Russian Technological Journal.
2025;13(1):38–48

ISSN 2782-3210 (Print)
ISSN 2500-316X (Online)

УДК 004.855.5:004.622
https://doi.org/10.32362/2500-316X-2025-13-1-38-48
EDN HJHQTR

НАУЧНАЯ СТАТЬЯ

Тематическое моделирование в потоке коротких сообщений на русском языке

Е.С. Мозаидзе @

Белгородский государственный технологический университет им. В.Г. Шухова, Белгород, 308012 Россия
@ *Автор для переписки, e-mail: mozaidze95@mail.ru*

Резюме

Цели. Работа посвящена тематическому моделированию коротких сообщений, поступающих посредством социальных сетей или другим способом в виде серии. Такая задача возникает в системах работы с населением в государственных и муниципальных структурах, в центрах опроса общественного мнения, а также в системах обслуживания клиентов и маркетинговых подразделениях. Цель работы – разработка и экспериментальная проверка набора алгоритмов тематической модели для автоматического определения основных тем обмена информацией и типичных сообщений, иллюстрирующих эти темы.

Методы. Используются методы переменных статистических распределений, примененных к статистике коллокаций, и подходы, характерные для решения задач тематического моделирования коротких текстов, но в применении к следующим друг за другом сообщениям. Таким образом, задачи онлайнового машинного обучения и тематического моделирования рассматриваются в совокупности.

Результаты. Рассмотрено построение тематической модели, в которой найденные кластеры с предъявлением их типичных представителей и текущего веса могут помочь человеку в принятии решений в соответствии с тематикой этих наиболее важных сообщений. Предложенный метод был экспериментально протестирован на корпусе реальных сообщений. Результаты тематического моделирования (построенные тематические модели) согласуются с результатами, полученными вручную: выбранные сообщения, иллюстрирующие проблемные темы с наибольшим весом, являются таковыми и с точки зрения экспертов.

Выводы. Предлагаемый алгоритм тематического моделирования позволяет автоматически выявлять наиболее важные темы в текущем общении, показывает посты, служащие индикаторами этих тем, что позволяет существенно упростить решение задачи.

Ключевые слова: тематическое моделирование, EM-алгоритм, скрытое размещение, метод поточной перенормировки

• **Поступила:** 25.03.2024 • **Доработана:** 30.09.2024 • **Принята к опубликованию:** 17.11.2024

Для цитирования: Мозаидзе Е.С. Тематическое моделирование в потоке коротких сообщений на русском языке. *Russian Technological Journal.* 2025;13(1):38–48. https://doi.org/10.32362/2500-316X-2025-13-1-38-48, https://elibrary.ru/HJHQTR

Прозрачность финансовой деятельности: Автор не имеет финансовой заинтересованности в представленных материалах или методах.

Автор заявляет об отсутствии конфликта интересов.

RESEARCH ARTICLE

Topic modeling in the stream of short messages in Russian

Elena S. Mozaidze @

V.G. Shukhov Belgorod State Technological University, Belgorod, 308012 Russia
@ *Corresponding author, e-mail: mozaidze95@mail.ru*

Abstract

Objectives. This work is devoted to the topic modeling of short messages received through social networks or in another way in the form of a series of short messages. This need arises in public relations systems in state and municipal structures, in public opinion polling centers, as well as in customer service systems and marketing departments. The aim of the work is to develop and experimentally test a set of algorithms for a thematic model for automatically determining the main topics of information exchange and typical messages illustrating these topics.

Methods. The work uses methods of variable statistical distributions applied to collocation statistics and approaches typical for resolving problems of topic modeling of short texts, but applied to successive messages. In this way, online machine learning and topic modeling are considered jointly.

Results. The work considered the construction of a thematic model in which clusters found with the presentation of their typical representatives and current weight can help decision-making in accordance with the subject of these most important messages. The proposed method was experimentally tested on a corpus of real messages. The results of topic modeling (the constructed thematic models) are consistent with the results obtained manually. The messages selected illustrate that the topics with the highest weight are seen as such from the point of view of human experts.

Conclusions. The proposed algorithm of topic modeling allows the most important topics in current communication to be automatically identified. It shows posts that serve as indicators of these topics, and thereby significantly simplifies the solution of the problem.

Keywords: topic modeling, EM-algorithm, hidden placement, streaming renormalization method

• **Submitted:** 25.03.2024 • **Revised:** 30.09.2024 • **Accepted:** 17.11.2024

For citation: Mozaidze E.S. Topic modeling in the stream of short messages in Russian. *Russian Technological Journal.* 2025;13(1):38–48. https://doi.org/10.32362/2500-316X-2025-13-1-38-48, https://elibrary.ru/HJHQTR

Financial disclosure: The author has no financial or proprietary interest in any material or method mentioned.

The author declares no conflicts of interest.

ВВЕДЕНИЕ

При работе с социальными сетями и мессенджерами почти всегда возникает задача автоматизированного поиска наиболее важной темы в обмене сообщениями. Это связано с многими причинами, среди которых необходимость модерации чата, выявление моментов, когда требуется вмешаться ответственному лицу, поиск наиболее важных на текущий момент тем общения людей в контексте тематики чата.

Исследуемый в работе случай относится к информационному обмену в социальных сетях города Белгород по той причине, что у авторов имеется возможность получить эти данные, однако, предлагаемая методика применима к любому предмету исследований подобного рода, для которого имеется достаточное количество данных.

Тематическое моделирование (topic modeling) – это способ научить машину (компьютер) выделять в текстах содержательные темы. Например, проанализировав массив новостных и публицистических текстов, можно выделить определенные темы. Конечно, компьютеры не могут понять смысл статей буквально, но если есть большая коллекция текстов с разными темами, то вероятности совместного употребления слов позволяют выделить отдельные тематические пласты.

