сетях / М.И. Коломейченко, А.А. Чеповский, А.М. Чеповский, // Фундаментальная и прикладная математика. — 2014. — №19:1. — С. 21–32.
11. Мамаев И.Д. Адаптация заимствованных слов к русской морфологии / И.Д. Мамаев, А.А. Зайцева // International Journal of Advanced Studies in Language and Communication. — 2019. — №2. — С. 106–113.
12. Матусевич А.А. Общение в социальных сетях: прагматический, коммуникативный, лингвостилистический аспекты характеристики: дис. … канд. филол. наук / А.А. Матусевич. — Киров, 2016. — 190 с.
13. Мейлахс П.А. Онлайновое сообщество СПИД-диссидентов в социальной сети «ВКонтакте»: структура и риторические стратегии / П.А. Мейлахс, Ю.Г. Рыков // XV апрельская международная научная конференция по проблемам развития экономики и общества: в 4-х книгах. Отв. ред.: Е. Г. Ясин. — Кн. 3. — М.: Издательский дом НИУ ВШЭ. — 2015. — С. 137-146.
14. Митрофанова О.А. Исследование структурной организации художественного произведения с помощью тематического моделирования: опыт работы с текстом романа «Мастер и Маргарита» М.А. Булгакова / О.А. Митрофанова // Труды международной конференции «Корпусная лингвистика–2019». — 2019. — С. 387–394.
15. Митрофанова О.А. Моделирование семантических связей в текстах социальных сетей с помощью алгоритма LDA (на материале русскоязычного сегмента Живого Журнала) / О.А. Митрофанова, А.С. Шиморина // Структурная и прикладная лингвистика: межвузовский сборник. — 2014. — №10. — С. 151–168.
16. Митрофанова О.А. Тематическое моделирование корпуса «Народных русских сказок А.Н. Афанасьева» / О.А. Митрофанова // Структурная и прикладная лингвистика. — 2015. — №11. — С. 146–154.
17. Митягин С.А. Исследование социальных сетей Интернет на предмет выявления сопутствующих интересов лиц, склонных к наркомании / С.А. Митягин, А.В. Якушев, А.В. Бухановский // Международный научно-исследовательский журнал. Технические науки. — 2012. — Вып. 6(6). — С. 59–64.
18. Морозова О.Н. Особенности Интернет-коммуникации: определение и свойства // О.Н. Морозова / Вестник ЛГУ им. А. С. Пушкина. Серия «Филология». — 2010. — №5. T.1. — С. 150–158.
19. Нашхоева М.Р. Взаимодействие пунктуационных знаков и эмотиконов в текстах форумов / М.Р. Нашхоева // Вестник Челябинского государственного педагогического университета. — 2011. — № 12. — С. 321–327.
20. Нокель М.А., Лукашевич Н.В. Тематические модели: добавление биграмм и учет сходства между униграммами и биграммами / М.А. Нокель, Н.В. Лукашевич // Вычислительные методы и программирование. — 2015. — №6 (2). — С. 215–234.
21. Орлова Н.В. Голосовые сообщения как источник сведений о коммуникативных нормах и ценностях / Н.В. Орлова // Экология языка и коммуникативная практика. — 2018. — №3 (14). — С. 57–66.
22. Рябцева Н.К. Интернет-коммуникация в лингвистическом и когнитивном аспекте / Н.К. Рябцева // Russian Linguistic Bulletin. — 2016. — №4 (8). — C. 70–73.
23. Смирнова Е.В. Гендерные различия в молодежном Интернет-дискурсе (на материале англоязычных Интернет-блогов и чатов) / Е.В. Смирнова // Лексикографическая копилка: сб. науч. ст. Вып. 8. — СПб.: Изд-во СПбГЭУ, 2019. — С. 73–79.
24. Смирнова О.С. Определение группы риска аккаунтов социальной сети «Вконтакте», попадающих под влияние квестовой игры суицидального характера / О.С. Смирнова // Современные информационные технологии и ИТ-образование. — 2017. — №13 (3). — С. 53–60.
25. Смородина А.А. Интернет-мемы как способ коммуникации человека в современном мире / А.А. Смородина // Международный журнал гуманитарных и естественных наук. — 2019. — №5 (3). — С. 78–82.
26. Хлопотов М.В. Исследование кластеров кинолюбителей и их тем в социальных сетях / М.В. Хлопотов, Н.В. Старцева, А.А. Макаренко // Вестник евразийской науки. — 2019. — №11 (2). — С. 1–11.
27. Холодковская Е.В. Особенности синтаксиса англоязычного интернет-комментария социальной сети Facebook / Е.В. Холодковская // Вестник Волгоградского государственного университета. Серия 2, Языкознание. — 2014. — №1 (20). — С. 79–83.
28. Хорошевский В.Ф. Семантические технологии в наукометрии: задачи, проблемы, решения и перспективы / В.Ф. Хорошевский, И.Е. Ефименко // Когнитивно-семиотические аспекты моделирования в гуманитарной сфере. Под редакцией В.Л. Стефанюка, Э.А. Тайсиной. — Академия наук РТ, Институт прикладной семиотики АН РТ. — Казань: Изд-во Академии наук РТ, 2017. — C. 222–266.
29. Щурина Ю.В. Интернет-мемы как феномен интернет-коммуникации / Ю.В. Щурина // Научный диалог. — 2011. — №3. — С. 160–172.
30. Ярцева В.Н. Лингвистический энциклопедический словарь / Гл. ред. В.Н. Ярцева. — М., 1990; DVD-версия. — М., 2008.
31. Acharya D. B. Community Detection Clustering via Gumbel Softmax / D.B. Acharya, H. Zhang // SN Computer Science. — 2020. — Vol. 1. — P. 1–11.
32. Aletras N. Representing topics using images / N. Aletras, M. Stevenson // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language. — 2013. — P. 158–167.
33. Arefyev N.V. Word2Vec not dead: predicting hypernyms of co-hyponyms is better than reading definitions / N.V. Arefyev, M.V. Fedoseev, A.V. Kabanov [et al.] // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (2020). — 2020. — Iss. 19. — P. 13–32.
34. Baumes J. Discovering Hidden Groups in Communication Networks / J. Baumes, M. Goldberg, M. Magdon-Ismail [et al.] // Intelligence and Security Informatics, Proceedings of the Second Symposium on Intelligence and Security Informatics, ISI 2004. — 2004. — P. 378-389.
35. Bhatia S. Automatic labelling of topics with neural embeddings / S. Bhatia, J.H. Lau, T. Baldwin // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. — 2016. — P. 953–963.
36. Blei D. Latent Dirichlet Allocation / D. Blei, A. Ng, M. Jordan // Journal of Machine Learning Research. — 2003. — Vol. 3. — P. 993–1022.
37. Bodrunova S. S. Topic Modeling for Twitter Discussions: Model Selection and Quality Assessment / S.S. Bodrunova, I.S. Blekanov, M. Kukarkin // Proceedings of the 6TH SWS International Scientific Conferences on Social Sciences 2019. — 2019. — P. 207–214.
38. Buchanan, M. Nexus: small worlds and the groundbreaking science of networks. — W.W. Norton & Company, 2002. — 235 р.
39. Chobe S. Advancing community detection using Keyword Attribute Search. / S. Chobe, J. Zhan // Journal of Big Data. — 2019. — Vol. 6. — P. 1–33.
40. Crystal D. Language and the Internet. Cambridge: Cambridge University Press. — 2001. — 272 p.
