ВВЕДЕНИЕ

Для современного информационного пространства характерен стремительный рост объёмов данных, среди которых присутствуют новостные сообщения, являющиеся важным источником информации о социальных, экономических и политических процессах. Новости не только оперативно отражают состояние экономики, но и формируют общественное мнение. Однако извлечение значимой информации из огромного массива текстовых данных требует применения специализированных инструментов, способных автоматизировать анализ и визуализировать динамику изменений. Использование новостных сообщений для формирования корпуса документов с целью текстового анализа обусловлено тесной взаимосвязью между актуальными социальными проблемами и тематическими приоритетами средств массовой информации (СМИ).

Целью бакалаврской работы является разработка web-приложения для определения тематических изменений в корпусах новостных сообщений.

Для достижения поставленной цели необходимо решить следующие задачи:

- сравнительный анализ методов извлечения ключевых слов;
- разработка подхода для сравнения корпусов новостных документов;
- выбор программного инструментария для разработки web-приложения;
- разработка web-приложения для оценки изменений в корпусе новостных сообщений;
- апробация приложения для региональных новостей по социально-экономической тематике.

Предметом исследования является набор методов оценки изменений в корпусе документов. Объектом исследования является совершенствование методов анализа текстов.

Работа структурирована следующим образом. В первом разделе рассматриваются различные методы извлечения ключевых слов, основанные на статистическом и графовом подходах. Во втором разделе описывается метод определения тематических изменений в новостных корпусах сообщений. В третьем разделе описывается выбор инструментальных средств для реализации метода, указанного во втором разделе, а также описываются разработанное приложение и его апробация на примере новостных сообщений, собранных для конкретных регионов за различные периоды.

Теоретическая значимость бакалаврской работы заключается в том, что предложена методика оценки динамики тематических изменений в корпусах новостных документов. Практическая значимость заключается в том, что разработанное приложение позволяет на основе сформированных срезов сравнивать наборы ключевых слов по периодам и по объектам.

В тексте использовано 9 рисунков, 4 таблицы, 30 источников.

Результаты исследований были опубликованы в статье: Чернышова Г.Ю., Таран Е.В., Коноров Д.А. Анализ инновационного развития регионов на основе новостного контента // Математические методы в технологиях и технике. 2024. № 11. С. 42-46.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Классификация методов извлечения ключевых слов» посвящен анализу статистических и графовых методов извлечения ключевых слов. Были рассмотрены такие статистические методы, как TF-IDF, RAKE, YAKE и другие [1-3]. Также были рассмотрены следующие графовые методы: TextRank, Topic Rank, Topical PageRank и другие [4-6].

Основным недостатком тематических методов является то, что темы слишком общие. Кроме того, при использовании методов, основанных на совместной встречаемости, теряется часть информации. То есть если два слова никогда не встречаются вместе в документе, то не будет ребер для их соединения в соответствующем графе слов, даже если они семантически связаны. При использовании методов, основанных на статистике, реальные значения слов в документе могут быть перегружены большим количеством внешних текстов, используемых для расчета статистической информации. Для решения этих проблем и включения семантики в процесс извлечения ключевых фраз предлагается система извлечения ключевых фраз, использующая графы знаний.

Метод RAKE показывает хорошие результаты для русскоязычных текстов. Далее предлагается модифицировать метод RAKE для использования в задачах сравнения корпусов документов. Кроме того, в качестве альтернативы предусмотрен вариант использования метода TF-IDF.

Второй раздел «Разработка подхода для сравнения корпусов текстов» посвящен разработке метода определения тематических изменений в новостных корпусах сообщений.

Введём понятие среза данных. Под срезом понимается набор документов,

собранных за определённый период для некоторого объекта.

В качестве объекта рассмотрен определённый регион, обозначенный i. Период времени, за который рассмотрен новостной контент, обозначен как j. Срез данных за определённый период времени j для конкретного объекта i обозначен как T_{ij}.