Тематический пласт, «отфильтрованный» из множества текстов – это просто набор слов, характерных для темы. Слова в таком наборе отсортированы

по важности для данной темы [1–3]. В терминах кластерного анализа тема – это результат бикластеризации, т.е. одновременной кластеризации и слов, и документов по их семантической близости.

В 1998 г. одними из первых интерес к теме вероятностной тематической модели проявили ученые К. Пападимитриу, Х. Томаки, С. Вемпала и П. Рагаван [4]. Их работа была посвящена скрытому семантическому индексированию (latent semantic indexing, LSI) – методу поиска информации, основанному на спектральном анализе базы документов.

Дальнейшее развитие этой темы отражено в работах зарубежных ученых.

Томас Хофман [5] изучал вероятностное скрытое семантическое индексирование. В отличие от стандартного скрытого семантического индексирования с помощью разложения по сингулярным значениям, вероятностный вариант имеет прочную статистическую основу и определяет надлежащую генеративную модель данных. Поисковые эксперименты на ряде тестовых коллекций показывают существенный прирост производительности по сравнению с методами прямого сопоставления терминов, а также с LSI.

Дэвид Блей [6–8] рассматривал контролируемое скрытое распределение Дирихле (spatial latent Dirichlet allocation, sLDA) или статистическую модель помеченных документов. В своих работах он иллюстрирует преимущества sLDA по сравнению с современной упорядоченной регрессией, а также по сравнению с неконтролируемым анализом (latent Dirichlet allocation, LDA), за которым следует отдельная регрессия.

Американский ученый в области информатики, доцент Стэндфордского университета, исследователь робототехники и машинного обучения, один из основателей платформы онлайн-обучения «Coursera»[1] Эндрю Ын давно предсказал[2] [3], что распознавание естественного языка станет основным способом взаимодействия человека с компьютером. В своей работе он обратил внимание на обучение с подкреплением, как на один из способов машинного обучения.

Также свой вклад в развитие этой темы внесли и российские ученые.

Воронцов К.В. [9] предложил в своей работе аддитивную регуляризацию тематических моделей (additive regularization of topic models, ARTM), которая основана на максимизации взвешенной суммы логарифма правдоподобия и дополнительных критериев – регуляризаторов. Это упрощает комбинирование тематических моделей и построение сколь угодно сложных многоцелевых моделей.

Потапенко А.А. [10] рассмотрен обобщенный EM-алгоритм[3] с эвристиками сглаживания, сэмплирования и разреживания, позволяющий при различных сочетаниях этих эвристик получать как известные тематические модели, так и новые.

Лукашевич Н.В. [11], Нокель М.А.[4] представили результаты экспериментов по добавлению биграмм в тематические модели и учету сходства между ними и униграммами. Они предложили новый алгоритм PLSA-SIM, являющийся модификацией алгоритма построения тематических моделей PLSA (probabilistic latent semantic analysis).

В статье Коршунова А.В., Гомзина А.Г. представлен сравнительный обзор различных моделей, описаны способы оценивания их параметров и качества результатов, а также приведены примеры открытых программных реализаций [12].

Разработаны программные библиотеки для тематического моделирования, такие как *Mallet*[5], *Gensim*[6] и *BigArtm*[7], позволяющие создавать вероятностное тематическое моделирование.

Несколько лет назад началось активное использование инструментов больших языковых моделей (large language model, LLM), в т.ч. для решения задач тематического моделирования. Появилось довольно большое число работ в этом направлении, из которых, применительно к целям настоящей работы, можно указать следующие. В работе [13] авторы исследуют ключевые события в новостных лентах. Рассматривается проблема их идентификации и связей. Исследование построено на использовании LLM для поиска и резюмирования, а собственно тематическое моделирование делается выбором топовой темы алгоритмом скользящего

[1] https://www.coursera.org. Дата обращения 02.12.2024. / Accessed December 02, 2024.

[2] Ng A.Y. *Shaping and Policy Search in Reinforcement Learning*. Ph.D. Thesis, UC Berkley, 2003.

[3] Expectation–maximization – алгоритм, используемый в математической статистике для нахождения оценок максимального правдоподобия параметров вероятностных моделей, в случае, когда модель зависит от некоторых скрытых переменных. [An expectation–maximization (EM) algorithm is an iterative method used in mathematical statistics to find maximum likelihood estimates of the parameters of probabilistic models when the model depends on some hidden variables.]

[4] Нокель М.А. *Методы улучшения вероятностных тематических моделей текстовых коллекций на основе лексико-терминологической информации*: автореферат дис. … канд. ф.-м. наук. М., 2015. 20 с. [Nokel M.A. *Methods for improving probabilistic topic models of the text collections based on lexicoterminological information*: Cand. Sci. Thesis (phys.-math.). Moscow, 2015. 20 p. (in Russ.).]

[5] http://mallet.cs.umass.edu/topics.php. Дата обращения 02.12.2024. / Accessed December 02, 2024.

[6] https://radimrehurek.com/gensim. Дата обращения 02.12.2024. / Accessed December 02, 2024.

[7] http://bigartm.org. Дата обращения 02.12.2024. / Accessed December 02, 2024.

окна. Несмотря на хорошие результаты в заявленной области, динамика распределения тем в работе явно не учитывается и, кроме того, не происходит дообучение модели. В статье [14] рассматривается интерпретация результатов тематических моделей с использованием большой языковой модели ChatGPT[8]. Любопытно, что результат работы показал, что данная LLM почти в половине случаев расходится в интерпретации с человеком. Целью работы было просто продемонстрировать способность ChatGPT описывать темы и предоставлять полезную информацию, но в самом тематическом моделировании LLM не помогала[9, 10]. Работы посвящены, соответственно, тематическому моделированию суммаризованных текстов и оценке контекстуализированной тематической согласованности. В первой работе показано, что LLM успешно транслирует смысл в суммаризованные тексты, но тематическое моделирование на них получается не всегда верным – качество зависит от контекста. Предложенная авторами второй работы автоматическая оценка согласованности тем хорошо работает с короткими документами и не подвержена влиянию бессмысленных, но высоко оцененных тем. Этот результат, безусловно, заслуживает рассмотрения и использования для дальнейших исследований.