41. Fortunato S. Community detection in networks: A user guide / S. Fortunato, D. Hric // Physics Reports. — 2016. — Vol. 659. — P. 1–44.
42. Gmati H. A new algorithm for communities detection in social networks with node attributes / H. Gmati, A. Mouakher, A. Gonzalez-Pardo [et al.] // Journal of Ambient Intelligence and Humanized Computing. — 2018. — P. 1–13.
43. He K. Hidden community detection in social networks / K. He, Y. Li, S. Soundarajan [et al.] // Information Sciences. — 2018. — Vol. 425. P. 92–106.
44. Herring S.C. Grammar and electronic communication. // The encyclopedia of applied linguistics. — 2012. — P. 2338–2346.
45. Jia Y. CommunityGAN: Community Detection with Generative Adversarial Nets / Y. Jia, Q. Zhang, W. Zhang [et al.] // Proceedings of the World Wide Web. — 2019. — P. 784–794.
46. Koltsova O. Redefining Media Agendas: Topic Problematization in Online Reader Comments / O. Koltsova, O. Nagornyy // Media and communications. — 2019. — Vol. 7, iss. 3. — P. 145–156.
47. Kriukova A. Explicit Semantic Analysis as a Means for Topic Labelling / A. Kriukova, A. Erofeeva, O. Mitrofanova [et al.] // Artificial Intelligence and Natural Language. AINL 2018. Communications in Computer and Information Science. — Springer, Cham, 2018. — Vol. 930. — P. 110–116.
48. Krzemień A. Detection of Jihadism in Social Networks Using Big Data Techniques Supported by Graphs and Fuzzy Clustering / A. Krzemień, C. Sánchez-Rebollo, C. Puente [et al.] // Advances in Complex Systems and Their Applications to Cybersecurity. — 2019. — Vol. 2019. — P. 1–13.
49. Lau J.H. Automatic labelling of topic models / J.H. Lau, K. Grieser, D. Newman [et al.] // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. — 2011. — P. 1536–1545.
50. Li Y.H. Classification of Text Documents / Y.H. Li, A.K. Jain // The Computer Journal. — 1998. — Vol. 41, №. 8. — P. 537–546.
51. Liu F. Deep learning for community detection: progress, challenges and opportunities / F. Liu, S. Xue, J. Wu [et al.] // Proceedings of the 29th International Joint Conference on Artificial Intelligence, IJCAI 2020. — 2020. — P. 4981–4987.
52. López-Rúa P. Teaching L2 vocabulary through SMS language: some didactic guidelines // ELIA. — 2007. — Vol. 7. — P. 165–188.
53. Magatti D. Automatic Labeling Of Topics / D. Magatti, S. Calegari, D. Ciucci [et al.] // Ninth International Conference on Intelligent Systems Design and Applications. — 2009. — P. 1227–1232.
54. Manika L. Author-Topic Modeling of DESIDOC Journal of Library and Information Technology (2008-2017), India [Электронный ресурс] / L. Manika, M. Margam // Library Philosophy and Practice (e-journal). — 2019. URL: <https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=5983&context=libphilprac> (дата обращения: 10.12.2020).
55. Mehdi A.A knowledge-based topic modeling approach for automatic topic labeling / A.A. Mehdi, P. Seyedamin, K. Kochut [et al.] // International Journal of Advanced Computer Science and Applications. — 2017. —Vol. 8(9). — P. 335–349.
56. Mei Q. Automatic labeling of multinomial topic models. / Q. Mei, X. Shen, C. Zhai // Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. — 2007. — P. 490–499.
57. Mirzagitova A. Automatic Assignment of Labels in Topic Modelling for Russian Corpora / A. Mirzagitova, O. Mitrofanova // Proceedings of 7th ExLing 2016. — 2016. — P. 115–118.
58. Mishra N. Clustering Social Networks. / N. Mishra, R. Schreiber, I. Stanton [et al.] // Algorithms and Models for the Web-Graph. WAW 2007. Lecture Notes in Computer Science. — Springer, Berlin, Heidelberg, 2007. — Vol. 4863. — P. 56–67.
59. Mitrofanova O. Topic Modelling of the Russian Corpus of Pikabu Posts: Author-Topic Distribution and Topic Labelling / O. Mitrofanova, V. Sampetova, I. Mamaev [et al.] // International conference «Internet and Modern Society» (IMS-2020) Proceedings, International Workshop «Computational Linguistics» (CompLing-2020). In press.
60. Panicheva P. Semantic feature aggregation for gender identification in Russian Facebook / P. Panicheva, A. Mirzagitova, Y. Ledovaya // Artificial Intelligence and Natural Language. AINL 2017. Communications in Computer and Information Science. — Springer, Cham, 2017. — Vol. 789. — P. 3–15.
61. Ramage D. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora / D. Ramage, D. Hall, R. Nallapati [et al.] // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. — 2009. — Vol. 1. — P. 248–256.
62. Reihanian A. Topic-oriented community detection of rating-based social networks / A. Reihanian, B. Minaei-Bidgoli, H. Alizadeh // Journal of King Saud University. Computer and Information Sciences. — 2016. — Vol. 28, iss. 3. — P. 303–310.
63. Rytsarev I.A. Clustering of social media content with the use of BigData technology / I.A. Rytsarev, A.V. Kupriyanov, D.V. Kirsh [et al.] // Journal of Physics: Conference Series. — 2018. — Vol. 1096. — P. 1–7.
64. Salz D. Hidden Community Detection in Online Forums / D. Salz, N. Benavides, J. Li // CS224W: Machine Learning with Graphs. — 2019. — P. 1–10.
65. Singh K. Clustering of people in social network based on textual similarity. / K. Singh, H. K. Shakya, B. Biswas // Perspectives in Science. — 2016. — Vol. 8. — P. 570–573.
66. Smoliarova A.S. Detecting pivotal points in social conflicts via topic modeling of twitter content / A.S. Smoliarova, S.S. Bodrunova, A.V. Yakunin [et al.] // Internet Science. INSCI 2018. Lecture Notes in Computer Science. — Springer, Cham, 2018. — Vol. 11551. — P. 61–71.
67. Squires L. Enregistering internet language // Language in Society. — 2010. — 457–492 pp.
68. Stevens K. Exploring Topic Coherence over Many Models and Many Topics / K. Stevens, P. Kegelmeyer, D. Andrzejewski [et al.] // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. — 2012. — P. 952–961.
69. Tikhomirov M.M. Combined Approach to Hypernym Detection for Thesaurus Enrichment / M.M. Tikhomirov, N.V. Loukachevitch, E.A. Parkhomenko // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (2020). — 2020. — Iss. 19. — P. 736–746.
70. Yadrintsev V.V. Distributional Models and Auxiliary Methods for Determining the Hypernyms of Words in Russian / V.V. Yadrintsev, A.A. Ryzhova, I.V. Sochenkov // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (2020). — 2020. — Iss. 19. — P. 762–772.