Текстовые новостные документы, из которых состоит срез T_{ij}, обозначены как t_{ij}^l. Тогда срез определяется следующим образом:

$$T_{ij} = \{t_{ij}^1, \ldots, t_{ij}^l, \ldots, t_{ij}^k\},$$

где $1 \leq i \leq n$, n — количество объектов, $1 \leq j \leq m$, m — количество периодов, $1 \leq l \leq k$, k — количество новостных документов в T_{ij}. При этом в разных срезах может быть разное количество новостных документов.

Рассмотрим дальнейшую обработку на примере среза T_{ij}. Работа со срезом T_{ij} ведётся как с единой строкой. Результат конкатенации новостных документов среза T_{ij} обозначим T'_{ij}. То есть $T'_{ij} = t_{ij}^1 \cdot \ldots \cdot t_{ij}^l \cdot \ldots \cdot t_{ij}^k$, где знак $\cdot$ означает конкатенацию.

Над текстом T'_{ij} производятся преобразования в следующем порядке:

— удаление знаков пунктуации, цифр, лишних пробелов, специальных символов;
— приведение символов текста к нижнему регистру;
— разбиение текста на слова;
— приведение слов к начальной форме;
— удаление стоп-слов.

Результатом проведенных преобразований будет список слов, который для объекта T'_{ij} обозначим как P_{ij}.

Формирование списка ключевых слов с помощью метода TF-IDF производится следующим образом. Пусть D – коллекция списков слов $\{P_{i1}, P_{i2}, \ldots, P_{ij}, \ldots, P_{im}\}$.

$$tf\,(w,\ P_{ij}) = \frac{n_w}{\sum_a n_a},$$

где n_w – это число вхождений слова w в P_{ij}, а в знаменателе — общее число слов в P_{ij}. Обозначим

$$idf(w, D) = ln(\frac{|D| + 1}{|\{P_{ij} \in D | w \in P_{ij}\}| + 1} + 1),$$

где $|D|$ — число списков слов в коллекции; $|\{P_{ij} \in D | w \in P_{ij}\}|$ – число списков из коллекции D, в которых встречается слово w. Тогда

$$tf - idf(w,\ P_{ij},\ D) = tf(w, P_{ij}) \cdot idf(w, D).$$

Для каждого слова w из списка P_{ij} вычисляется $tf - idf(w, P_{ij}, D)$. Формируется матрица V, состоящая из элементов $v_{su} = tf - idf\,(w_s,\ d_u,\ D)$, $s = 1, \ldots, b$, $u = 1, \ldots, c$, где b — количество уникальных слов среди всех слов из коллекции D, $c = |D|$. Ключевыми словами считаются те слова, у которых значение tf наибольшее. Векторное представление для среза T_{ij}, полученное с помощью метода TF-IDF, представляет собой столбец u из матрицы V, соответствующий этому срезу.

Индекс RAKE вычисляется следующим образом. Текст T'_{ij} обрабатывается как и при получении P_{ij}, однако не производится удаление стоп-слов и знаков пунктуации. Результатом обработки является список, элементами которого являются слова и знаки пунктуации.

Далее слова из этого списка объединяются в кандидаты в ключевые фразы (ККФ) следующим образом. Элементы списка маркируются как релевантные и нерелевантные. Нерелевантные элементы рассматриваются как разделители при разбиении на ККФ. Релевантными считаются существительные и прилагательные, не являющиеся стоп-словами, а нерелевантными считаются все остальные слова и знаки пунктуации. ККФ является продолжительная последовательность релевантных элементов, не содержащая нерелевантных. Результатом является список ККФ, который обозначим K_{ij}.

Для каждого слова w, встречающегося в K_{ij}, определяются характеристики: степень $d(w)$ и частота $f(w)$. $f(w)$ определяется как частота встречаемости слова w в K_{ij}. При этом слово может быть как отдельным ККФ, так и частью какого-либо кандидата. Степень $d(w)$ определяется так:

$$d(w) = \sum_{w \in phrase} (length(phrase) - 1),$$

где $phrase$ – ККФ, $length(phrase)$ – количество слов в ККФ.