Постановка задачи и концепция исследования текстов, поступающих в адрес государственных структур, для распределения их по структурным единицам управления, соответствующим тематикам текстов, были целью работы [1], результаты которой используются в настоящем исследовании.

Целями данной работы являются разработка и экспериментальная проверка набора алгоритмов тематического моделирования, который приведет к построению набора кластеров с предъявлением их типичных представителей и текущего веса, при соблюдении обычных для тематического моделирования условий нормировки и распределения.

1. МАТЕРИАЛЫ И МЕТОДЫ

1.1. Постановка задачи

Пусть имеется постоянно действующая система, принимающая сообщения короткой и средней длины (ремарки, обращения). Это может быть персональная страница в социальной сети, веб-сервис по приему обращений, электронный почтовый ящик с автоматизированной выгрузкой текстов, система управления отношениями с клиентами и т.д.

Чаще всего для собранной таким образом информации возникает одна из двух задач: 1) распределить сообщения по заранее заданным группам (классам), или 2) сгруппировать сообщения в заранее неопределенные группы (кластеры) с близкой семантикой. Будем рассматривать вторую задачу на потоке сообщений, а именно: каждому вновь пришедшему сообщению сопоставим вектор, координаты которого представляют собой вероятности принадлежности данного сообщения к сформированным к этому моменту кластерам.

Указанная задача является задачей тематического моделирования или, иначе говоря, мягкой бикластеризации. В данной формулировке задача усложняется тем, что множество сообщений не является ограниченным, и необходимо либо каждый раз определять число кластеров, либо зафиксировать его и расформировывать лишние кластеры.

В случае удовлетворительного решения описанная задача может применяться в системах работы с населением для государственных и муниципальных структур, в центрах опроса общественного мнения, а также в системах CRM (customer relationship management – управление взаимоотношениями с клиентами) и маркетинговых службах корпораций.

1.2. Метод решения

Методы тематического моделирования основаны, как правило, на вычислениях частотностей слов в документах, а также слов и документов в темах. Чаще всего для тематического моделирования используются EM-алгоритм[11, 12] или скрытое размещение Дирихле[13]. Отличительными чертами и того, и другого является необходимость взвешивания всех слов в сообщениях без принятия во внимание смысловых связей. Смыслы восстанавливаются по известным смыслам в больших массивах текста. По сути, тематическое моделирование только осуществляет бенчмаркинг смысла.

В поставленной задаче работа со смыслом описанным способом (похожим на бенчмаркинг) не представляется возможной, т.к. сообщения, как правило, короткие и часто содержат грамматические ошибки. В таком случае крайне тяжело

[8] https://chat-gpt.org/. Дата обращения 02.12.2024. / Accessed December 02, 2024.

[9] https://arxiv.org/abs/2403.15112. Дата обращения 02.12.2024. / Accessed December 02, 2024.

[10] https://arxiv.org/abs/2305.14587. Дата обращения 02.12.2024. / Accessed December 02, 2024.

[11] https://pythobyte.com/python-for-nlp-topic-modeling-8fb3d689/?ysclid=lgdpql4ef3963911399 (in Russ.). Дата обращения 02.12.2024. / Accessed December 02, 2024.

[12] https://mathprofi.com/userinfo/14285/ (in Russ.). Дата обращения 02.12.2024. / Accessed December 02, 2024.

[13] https://digitrain.ru/articles/252142/ (in Russ.). Дата обращения 02.12.2024. / Accessed December 02, 2024.

сориентировать их на какую-то тему бенчмаркингом – для этого требуется больший блок текста.

В работе [1] предложена методика работы с сообщениями на основе пар «существительное – глагол», набор которых ставится в соответствие каждому сообщению, имеющему хоть одну такую пару. В данной работе будем считать, что такое отображение уже произведено и каждое сообщение имеет идентификатор и набор соответствующих ему пар «существительное – глагол».

Далее предполагается, что:

- сообщения следуют друг за другом и каждое имеет идентификатор;
- если в сообщении нет глаголов или существительных, оно считается несущественным, не принимается в расчет, но идентификатор ему присваивается;
- для скользящего окна сообщений имеется журнал пар «существительное – глагол» с соответствующим идентификатором сообщения;
- темы идентифицируются существительными, входящими в сообщения;
- каждая пара «существительное – глагол» имеет вероятность (число от 0 до 1) вхождения в тему таким образом, что сумма вероятностей для пары по всем темам равна 1;
- каждое сообщение имеет вероятность вхождения в тему, и сумма вероятностей сообщения по всем темам равна 1;
- максимальное количество тем фиксировано, удаление тем производится на основе показателя веса темы, рассчитываемого как произведение средней вероятности пары в теме на число сообщений, имеющих отношение к теме.

Такие предположения позволяют воспользоваться методом поточной перенормировки – регулярного отбрасывания элементов с малым весом.

Далее рассмотрим поочередно шаги процесса анализа сообщений в движущемся окне.

1.2.1. Препроцессинг

На вход анализатора поступают сообщения, представляющие собой текст на русском языке из одного или нескольких предложений. Для подготовки к моделированию над текстом сообщения производятся следующие операции:

- токенизация предложений (разделение на предложения) средствами *nltk*[14];
- токенизация слов в предложении средствами *nltk*;
- определение частей речи токенов средствами *nltk*;
- лемматизация существительных и глаголов средствами *pymorphy2*[15];
- составление пар «существительное – глагол» по правилу «ближайшие соседи».