# **СПИСОК ЭЛЕКТРОННЫХ РЕСУРСОВ**

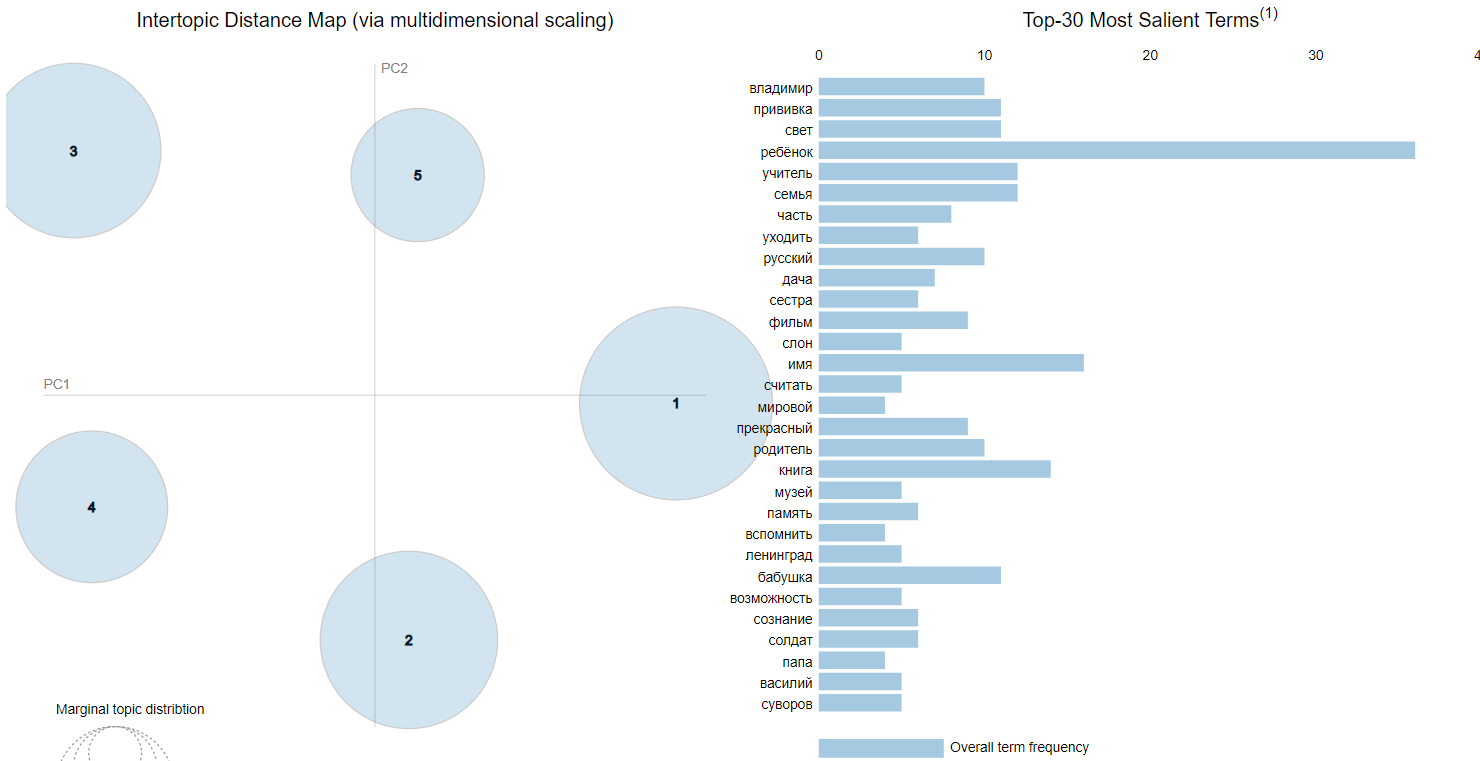
1. Ваша сеть ВКонтакте. Одной картинкой [Электронный ресурс]. URL: <http://www.yasiv.com/vk> (дата обращения: 12.04.2020)
2. Викисловарь [Электронный ресурс]. URL: <https://ru.wiktionary.org/wiki/%D0%97%D0%B0%D0%B3%D0%BB%D0%B0%D0%B2%D0%BD%D0%B0%D1%8F_%D1%81%D1%82%D1%80%D0%B0%D0%BD%D0%B8%D1%86%D0%B0> (дата обращения: 11.09.2020)
3. ВКонтакте [Электронный ресурс]. URL: <https://vk.com/> (дата обращения: 22.07.2020)
4. Жизнь в Кудрово [Электронный ресурс] // ВКонтакте. URL: <https://vk.com/kudrovolife> (дата обращения: 18.01.2020)
5. Морфологический анализатор pymorphy2 [Электронный ресурс]. URL: <https://pymorphy2.readthedocs.io/en/latest/> (дата обращения: 17.04.2020)
6. Национальный корпус русского языка [Электронный ресурс]. URL: <https://ruscorpora.ru/new/> (дата обращения: 12.04.2020)
7. Новости для прессы [Электронный ресурс] // ВКонтакте. URL: <https://vk.com/press/press-vk-12> (дата обращения: 12.04.2020)
8. Новый частотный словарь русской лексики [Электронный ресурс]. URL: <http://dict.ruslang.ru/freq.php> (дата обращения: 06.03.2020)
9. Описание методов API [Электронный ресурс] // VK Developers. URL: <https://vk.com/dev> (дата обращения: 10.04.2020)
10. Открытый корпус [Электронный ресурс]. URL: <http://opencorpora.org/> (дата обращения: 10.04.2020)
11. Программное обеспечение [Электронный ресурс] // Лаборатория интернет-исследований. URL: <https://linis.hse.ru/soft-linis/> (дата обращения: 03.04.2020)
12. Реставрация книг, переплётное дело [Электронный ресурс] // Alib.ru - букинистические книги. URL: <https://www.alib.ru/forumarhtema.phtml?forumbeg=150&arhtema=apteka> (дата обращения: 11.11.2020)
13. Тезаурус русского языка RuWordNet [Электронный ресурс]. URL: <https://ruwordnet.ru/ru> (дата обращения: 23.04.2020)
14. Beautiful Soup Documentation [Электронный ресурс]. URL: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/> (дата обращения: 21.04.2020)
15. Facebook [Электронный ресурс]. URL: <https://www.facebook.com/> (дата обращения: 12.05.2020)
16. Gensim. Topic modelling for humans [Электронный ресурс]. URL: <https://radimrehurek.com/gensim/> (дата обращения: 12.04.2020)
17. Gephi. Makes graphs handy [Электронный ресурс]. URL: <https://gephi.org/> (дата обращения: 15.04.2020)
18. GitHub [Электронный ресурс]. URL: <https://github.com/> (дата обращения: 06.12.2020)
19. Google Формы [Электронный ресурс]. URL: <https://www.google.com/intl/ru_ru/forms/about/> (дата обращения: 17.04.2020)
20. Matplotlib. Version 3.3.3 [Электронный ресурс]. URL: <https://matplotlib.org/> (дата обращения: 02.07.2020)
21. Orange [Электронный ресурс]. URL: <https://orangedatamining.com/> (дата обращения: 10.07.2020)
22. Python 3.7.0 [Электронный ресурс] // Python. URL: <https://www.python.org/downloads/release/python-370/> (дата обращения: 01.02.2020)
23. Requests: HTTP for Humans [Электронный ресурс]. URL: <https://requests.readthedocs.io/en/master/> (дата обращения: 09.03.2020)
24. Scikit-learn. Machine learning in Python [Электронный ресурс]. URL: <https://scikit-learn.org/stable/> (дата обращения: 15.03.2020)
25. Stanford Topic Modeling Toolbox [Электронный ресурс] // The Stanford Natural Language Processing Group. URL: <https://nlp.stanford.edu/software/tmt/tmt-0.4/> (дата обращения: 10.07.2020)
26. The Ideal Length for Instagram, Facebook, Twitter, & LinkedIn Posts [Электронный ресурс] // HubSpot. URL: <https://blog.hubspot.com/marketing/character-count-guide> (дата обращения: 25.03.2020)
27. Topic Modeling [Электронный ресурс] // MAchine Learning for LanguagE Toolkit. URL: <http://mallet.cs.umass.edu/topics.php> (дата обращения: 10.07.2020)
28. Vandrouki | Путешествия почти бесплатно (RU) [Электронный ресурс] // ВКонтакте. URL: <https://vk.com/vandroukiru> (дата обращения: 10.04.2020)

# **ПРИЛОЖЕНИЕ 1. ПРИМЕРЫ НЕКОТОРЫХ ВИЗУАЛИЗИРОВАННЫХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ**

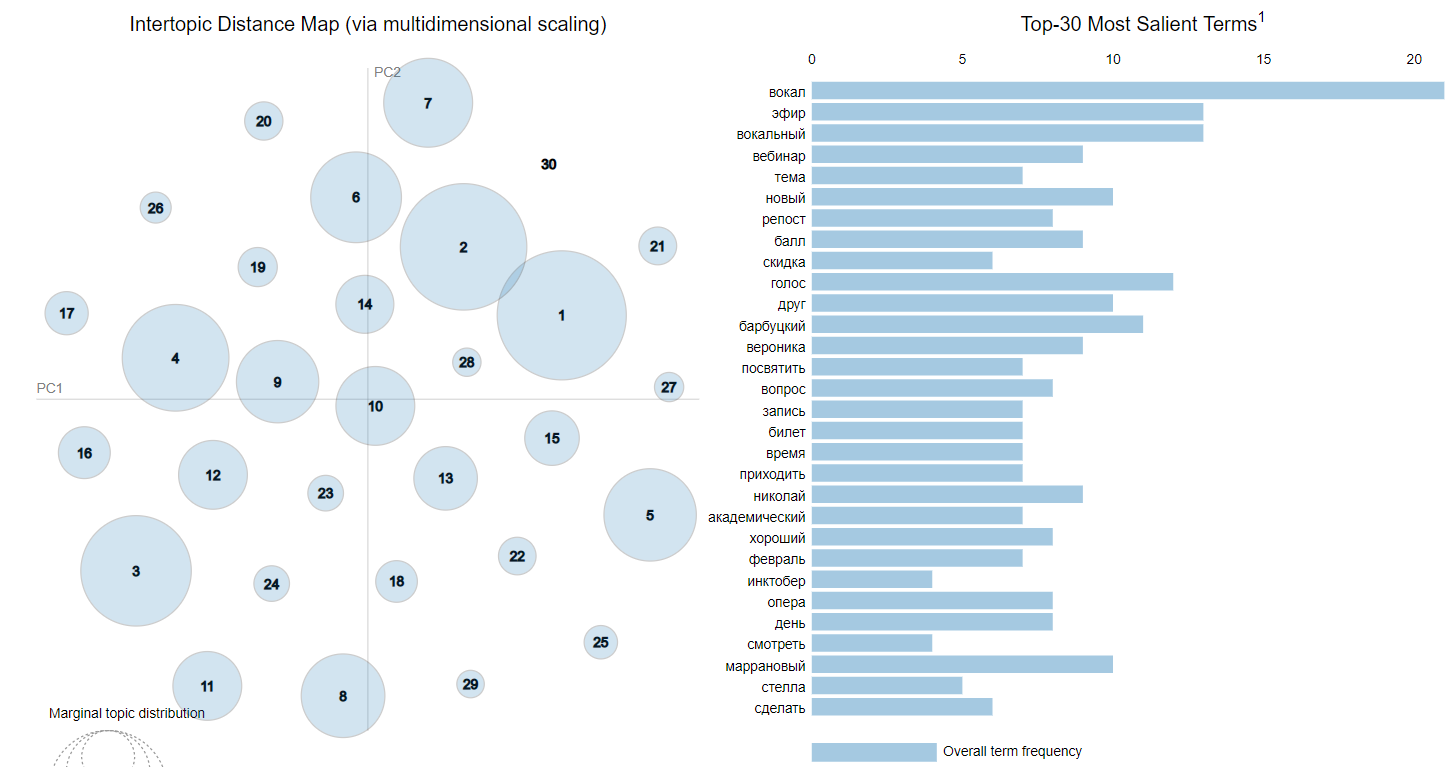
Темы пользователя с ID 1449.