Для каждого слова w вычисляется индекс RAKE, определяемый следующим образом:

$$score(w) = \frac{d(w)}{f(w)}.$$

Индекс RAKE для ККФ $phrase$ представляет собой

$$score(phrase) = \sum_{w \epsilon phrase} score(w).$$

Создается копия списка K_{ij}, который обозначим как K'_{ij}. Задается некоторое минимальное значение min_freq. Если частота встречаемости ККФ меньше min_freq, то кандидат удаляется из списка ККФ K'_{ij}. Если ККФ входит в состав другого кандидата, то частота его встречаемости в другом кандидате не учитывается. Из-за этого становится возможен случай, в котором, например, кандидат «учебное заведение» встречается реже кандидата «высшее учебное заведение». Задается некоторое максимальное значение $ngram\_max$. Если количество слов в кандидате больше, чем $ngram\_max$, то кандидат удаляется из списка ККФ K'_{ij}. ККФ, у которых наибольший индекс RAKE, считаются ключевыми фразами.

Пусть K''_{ij} — таблица, первым столбцом которой является список уникальных ККФ из списка K'_{ij}, а вторым — индекс RAKE для соответствующих кандидатов. Обозначим $K''_{ij}(phrase)$ — значение индекса RAKE для фразы $phrase$ из таблицы K''_{ij}. Векторным представлением ККФ для среза T_{ij} является список индексов RAKE $K''_{ij}(phrase)$ для каждой фразы.

В результате получены векторные представления среза T_{ij} методом TF-IDF и методом RAKE. Рассмотрим задачу определения схожести тематик текстов, представленных в двух срезах T_{ij} и T_{iz}. Для определения схожести тематик текстов, представленных в срезах, проведем оценку меры схожести срезов на основе их векторных представлений. Для этого используется косинусная мера близости.

Пусть $R = (r_e)$ — векторное представление среза T_{ij}, а $S = (s_e)$ — векторное представление среза T_{iz}. Рассмотрим векторные представления, полученные с помощью метода TF-IDF. Тогда $e = 1, \ldots, h$, $1 \leq z \leq m$, h — количество уникальных слов среди всех слов из коллекции D, $D = \{P_{i1}, P_{i2}, ..., P_{ij}, ..., P_{im}\}$.

Если векторные представления вычисляются с помощью метода RAKE, то $e = 1, \ldots, h$, $1 \leq z \leq m$, h — количество уникальных кандидатов из объединения множеств K'_{ij} и K'_{iz}. Если фраза присутствует в K'_{ij}, но отсутствует в K'_{iz}, то индекс RAKE для этой фразы в векторе S будет равен нулю. Аналогично, если фраза есть в K'_{ij}, но она отсутствует в K'_{iz}, то индекс RAKE для этой фразы в векторе R будет равен нулю.

Косинусная мера сходства вычисляется следующим образом:

$$cos(R,S) = \frac{\sum_{e=1}^{h} r_e \cdot s_e}{\sqrt{\sum_{e=1}^{h} (r_e)^2} \cdot \sqrt{\sum_{e=1}^{h} (s_e)^2}}.$$

Если косинусная мера сходства двух срезов близка к 1, значит, их векторные представления близки. Это показывает, что списки слов или фраз, выделенные методом TF-IDF или методом RAKE как значимые, похожи. Если косинусная мера сходства близка к 0, то векторные представления срезов почти не имеют сходств.

Полезным инструментом для оценки динамики изменений в новостных документах является тренд-карта. При построении тренд-карты по оси абсцисс для каждого слова или фразы откладывается значение динамики ($dynamism$), а по оси ординат — значимость ($significance$).