Все это реализовано в программном модуле *appeals_processing.ipynb*[16] и может работать в реальном времени в потоке сообщений.

1.2.2. Тематическое моделирование

На выходе препроцессора *appeals_processing.ipynb* получаем множества лемматизированных существительных (nouns) и глаголов (verbs), содержащихся в окне сообщений, длиной *A* строк, а также .csv или .json файл (в настоящее время реализовано получение .csv, но перенастройка в .json не представляет труда), в каждой строке которого представлен список из трех элементов:

- строка вида «[(сущ1, глаг1), …, (сущN, глагN)]»;
- номер файла;
- номер сообщения в файле.

Таким образом, набор пар «существительное – глагол» с одной стороны может быть вложен в числовое множество, а с другой стороны – однозначно связан с сообщением, по которому он построен.

Далее для каждой пары в каждом сообщении реализуется следующий алгоритм, шаги которого рассматриваются подробнее в разделе «Результаты»:

1. Если существительное содержится в темах, то пара включается в эти темы, а также во вновь созданные для данного сообщения темы с равными вероятностями.
2. Если существительное не содержится в темах, но глагол содержится в темах, то пара включается в эти темы с вероятностями, уменьшенными множителем μ (задаваемый постоянный параметр), а также создается новая тема по существительному, и пара включается в нее и во вновь созданные для данного сообщения темы с равными вероятностями.
3. Если ни существительное, ни глагол не содержатся в темах, то формируется новая тема по существительному, и пара включается в нее, а также во вновь созданные для данного сообщения темы с равными вероятностями.

После того, как все пары всех сообщений обработаны этим алгоритмом, создается .json файл тематической модели: каждому ключу темы (существительному) ставится в соответствие .json объект с ключами – парами «существительное – глагол», а значениями – вероятностями вхождения пары в тему.

[14] https://www.nltk.org/. Дата обращения 02.12.2024. / Accessed December 02, 2024.

[15] https://github.com/pymorphy2/pymorphy2. Дата обращения 02.12.2024. / Accessed December 02, 2024.

[16] https://disk.yandex.ru/d/8LPWy3ZP-7V30Q (in Russ.). Дата обращения 02.12.2024. / Accessed December 02, 2024.

Итогом тематического моделирования является свернутый по значениям вероятностей пар .json объект тематической модели: в нем каждому ключу темы соответствует числовое значение веса этой темы. Отсортированный по убыванию значений этот объект даст искомые кластеры с их весами.

1.3. Материалы

Данные для настоящего исследования были подготовлены из входящих сообщений, пришедших в администрацию города Белгород, путем удаления собственных имен и других признаков, демаскирующих персональные данные. Обработанный таким образом датасет представлял собой .csv файл с заголовком вида «Дата; Время; Вопрос», любое сообщение можно восстановить по дате и времени получения. Вопросы и являлись теми самыми сообщениями (обращениями), которые были подвергнуты обработке со следующими целями: построить облако тегов (которое в данном случае являлось облаком тем) и выявить наиболее важные вопросы.

В датасет входило 3621 сообщение, часть из которых повторялись (люди копировали вопрос и отправляли снова), а некоторые не содержали пар «существительное – глагол» или были написаны с нарушением орфографии так, что слова не находились в базовых словарях.

Просмотр датасета был организован скользящим окном по 300 сообщений. Это средний объем сообщений за месяц. Таким образом, система анализировала контент обращений за прошедший месяц и выдавала темы и наиболее важные вопросы в сообщениях за этот период. Так как окно скользящее, то результат работы мог меняться каждый день и даже с каждым новым сообщением.

В настоящее время специалисты администрации определяют приоритет рассмотрения обращений вручную. Предлагаемая программная система сможет делать это автоматически. В работе предлагается метрика для измерения качества сортировки обращений машиной в сравнении с сортировкой людьми.

2. РЕЗУЛЬТАТЫ

2.1. Вычисление вероятностей в тематической модели

На входе в алгоритм тематического моделирования имеются:

- .csv-файл или другой источник обращений, содержащий дату и время получения обращения и его текст;
- словари существительных и глаголов русского языка;
- построенный к этому моменту (по сообщению номер m) словарь T_m тематической модели: по ключу-существительному находится словарь с ключом – парой «существительное – глагол» и значением в виде вероятности принадлежности пары к теме; в начале работы этого словаря еще нет, он инициируется по первому существенному (содержащему существительное и глагол) обращению.

В словаре тематической модели сумма значений вероятностей (далее – просто значений) по всем темам для данной пары должна равняться 1. То есть,

$$T_m = \{t : \{s : p_{st}\}\}, \sum_t p_{st} = 1 \ \forall s \in S_m,$$

где t – существительное (тема), s – пары «существительное – глагол», построенные из набора S_m, содержащего m обращений, p_{st} – вероятность того, что пара s принадлежит теме t.

Выход алгоритма представляет собой новый словарь T_{m+1} тематической модели, удовлетворяющий тому же требованию.

Возможные случаи для каждой пары s из обращения $m + 1$:

1. Если в T_m уже имеется пара s, то все ее значения заменяются на величины $\hat{p}(1 - p_0 n_0)$, где $\hat{p}$ – текущее значение, а p_0, n_0 – значение для пары s в новых темах и их количество, соответственно;
2. Если в T_m нет пары s, но имеется тема, совпадающая с существительным из пары s, то s вписывается в словарь этой темы со значением, равным p_1; кроме того, пара s записывается во все вновь созданные темы с вероятностями p_4;
3. Если в T_m нет темы, совпадающей с существительным из s, но в каких-то темах есть пара с глаголом из s, то s вписывается в словари этих тем со значением, равным p_2; кроме того, пара s записывается во все вновь созданные темы с вероятностями p_5;
4. Если ни существительное, ни глагол из пары s не присутствуют в T_m, то в T_{m+1} создается новая тема (по существительному из s), и s записывается в нее с вероятностью p_3; во все остальные вновь созданные темы s записывается с вероятностью p_6.