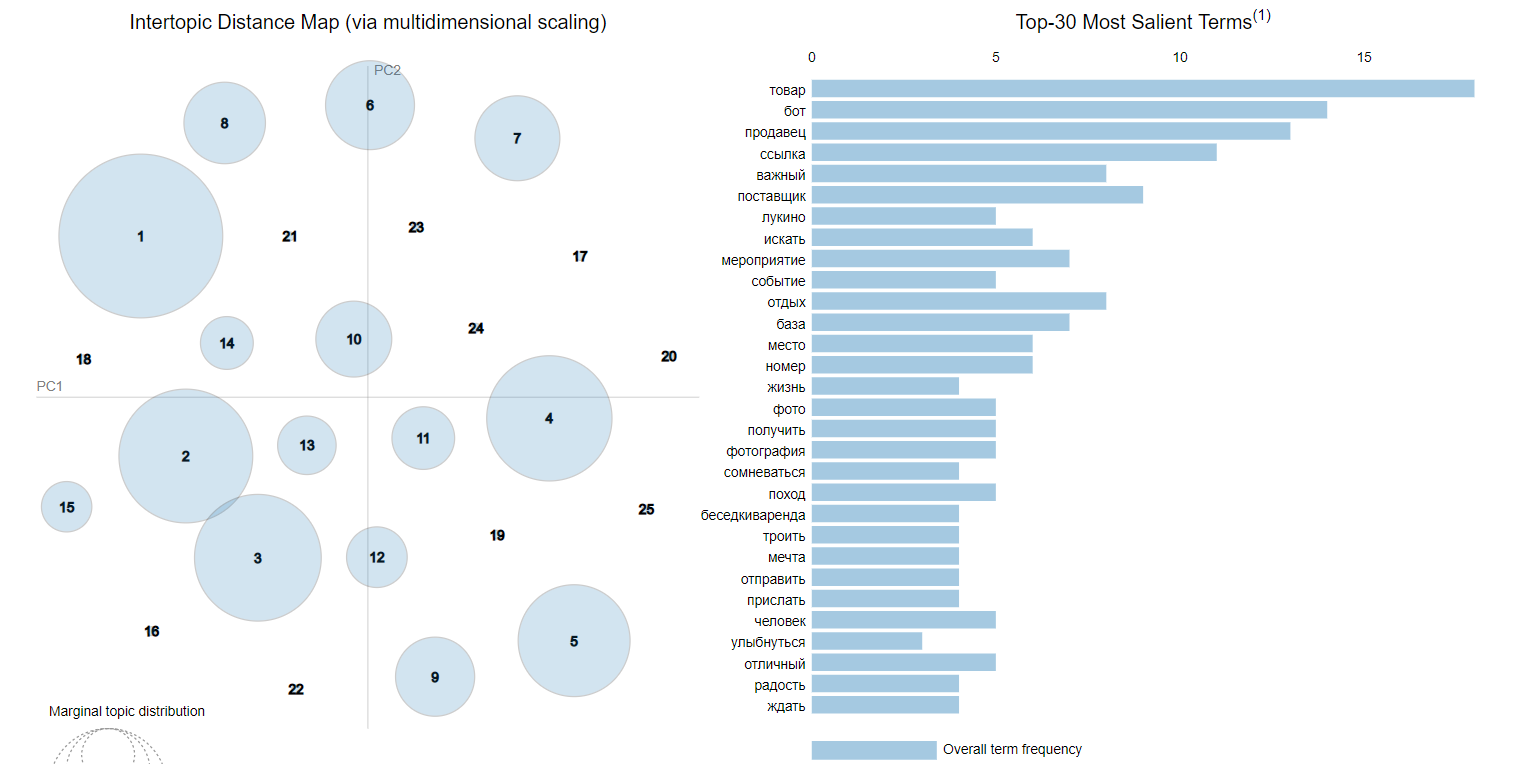
Темы пользователя с ID 103770.

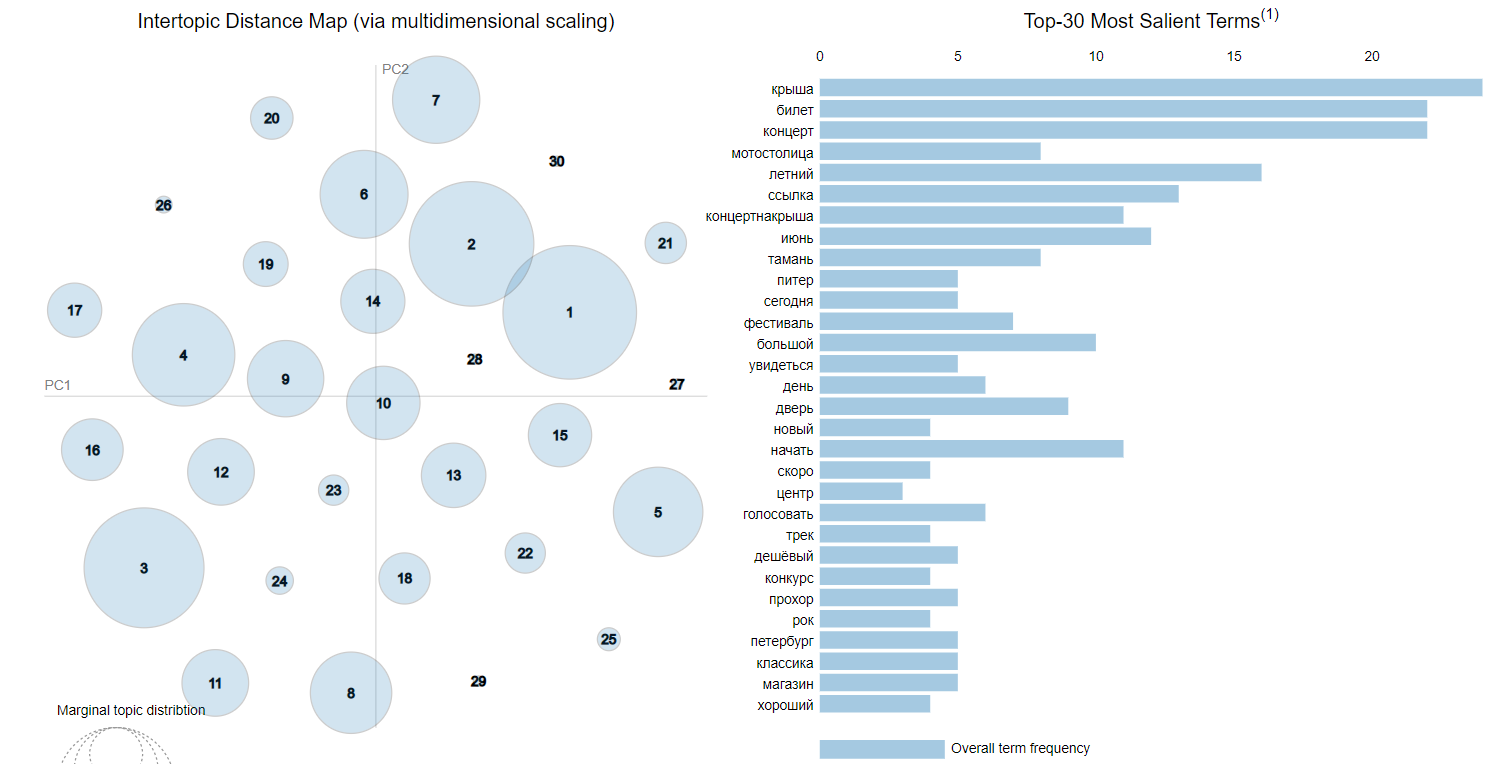


Темы пользователя с ID 138249.