Тренд-карта для объекта i строится на основе данных из срезов $T_{i1}, ..., T_{ij}, ..., T_{im}$. Значимостью для фразы $phrase$ для региона i является сумма индексов RAKE данной фразы, полученная для всех периодов для этого региона, а динамикой является средний абсолютный прирост индекса RAKE для данной фразы:

$$significance(phrase) = \sum_{j=1}^{m} K_{ij}''(phrase),$$

$$dynamism(phrase) = \frac{K_{im}''(phrase) - K_{i1}''(phrase)}{(m - 1)}.$$

Для метода TF-IDF динамика и значимость для объекта i вычисляются аналогичным образом. Однако значимостью в данном случае является частота встречаемости слова в коллекции D, а динамикой — средний абсолютный прирост частоты встречаемости этого слова:

$$significance(w) = \sum_{j=1}^{m} tf(w, P_{ij}),$$

$$dynamism(w) = \frac{tf(w, P_{im}) - tf(w, P_{i1})}{(m - 1)}.$$

Тренд-карта помогает определить новые активно обсуждаемые темы, освещаемые СМИ. Высокая динамика слова или фразы означает то, что слово или фраза стали более значимыми, чем раньше. Если динамика отрицательная, то

это значит, что слово или фраза обладают меньшей значимостью, чем ранее. Если динамика равна нулю, то значимость слова или фразы не изменилась за рассмотренные периоды. Активно обсуждаемые темы обладают высокой значимостью. Наибольший интерес представляют слова и фразы, обладающие высокими показателями динамики и значимости.

Для их визуализации осуществляется нормализация показателей значимости и динамики, то есть их отображение на отрезки от 0 до 1 по осям значимости и динамики. Термины с наибольшей суммой нормализованных показателей динамики и значимости будут отражены на графике. В приложении реализовано построение тренд-карт и матриц косинусного сходства в соответствии с представленными методиками.

Третий раздел «Разработка web-приложения для анализа региональных новостных источников» посвящен выбору инструментальных средств для реализации метода, указанного во втором разделе, а также описанию разработанного приложения и его апробации на примере новостных сообщений, собранных для конкретных регионов за различные периоды.

В качестве основного инструментального средства выбран язык программирования R [7]. Он применяется в качестве ключевого инструмента в Text Mining, сочетающего гибкость open-source экосистемы с ориентированностью на статистическую строгость и визуальную интерпретацию данных. В работе будет использован язык R версии 4.2.3. Для реализации приложения был использован web-фреймворк Shiny с открытым исходным кодом [8].

Одной из основных целей является создание удобного интерфейса, позволяющего пользователю, не используя программный код, обработать текстовые данные, получить списки ключевых слов и облака слов, а также оценить динамику изменений в новостных сообщениях, собранных за разные периоды времени.

Приложение позволяет:

- производить загрузку срезов в виде файлов;
- выбирать срез, для которого будет сформирован список ключевых слов и построено облако слов;
- выбирать метод извлечения ключевых слов, с помощью которого будет производиться извлечение ключевых слов;
- строить тренд-карту с помощью выбранного метода для проанализированных срезов;

— строить матрицу косинусного подобия с помощью выбранного метода для проанализированных срезов;
— просматривать результаты вычислений;
— сохранять результаты вычислений.

При разработке приложения был сформирован список стоп-слов, который включал в себя русскоязычные союзы, частицы, предлоги, имена, фамилии, названия регионов, для которых применялась программа, и слова, не несущие смысловой нагрузки в контексте экономической тематики. Список стоп-слов был настроен для конкретных регионов.

При вычислении индекса RAKE определяется минимальная частота встречаемости ККФ min_freq, равная 10. Максимальное количество слов в ККФ $ngram\_max$ определяется равным 3.

Разработанное приложение было протестировано на примере следующих регионов: Астраханская область, Республика Татарстан, Омская область. Эти регионы представлены в рейтинге устойчивого развития регионов [9], который охватывает 85 субъектов Российской Федерации и 250 городов России.