Таким образом, имеем непересекающиеся возможности для пары s и в итоге должны выполняться следующие соотношения:

Случай 1. Сумма по всем темам, содержащим s, ее новых значений, должна равняться 1:

$$p_0 n_0 + \sum_s \hat{p}_s (1 - p_0 n_0) = 1,$$

что выполняется тождественно. Но поскольку p_0 имеет смысл вероятности, то

$$p_0 \le \frac{1}{n_0},$$

Случай 2. Сумма по всем старым и новым темам равна 1:

$$p_1 n_{\text{noun}} + n_0 p_4 = 1,$$

где n_{noun} – количество имеющихся (старых) тем с существительным из пары s. Отсюда получаем

$$p_1 = \frac{1 - n_0 p_4}{n_{\text{noun}}}, \quad p_4 \le \frac{1}{n_0}.$$

Случай 3. Сумма по всем старым и новым темам равна 1:

$$p_2 n_{\text{verb}} + n_0 p_5 = 1,$$

где n_{verb} – количество старых тем с глаголом из пары s. Отсюда получаем

$$p_2 = \frac{1 - n_0 p_5}{n_{\text{verb}}}, \quad p_5 \le \frac{1}{n_0}.$$

Случай 4. Сумма по всем новым темам равна 1:

$$p_3 n_{\text{new}} + (n_0 - n_{\text{new}}) p_6 = 1,$$

где n_{new} – количество новых тем с существительным из пары s. Отсюда получаем

$$p_3 = \frac{1 - (n_0 - n_{\text{new}}) p_6}{n_{\text{new}}}, \; p_6 \le \frac{1}{n_0 - n_{\text{new}}}.$$

Так как все вероятности p_4, p_5, p_6 соответствуют включению пары в тему только по признаку присутствия в одном обращении, будем считать, что они равны:

$$p_4 = p_5 = p_6 \equiv q \le \frac{1}{n_0}.$$

Вероятности p_1 и p_5 будем считать равными, т.к. они соответствуют включению в тему по существительному.

Вероятность включения в тему по существительному свяжем с вероятностью включения в тему по глаголу:

$$p_2 = kp_1.$$

Обозначим $p_1 = p_3 = p$, и теперь единичные суммы перепишутся в виде:

$$pn_{\text{noun}} + n_0 q = 1,$$

$$kpn_{\text{verb}} + n_0 q = 1,$$

$$pn_{\text{new}} + (n_0 - n_{\text{new}}) q = 1.$$

Решим эту систему уравнений и определим p, q, k:

$$p = \frac{n_{\text{new}}}{n_{\text{new}} n_0 - n_{\text{noun}} n_0 + n_{\text{noun}} n_{\text{new}}},$$

$$q = \frac{n_{\text{new}} - n_{\text{noun}}}{n_{\text{new}} n_0 - n_{\text{noun}} n_0 + n_{\text{noun}} n_{\text{new}}},$$

$$k = \frac{n_{\text{noun}}}{n_{\text{verb}}},$$

если все n_{verb} и $n_{\text{new}} \cdot n_0 - n_{\text{noun}} \cdot n_0 + n_{\text{noun}} \cdot n_{\text{new}}$ отличны от 0.

Рассмотрим теперь случаи, когда одно или оба эти значения равны 0.

Если $n_{\text{verb}} = 0$, то в сообщении нет пар, соответствующих Случаю 3, а значит, имеется только два уравнения и неизвестные – только p и q. Решения для них уже найдены выше.

Если

$$n_{\text{new}} \cdot n_0 - n_{\text{noun}} \cdot n_0 + n_{\text{noun}} \cdot n_{\text{new}} = 0,$$

то из уравнений следует $n_{\text{new}} = n_{\text{noun}} = 0$, т.е. этот случай соответствует описанному Случаю 1, когда пара s уже содержится в T_m, а величина p_0 может принимать любое значение.

Так как в новых темах в этом случае точно нет существительного из s, то вероятность пары в них должна быть меньше или равна вероятности любой пары из данного обращения в старом словаре:

$$\hat{p}_s (1 - p_0 n_0) \le \min_s \hat{p}_s \Big|_{\hat{p}_s \ne 0}.$$

В итоге, для нахождения нового словаря T_{m+1} для каждой пары из обращения требуется найти числа

$$n_{\text{noun}}, n_0, n_{\text{new}}, n_{\text{verb}}, p_0,$$

причем n_0 одинаково для всех пар s из данного обращения.

При построении нового словаря T_{m+1} следует использовать значения вероятностей

$$p_0 = \frac{1}{n_0}\left(1 - \frac{1}{\hat{p}} \min_l \hat{p}_l \Big|_{\hat{p}_l \ne 0}\right),$$

$$p_1 = p_3 = \frac{n_{\text{new}}}{n_{\text{new}} n_0 - n_{\text{noun}} n_0 + n_{\text{noun}} n_{\text{new}}},$$

$$p_2 = \frac{1}{n_{\text{verb}}} \cdot \frac{n_{\text{new}} n_s}{n_{\text{new}} n_0 - n_{\text{noun}} n_0 + n_{\text{noun}} n_{\text{new}}},$$

$$p_4 = p_5 = p_6 = \frac{n_{\text{new}} - n_{\text{noun}}}{n_{\text{new}} n_0 - n_{\text{noun}} n_0 + n_{\text{noun}} n_{\text{new}}}.$$

2.2. Алгоритм тематического моделирования обращений, представленных наборами пар «существительное – глагол»

Вход: словарь T_m тематической модели (может быть пустым), содержащий не более w различных пар «существительное – глагол» (w – размер «окна»); обращение, представленное в виде списка упорядоченных пар «существительное – глагол» (существительное на первом месте, токены лемматизированы).