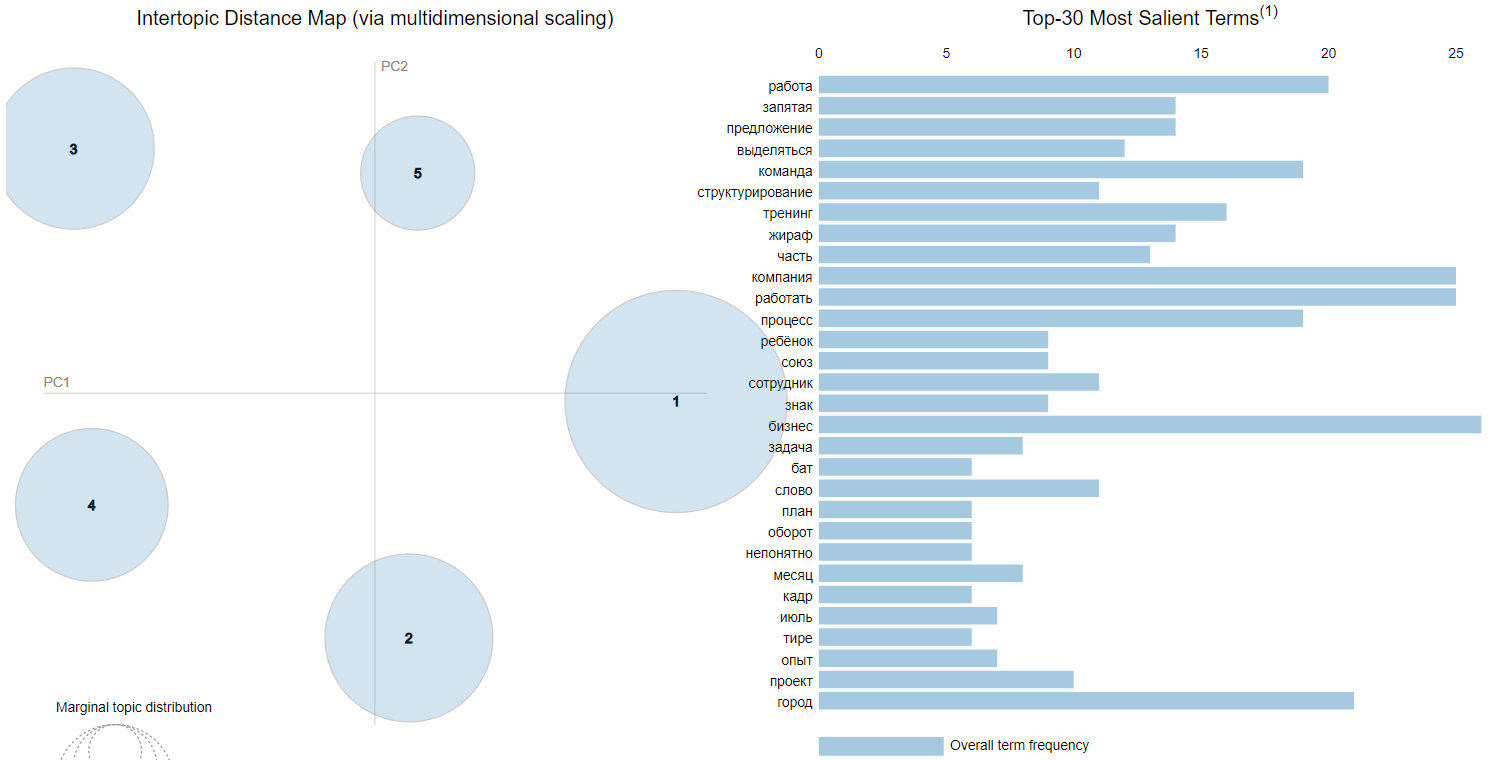


Темы пользователя с ID 1218605.

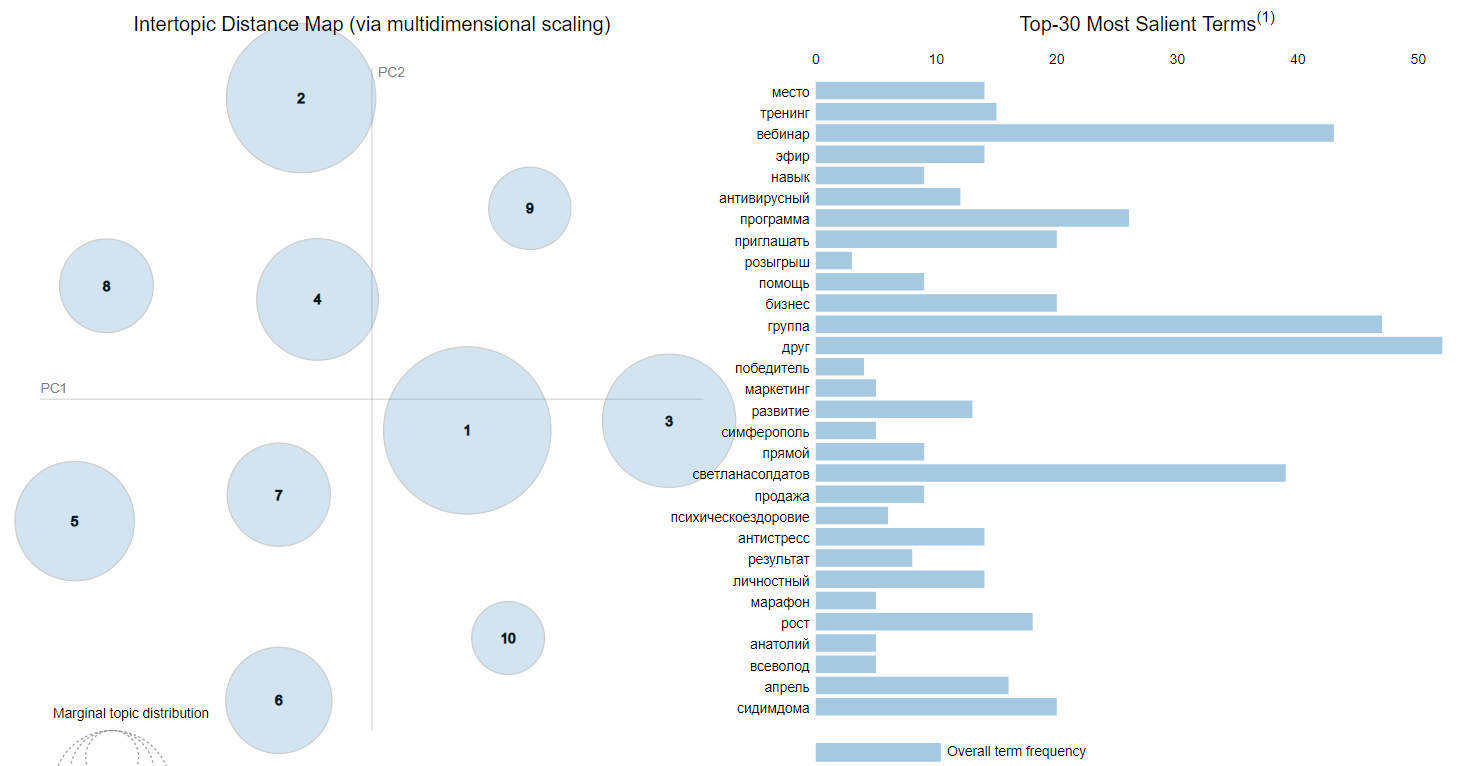


Темы пользователя с ID 124052.