Выход: словарь T_{m+1}, содержащий не более w различных пар, в котором все пары из обращения $m + 1$ получили вероятности принадлежности к темам, выраженным существительными.

Процедура:

1. Устанавливаем $n_0 = n_{\text{new}} = 0$.
2. Для всех пар s в обращении:
3. Просматриваем словарь T_m и находим:
   а) значения $\hat{p}_s$, которые s имеет в T_m, и соответствующие темы $t_{sp}^1, ..., t_{sp}^n$;
   б) темы $t_s^1, ..., t_s^{n_{\text{noun}}}$, совпадающие с существительным из s, но не содержащие s целиком;
   в) темы $t_{\text{verb}}^1, ..., t_{\text{verb}}^{n_{\text{verb}}}$, не совпадающие с указанными в пп. а) и б), в которых имеется хотя бы один глагол из s;
4. Если шаг 3 не дал ни одного совпадения, увеличиваем n_0 на 1; если, кроме того, среди уже созданных новых тем нет существительного из s (или пока нет новых тем), то увеличиваем n_{new} на 1 и создаем новую тему $t_{\text{new}}^{n_{\text{new}}}$;
5. По завершению просмотра словаря T_m создаем из него копированием словарь T_{m+1} и далее работаем с ним.
6. Находим

$$p_0 = \frac{1}{n_0}\left(1 - \frac{1}{\hat{p}} \min_l \hat{p}_l \Big|_{\hat{p}_l \neq 0}\right);$$

7. В темы $t_{sp}^1, ..., t_{sp}^n$ вписываем с вероятностями p_0 пары s, найденные на шаге 3а).
8. В темы $t_s^1, ..., t_s^{n_{\text{noun}}}$ вписываем с вероятностями $p = \frac{n_{\text{new}}}{n_{\text{new}} n_0 - n_{\text{noun}} n_0 + n_{\text{noun}} n_{\text{new}}}$ пары s, найденные на шаге 3б).
9. В темы $t_{\text{verb}}^1, ..., t_{\text{verb}}^{n_{\text{verb}}}$ вписываем с вероятностями $p_2 = \frac{1}{n_{\text{verb}}} p$ пары s, найденные на шаге 3в).
10. В темы $t_{\text{new}}^1, ..., t_{\text{new}}^{n_{\text{new}}}$ вписываем:
    а) с вероятностями p пары, существительные которых совпадают с темой;
    б) с вероятностями
    $$q = \frac{n_{\text{new}} - n_{\text{noun}}}{n_{\text{new}} n_0 - n_{\text{noun}} n_0 + n_{\text{noun}} n_{\text{new}}}$$
    все остальные пары обращения.
11. Если число пар W в словаре больше, чем w, то находим тему с наименьшим весом и удаляем из нее все или $(W - w)$ пар (что меньше); делаем это, пока все $(W - w)$ лишние пары не будут удалены из словаря.
12. Вычисляем веса оставшихся тем и проводим их сортировку.

Следуя этому алгоритму, будем получать в каждой итерации наиболее весомые темы, отсортированные по весу.

2.3. Результаты тематического моделирования

Для тестирования алгоритма был выбран упомянутый ранее датасет, из которого были удалены повторяющиеся сообщения и сообщения, не содержащие существительных вместе с глаголами. Рабочий датасет содержал около 2700 сообщений, направленных в адрес мэра Белгорода и городских департаментов за год. Данные анонимизированы. Размер окна (в парах, одновременно находящихся в словаре) принят равным $w = 300$.

На рисунке представлен график изменения веса топовых тем словаря на протяжении всего года.

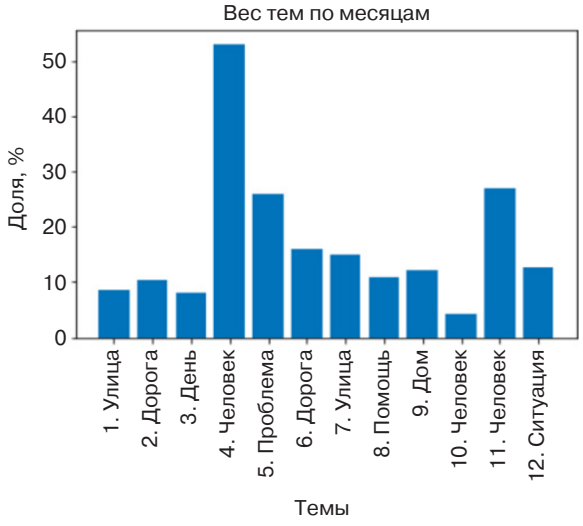

Рисунок. Темы, наиболее популярные в обращениях в мэрию г. Белгорода

В целом, выбранные моделью темы соответствуют наиболее важным вопросам, волнующим горожан в определенный период времени. Непосредственно по темам этого не видно, т.к. тема в контексте настоящей работы – это одно слово (существительное), по которому трудно оценить важность сообщений. Но тема является ключом к множеству сообщений и, если это множество таково, и изменяется так, что вес темы увеличивается, то сообщения в множестве заслуживают внимания. По сгруппированным вокруг тем-лидеров сообщениям администрация работала и далее в топ вышли другие темы. Это означает, что люди (администрация) и машина (предлагаемая система) верно отреагировали на настроения в городе, иначе нашлись бы темы, которые оставались в топе продолжительное время. По оценке специалистов администрации города, тематическое моделирование проведено верно.