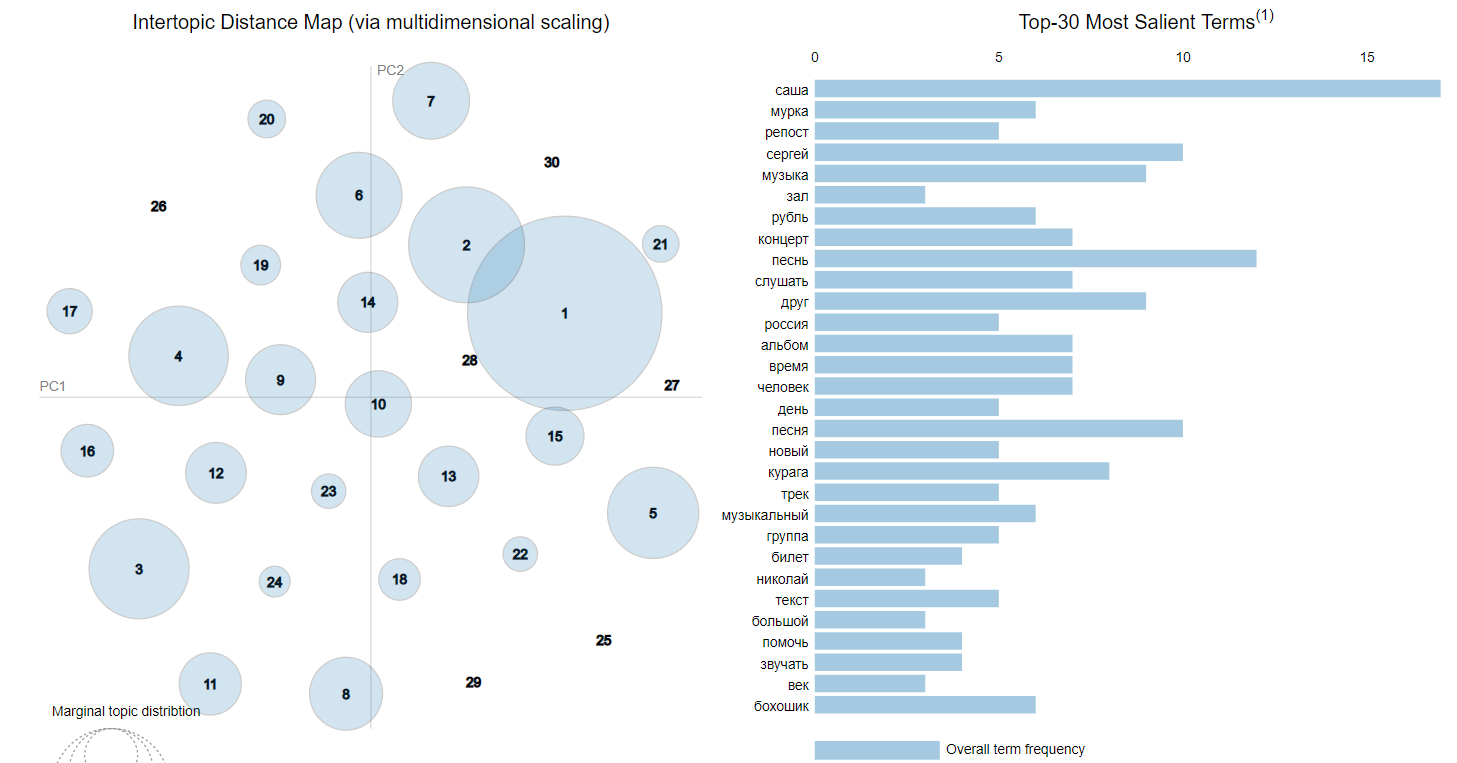
Темы пользователя с ID 14560.



Темы пользователя с ID 13977.



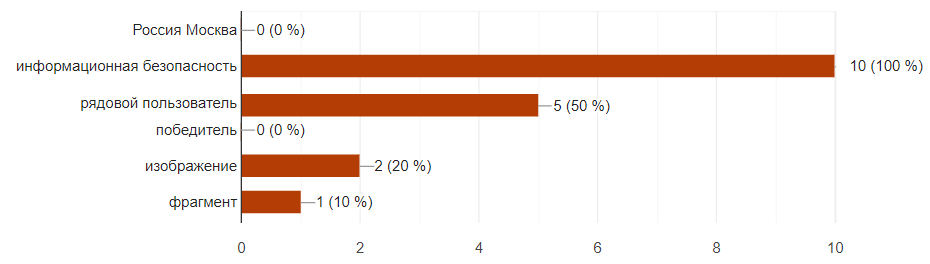
Темы пользователя с ID 1059933.



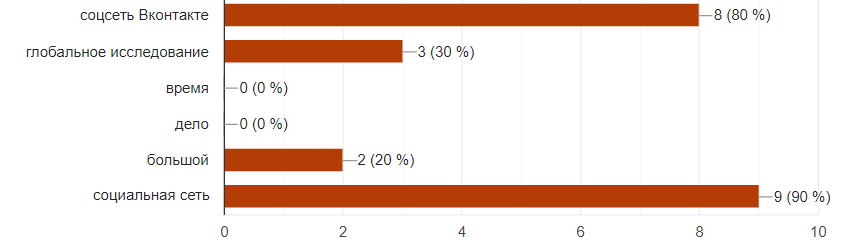
# **ПРИЛОЖЕНИЕ 2. РЕЗУЛЬТАТЫ ЭКСПЕРТНЫХ ОЦЕНОК ДЛЯ КАНДИДАТОВ В МЕТКИ ТЕМ**

В данном приложении можно ознакомиться с результатами экспертных оценок автоматически извлечённых меток тем для случайной выборки пользователей. Оценочная форма расположена на сайте Google Формы[[41]](#footnote-41).

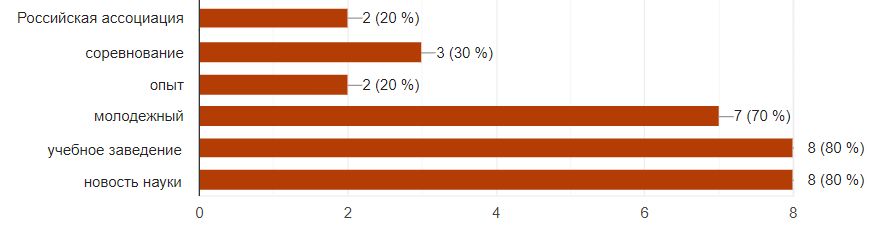
Пользователь 834: *адрес, внимание, страница, надёжно, ссылка, взламывать, скоро, перевод, фотография, страничка*.



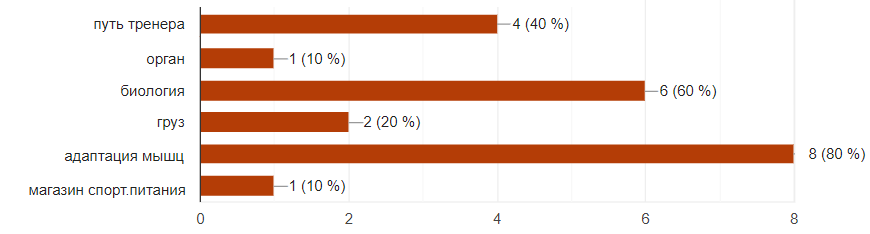
Пользователь 878: *книга, вконтакте, год, мочь, большой, новый, реклама, соцсеть, аудитория, канал*.



Пользователь 57296: *конкурс, университет, государственный, молодёжный, проект, научный, российский, год, имя, студенческий.*



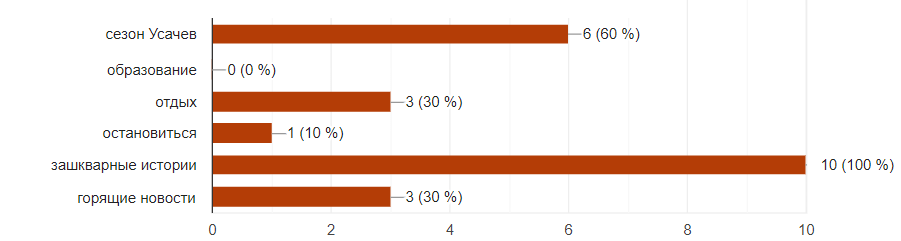
Пользователь 94855: *мышца, рост, минимальный, адаптационный, нагрузка, максимум, выяснить, спешить, гипертрофия, учёный.*



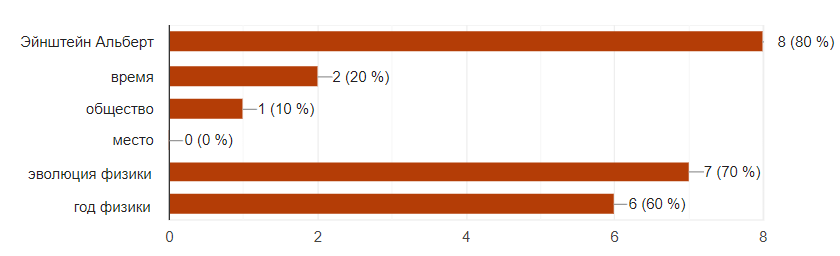
Пользователь 99607: *фотограф, причёска, макияж, яркий, образ, фото, маргарита, пивоваров, прекрасный, фотостудия.*



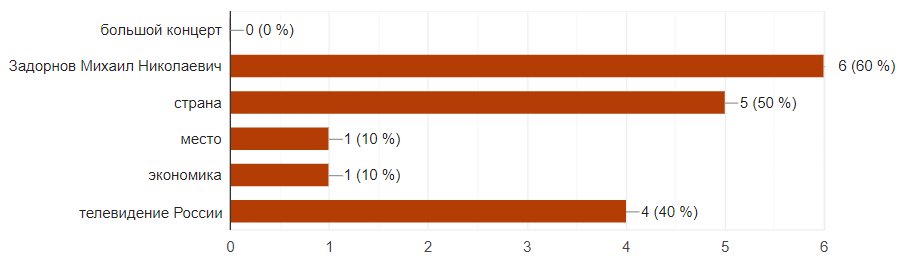
Пользователь 190868: *кликклак, выпуск, премьера, просмотр, забывать, приятный, руслан, усачев, кшиштовский, история.*



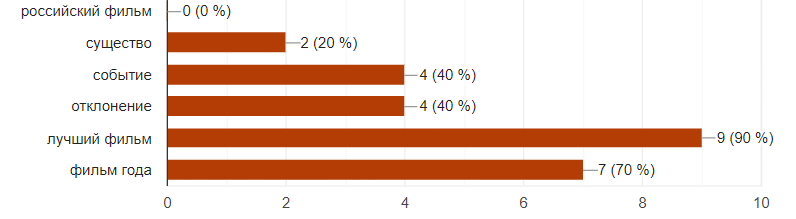
Пользователь 385450: *цивилизация, год, эйнштейн, квантовый, картина, мир, человек, место, город, религия.*



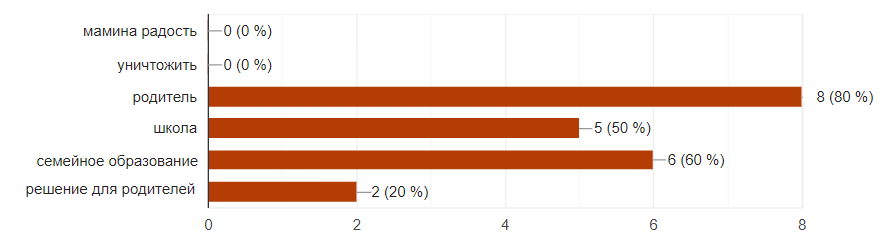
Пользователь 554429: *год, россия, страна, человек, вакцина, народ, задорнов, роман, мир, русский.*

**

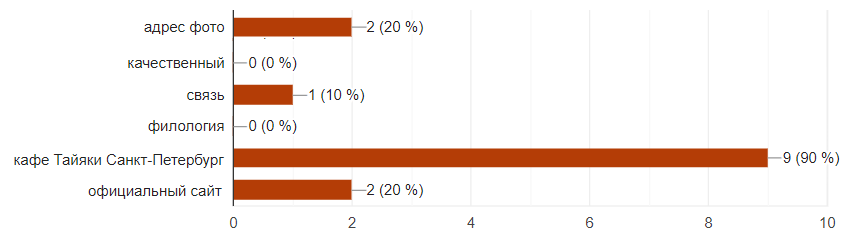
Пользователь 820468: *человек, джокер, оказаться, обстоятельство, убить, язык, несколько, изучение, выбрать, жертва.*

**

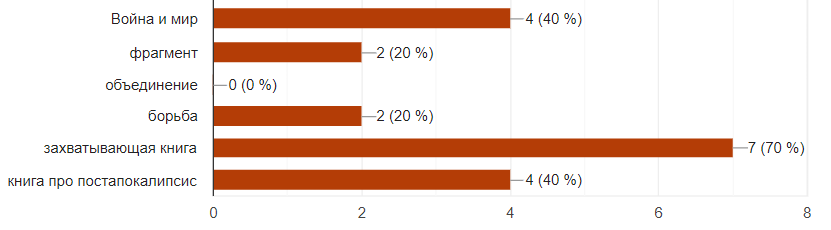
Пользователь 1085454: *разрушать, мама, класс, ребёнок, взрослый, возраст, вк, посмотреть, школа, напиться.*

**

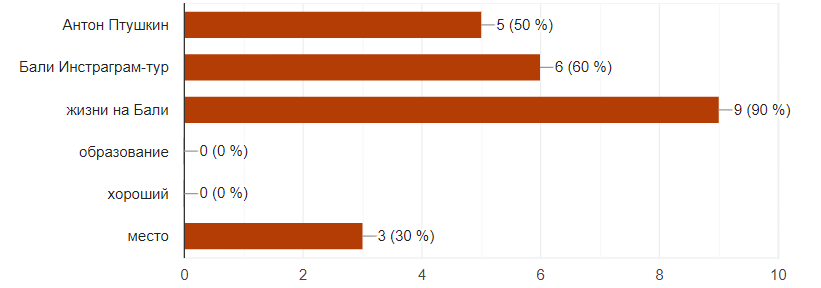
Пользователь 1888899: *гость, кафе, тайяки, дорогой, тайякии, хотеть, друг, японский, время, марат.*

**

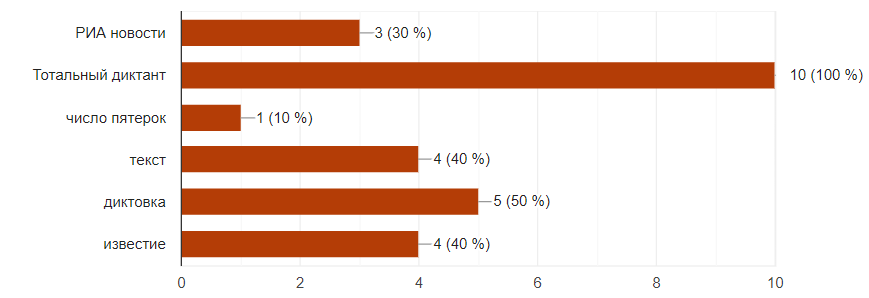
Пользователь 1955327: *человек, роман, год, книга, внешность, земля, хороший, мир, человечество, война.*



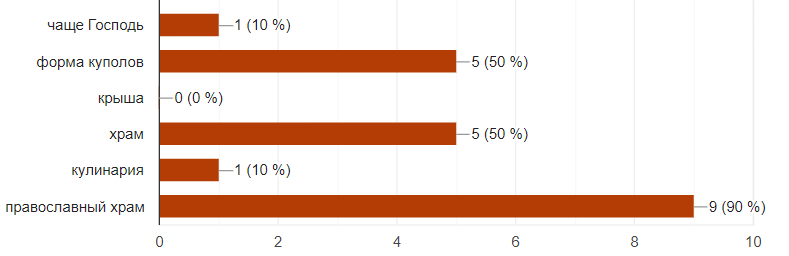
Пользователь 2394861: *выпуск, красивый, место, новый, бали, год, вид, антоха, дружище, инстаграм.*

**

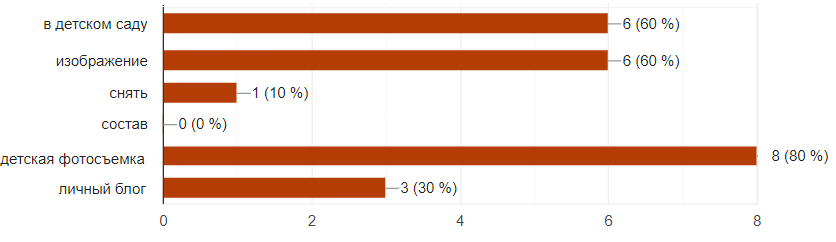
Пользователь 3854113: *год, диктант, тотальный, язык, акция, два, вид, новость, человек, жизнь.*