Для количественной оценки качества тематического моделирования можно использовать метрики, например, BCubed-метрику[17] [15], применение которой оправдано по построению алгоритма. Если обозначить через $p_{ts.mon}$ вероятность того, что сообщение s, относящееся к месяцу mon, принадлежит теме $t\left(\sum_t p_{ts.mon}=1\right)$, то BC-точность (BC precision, BCP) и BC-полнота (BC recall, BCR) для темы t в месяце mon будут определяться, соответственно, формулами:

$$BCP(t,mon)=\frac{\sum_s p_{ts.mon}}{\sum_{s,k} p_{tsk}},$$

$$BCR(t,mon)=\frac{\sum_s p_{ts.mon}}{\sum_{r,k} p_{rsk}}.$$

Для датасета из эксперимента первый показатель топовых тем всегда превышал 55%, а второй варьировался от 27% до 83% (соответственно весу тем, показанному на рисунке). С учетом того, что число тем в месяце никогда не было меньше 70, качество кластеризации можно признать высоким и по метрике: случайный выбор приведет к показателю около 1.5%.

ЗАКЛЮЧЕНИЕ

В работе предложены метод и алгоритм тематического моделирования для коротких сообщений. Кластеризация коротких сообщений является сложной задачей, поскольку такие сообщения очень трудно сопоставить какому-либо контексту, т.е., большие корпуса текста, на которых можно обучить модель, не дают, как правило, удовлетворительных описаний тем.

Работа основана на предложенной ранее методике выделения контура смысла сообщения в текстах на русском языке на основе пар «существительное – глагол».

По предложенной методике построена тематическая модель и проведено моделирование на реальных данных. Построенная кластеризация показала относительно высокое качество по метрике BCubed. В то же время результат виден и при качественной оценке: если тема вычисляется как топовая в конкретном периоде, то вопросы, поднятые в ней, заслуживают первоочередной оценки и принятия мер. На примере датасета, используемого в эксперименте, такая корреляция (между предлагаемым набором сообщения для первоочередного реагирования и выбранными людьми сообщениями) была больше 70%. Однако даже без количественной оценки сотрудники организации – владельца датасета высказали мнение о готовности использовать программное обеспечение, построенное на предлагаемом алгоритме, как систему поддержки принятия решений (точнее, советующую систему для первоочередного реагирования на сообщения). По их мнению, это намного сократит время обработки сообщений.

БЛАГОДАРНОСТИ

Работа выполнена в рамках реализации федеральной программы поддержки университетов «Приоритет 2030» с использованием оборудования на базе Центра высоких технологий Белгородского государственного технологического университета им. В.Г. Шухова.

ACKNOWLEDGMENTS

This work was realized in the framework of the Priority 2030 Program using the equipment of High Technology Center at the V.G. Shukhov Belgorod State Technological University.

[17] Семейство метрик BCubed реализовано в библиотеке https://pypi.org/project/bcubed-metrics/. Дата обращения 02.12.2024. [The BCubed family of metrics is implemented in the library https://pypi.org/project/bcubed-metrics/. Accessed December 02, 2024.]

СПИСОК ЛИТЕРАТУРЫ

1. Брусенцев А.Г., Зуева Е.С. Тематические модели и инструменты обработки естественного языка в применении к задачам муниципальных структур. В сб.: *Актуальные теоретические и прикладные вопросы управления социально-экономическими системами: материалы II Международной научно-практической конференции.* М.: Институт развития дополнительного профессионального образования; 2020. Т. 2. С. 262–269. https://elibrary.ru/fkgyxn
2. Зуева Е.С. Вероятностная классификация входящих обращений на основе алгоритма управляемых рекуррентных нейронов. В сб.: *Материалы Международной научно-технической конференции молодых ученых БГТУ им. В.Г. Шухова.* Белгород: БГТУ им. В.Г. Шухова; 2021. С. 3564–3575. https://www.elibrary.ru/nhlzpv
3. Поляков В.М., Мозаидзе Е.С. Алгоритм коллаборативной фильтрации как возможный инструмент выявления опасного твита (короткого сообщения) в социальных сетях представительства органа государственной власти Белгородской области. В сб.: Современные вопросы устойчивого развития общества в эпоху трансформационных процессов: *материалы IV международной научно-практической конференции.* М.: ООО «ИРОК»; 2022. С. 136–148. https://doi.org/10.34755/IROK.2022.14.90.027, https://www.elibrary.ru/mzrsgm
4. Papadimitriou C.H., Tamaki H., Raghavan P., Vempala S. Latent semantic indexing: A probabilistic analysis. In: *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems.* ACM; 1998. P. 159–168. https://doi.org/10.1145/275487.275505
5. Hofmann T. Probabilistic latent semantic indexing. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.* ACM; 1999. P. 50–57. https://doi.org/10.1145/312624.312649
6. Blei D., McAuliffe J. Supervised topic models. In: *Advances in Neural Information Processing Systems 20* (*NIPS 2007*). 2008. P. 121–128.
7. Blei D.M., Lafferty J.D. Dynamic topic models. In: *Proceedings of the 23rd International Conference on Machine learning* (*ICML '06*). ACM; 2006. P. 113–120. https://doi.org/10.1145/1143844.1143859
8. Blei D.M. Probabilistic topic models. *Communications of the ACM*. 2012;55(4):77–84. https://doi.org/10.1145/2133806.2133826
9. Воронцов К.В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. *Доклады академии наук*. 2014;456(3):268–271. https://doi.org/10.7868/S0869565214090096
10. Воронцов К.В., Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования. *Машинное обучение и анализ данных*. 2013;1(6):657–686.
11. Нокель М.А., Лукашевич Н.В. Тематические модели: добавление биграмм и учет сходства между униграммами и биграммами. *Вычислительные методы и программирование*. 2015;16(2):215–234. https://doi.org/10.26089/NumMet.v16r222
12. Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке. *Труды Института системного программирования РАН* (*Труды ИСП РАН*). 2012;23:215–242. https://doi.org/10.15514/ISPRAS-2012-23-13
13. Nakshatri N., Liu S., Chen S., Roth D., Goldwasser D., Hopkins D. Using LLM for Improving Key Event Discovery: Temporal-Guided News Stream Clustering with Event Summaries. *Findings of the Association for Computational Linguistics: EMNLP*. 2023:4162–4173. https://doi.org/10.18653/v1/2023.findings-emnlp.274
14. Rijcken E., Scheepers F., Zervanou K., Spruit M., Mosteiro P., Kaymak U. Towards Interpreting Topic Models with ChatGPT. *2023. Paper presented at The 20th World Congress of the International Fuzzy Systems Association*, Daegu, Republic of Korea. 2023. V. 5. URL: https://pure.tue.nl/ws/portalfiles/portal/300364784/IFSA_InterpretingTopicModelsWithChatGPT.pdf
15. Amigo E., Gonzalo J., Artiles J., Verdejo F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*. 2009;12(4):461486.