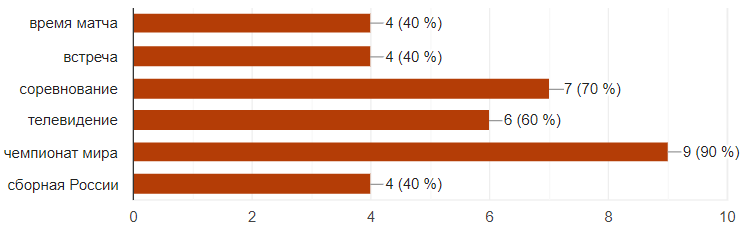
Пользователь 4817853: *купол, собор, храм, церковь, год, построить, рецепт, форма, икра, сказать.*

**

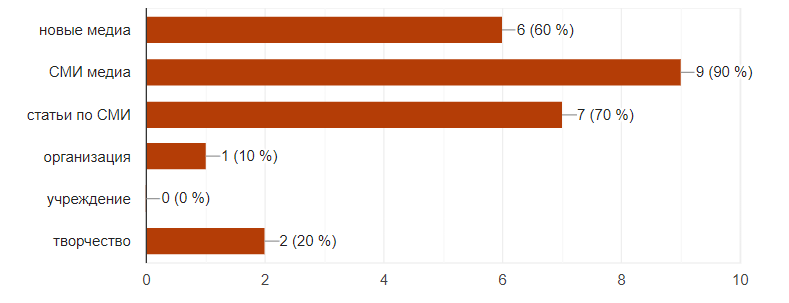
Пользователь 5019159: *съёмка, ребёнок, заранее, родитель, фото, воспитатель, одежда, снимка, коллаж, группа.*

**

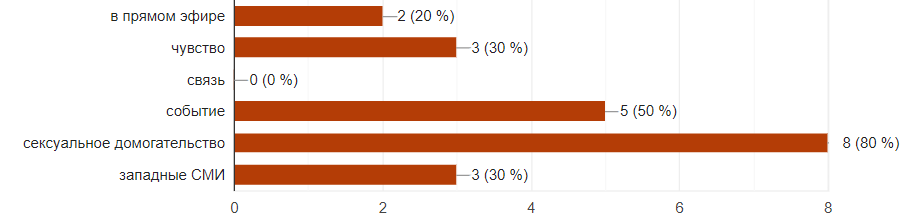
Пользователь 5068846: *матч, финляндия, сборная, чемпионат, журналист, мир, день, хоккей, тв, игра.*

**

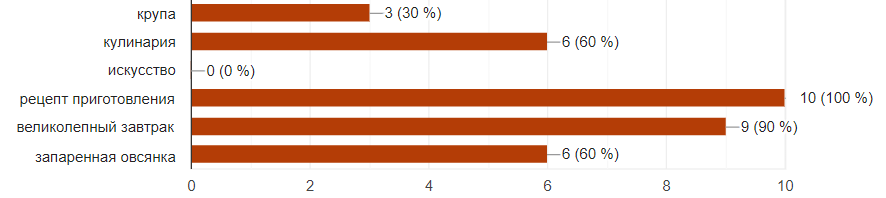
Пользователь 5424128: *медиа, журналистика, читатель, год, модель, сайт, баньян, кооператив, журналистский, журналист.*



Пользователь 6225105: *домогательство, сексуальный, твиттер, джульетта, пользователь, действие, болельщик, поцеловать, чемпионат, щека.*

**

Пользователь 41003889: *овсянка, банка, ложка, йогурт, молоко, крышка, стакан, закрывать, добавлять, мёд.*



1. <https://vk.com/> [↑](#footnote-ref-1)
2. <https://radimrehurek.com/gensim/> [↑](#footnote-ref-2)
3. <https://pymorphy2.readthedocs.io/en/latest/> [↑](#footnote-ref-3)
4. <https://gephi.org/> [↑](#footnote-ref-4)
5. <https://scikit-learn.org/stable/> [↑](#footnote-ref-5)
6. <https://matplotlib.org/> [↑](#footnote-ref-6)
7. <https://github.com/Wheatley961/Hidden_communities_Master_Degree> [↑](#footnote-ref-7)
8. <https://vk.com/kudrovolife?w=wall-51766355_2040391> [↑](#footnote-ref-8)
9. <https://vk.com/kudrovolife?w=wall-51766355_2039568> [↑](#footnote-ref-9)
10. <https://vk.com/kudrovolife?w=wall-51766355_2039568> [↑](#footnote-ref-10)
11. <https://vk.com/vandroukiru?w=wall-63731512_631979> [↑](#footnote-ref-11)
12. <https://www.alib.ru/forumarhtema.phtml?forumbeg=150&arhtema=apteka> [↑](#footnote-ref-12)
13. <https://vk.com/vandroukiru?w=wall-63731512_1369484> [↑](#footnote-ref-13)
14. <https://vk.com/vandroukiru?w=wall-63731512_1369750> [↑](#footnote-ref-14)
15. <https://www.facebook.com/> [↑](#footnote-ref-15)
16. [https://nlp.stanford.edu/software/tmt/tmt-0.4/](https://vk.com/away.php?to=https%3A%2F%2Fnlp.stanford.edu%2Fsoftware%2Ftmt%2Ftmt-0.4%2F&post=-193062269_86&cc_key=) [↑](#footnote-ref-16)
17. [http://mallet.cs.umass.edu/topics.php](https://vk.com/away.php?to=http%3A%2F%2Fmallet.cs.umass.edu%2Ftopics.php&post=-193062269_86&cc_key=) [↑](#footnote-ref-17)
18. [https://radimrehurek.com/gensim/](https://vk.com/away.php?to=https%3A%2F%2Fradimrehurek.com%2Fgensim%2F&post=-193062269_86&cc_key=) [↑](#footnote-ref-18)
19. <https://orange.biolab.si/> [↑](#footnote-ref-19)
20. [https://ru.wiktionary.org](https://ru.wiktionary.org/wiki/%D0%97%D0%B0%D0%B3%D0%BB%D0%B0%D0%B2%D0%BD%D0%B0%D1%8F_%D1%81%D1%82%D1%80%D0%B0%D0%BD%D0%B8%D1%86%D0%B0) [↑](#footnote-ref-20)
21. <https://linis.hse.ru/soft-linis/> [↑](#footnote-ref-21)
22. <https://www.python.org/downloads/release/python-370/> [↑](#footnote-ref-22)
23. <https://www.crummy.com/software/BeautifulSoup/bs4/doc/> [↑](#footnote-ref-23)
24. <https://vk.com/dev/methods> [↑](#footnote-ref-24)
25. <https://requests.readthedocs.io/en/master/> [↑](#footnote-ref-25)
26. <https://blog.hubspot.com/marketing/character-count-guide> [↑](#footnote-ref-26)
27. <https://pymorphy2.readthedocs.io/en/stable/> [↑](#footnote-ref-27)
28. <http://opencorpora.org/> [↑](#footnote-ref-28)
29. <https://ruscorpora.ru/new/> [↑](#footnote-ref-29)
30. <http://dict.ruslang.ru/freq.php> [↑](#footnote-ref-30)
31. <https://radimrehurek.com/gensim/> [↑](#footnote-ref-31)
32. <https://github.com/bmabey/pyLDAvis> [↑](#footnote-ref-32)
33. <https://ruwordnet.ru/ru> [↑](#footnote-ref-33)
34. <https://pymorphy2.readthedocs.io/en/latest/> [↑](#footnote-ref-34)
35. <https://www.google.com/intl/ru_ru/forms/about/> [↑](#footnote-ref-35)
36. <https://gephi.org/> [↑](#footnote-ref-36)
37. <https://vk.com/press/press-vk-12> [↑](#footnote-ref-37)
38. <http://www.yasiv.com/vk> [↑](#footnote-ref-38)
39. <https://scikit-learn.org/stable/> [↑](#footnote-ref-39)
40. <http://dict.ruslang.ru/freq.php> [↑](#footnote-ref-40)
41. <https://docs.google.com/forms/d/1VX8rBmJR2tf-22M7dRP_BvuZCzXvPgBORqy59k9RQM4> [↑](#footnote-ref-41)