REFERENCES

1. Brusentsev A.G., Zueva E.S. Thematic models and tools for processing the natural language in application to the problems of municipal structures. In: *Actual Theoretical and Applied Issues of the Socio-Economic Systems Management: Proc. Second International Scientific and Practical Conference.* Moscow; 2020. V. 2. P. 262–269 (in Russ.). https://elibrary.ru/fkgyxn
2. Zueva E.S. Probabilistic classification of incoming calls based on a controlled recurrent neurons algorithm. In: *Proc. International Scientific and Technical Conference of Young Scientists of V.G. Shukhov BSTU.* Belgorod: V.G. Shukhov BSTU; 2021. P. 3564–3575 (in Russ.). https://www.elibrary.ru/nhlzpv
3. Polyakov V.M., Mozaidze E.S. Collaborative filtering algorithm as a possible tool for detecting a dangerous tweet (short message) in social networks of a representative office of the government of the Belgorod region. In: *Modern Issues of Sustainable Development of Society in the Era of Transformation Processes*: *Collection of Materials of the 4th International Scientific and Practical conference*. Moscow; 2022. P. 136–148 (in Russ.). https://doi.org/10.34755/IROK.2022.14.90.027, https://www.elibrary.ru/mzrsgm
4. Papadimitriou C.H., Tamaki H., Raghavan P., Vempala S. Latent semantic indexing: A probabilistic analysis. In: *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems.* ACM; 1998. P. 159–168. https://doi.org/10.1145/275487.275505
5. Hofmann T. Probabilistic latent semantic indexing. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM; 1999. P. 50–57. https://doi.org/10.1145/312624.312649

6. Blei D., McAuliffe J. Supervised topic models. In: *Advances in Neural Information Processing Systems 20* (*NIPS 2007*). 2008. P. 121–128.
7. Blei D.M., Lafferty J.D. Dynamic topic models. In: *Proceedings of the 23rd International Conference on Machine learning* (*ICML '06*). ACM; 2006. P. 113–120. https://doi.org/10.1145/1143844.1143859
8. Blei D.M. Probabilistic topic models. *Communications of the ACM*. 2012;55(4):77–84. https://doi.org/10.1145/2133806.2133826
9. Vorontsov K.V. Additive regularization for topic models of text collections. *Dokl. Math.* 2014;89(3):301–304. https://doi.org/10.1134/S1064562414020185
[Original Russian Text: Additive regularization for topic models of text collections. *Doklady Akademii Nauk*. 2014;456(3): 268–271 (in Russ.). https://doi.org/10.7868/S0869565214090096]
10. Vorontsov K.V., Potapenko A.A. EM-like algorithms for probabilistic topic modeling. *Mashinnoe obuchenie i analiz dannykh = Machine Learning and Data Analysis*. 20131(6):657–686 (in Russ).
11. Nokel M.A., Lukashevich N.V. Topic Models: Adding Bigrams and Taking Account of the Similarity between Unigramsand Bigrams. *Vychislitel'nye metody i programmirovanie = Numerical Methods and Programming*. 2015;16(2):215–234 (in Russ.). https://doi.org/10.26089/NumMet.v16r222
12. Korshunov A., Gomzin A. Topic modeling in natural language texts. *Trudy Instituta sistemnogo programmirovaniya RAN* (*Trudy ISP RAN*) = *Proceedings of the Institute for System Programming of the RAS* (*Proceedings of ISP RAS*). 2012;23: 215–240 (in Russ.). https://doi.org/10.15514/ISPRAS-2012-23-13
13. Nakshatri N., Liu S., Chen S., Roth D., Goldwasser D., Hopkins D. Using LLM for Improving Key Event Discovery: Temporal-Guided News Stream Clustering with Event Summaries. *Findings of the Association for Computational Linguistics: EMNLP*. 2023:4162–4173. https://doi.org/10.18653/v1/2023.findings-emnlp.274
14. Rijcken E., Scheepers F., Zervanou K., Spruit M., Mosteiro P., Kaymak U. Towards Interpreting Topic Models with ChatGPT. *2023. Paper presented at The 20th World Congress of the International Fuzzy Systems Association*, Daegu, Republic of Korea. 2023. V. 5. Available from URL: https://pure.tue.nl/ws/portalfiles/portal/300364784/IFSA_InterpretingTopicModelsWithChatGPT.pdf
15. Amigo E., Gonzalo J., Artiles J., Verdejo F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*. 2009;12(4):461486.

Об авторе

Мозаидзе Елена Сергеевна, аспирант, кафедра программного обеспечения вычислительной техники и автоматизированных систем, ФГБОУ ВО «Белгородский государственный технологический университет им. В.Г. Шухова» (308012, Россия, Белгород, ул. Костюкова, д. 46). E-mail: mozaidze95@mail.ru. https://orcid.org/0000-0002-7919-7963

About the author

Elena S. Mozaidze, Postgraduate Student, Department of Computer Software and Automated Systems, V.G. Shukhov Belgorod State Technological University (46, Kostyukova ul., Belgorod, 308012 Russia). E-mail: mozaidze95@mail.ru. https://orcid.org/0000-0002-7919-7963