На 2024 год Республика Татарстан занимает пятое место этого рейтинга и относится к группе регионов-лидеров. Омская область занимает 41-е место и является представителем среднего сегмента рейтинга. Астраханская область на 2024 год занимает 77-е место и относится к отстающим регионам. Данные регионы были выбраны с целью выявления возможного влияния места региона в рейтинге на тематику новостного контента.

В процессе сбора данных из новостных ресурсов в социальной сети ВКонтакте использовался набор ключевых слов: «экономика», «конкурентоспособность», «инновации» и «инвестиции». Если ни одно из этих ключевых слов не встречается в новостном посте, то он отсеивается. Эта фильтрация позволила сузить тематику сообщений. В качестве источников выбирались региональные ресурсы ВКонтакте на основании рейтинга популярности API VK, включая официальные правительственные страницы регионов. Новостные сообщения были собраны за 2020, 2021, 2022, 2023 и 2024 годы. В настоящее время сеть ВКонтакте является популярной, общедоступной и широко распространённой социальной сетью в Российской Федерации. Были получены результаты работы программы для трех регионов за указанные годы.

В качестве примера использования приложения рассмотрим матрицу ко-

Таблица 1. Матрица косинусного сходства тематик новостного контента, полученная для трех регионов за 2020 год с помощью метода RAKE.

	Татарстан	**Омская обл.**	**Астраханская обл.**
Татарстан	1	0.28	0.18
Омская обл.	0.28	1	0.22
Астраханская обл.	0.18	0.22	1

синусного сходства тематик новостного контента, полученную для трех регионов за 2020 год с помощью метода RAKE. Таблица 1 показывает, что темы, освещаемые в 2020 году в Республике Татарстан, больше схожи с темами в Омской области, чем с темами в Астраханской области. При этом темы, освещаемые в Омской области, больше схожи с темами в Республике Татарстан, чем с темами в Астраханской области.

ЗАКЛЮЧЕНИЕ

В рамках бакалаврской работы были рассмотрены различные методы извлечения ключевых слов. Был разработан подход для оценки динамики изменений в корпусе новостных сообщений, в ходе которого были применены некоторые из рассмотренных методов извлечения ключевых слов: RAKE, TF, TF-IDF. Были выбраны инструментальные средства, необходимые для разработки web-приложения и для реализации разработанного подхода, которые были применены для новостных сообщений, собранных из сообществ социальной сети «ВКонтакте» для Омской области, для Астраханской области и для Республики Татарстан за 2020, 2021, 2022, 2023 и 2024 годы. Для этих данных были получены списки ключевых слов, вычисленные с помощью методов TF и RAKE, тренд-карты и матрицы косинусного подобия тем.

Дальнейшим направлением развития приложения является добавление различных методов извлечения ключевых слов в приложение и возможностей настройки методов извлечения ключевых слов пользователем. Результаты исследований были опубликованы в «Чернышова Г.Ю., Таран Е.В., Коноров Д.А. Анализ инновационного развития регионов на основе новостного контента // Математические методы в технологиях и технике. 2024. № 11. С. 42-46.»

Основные источники информации:

1 Salton, G., Yang, C.S. On the specification of term values in automatic indexing // J Document. — 1973. — Т.29 № 4. — С. 351–372.

2 Rose, S., Engel, D., Cramer, N., Cowley, W. Automatic keyword extraction from individual documents // Text Mining: Applications and Theory. — Chichester, U.K.: Wiley, 2010. — С. 1–20.

3 Campos, R., Mangaravite, V., Pasquali, A., Jorge, A.M., Nunes, C., Jatowt, A. YAKE! Keyword extraction from single documents using multiple local features // Information Sciences. — 2020. — Т. 509. — С. 257–289.

4 Mihalcea, R., Tarau, P. TextRank: Bringing Order into Text // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. — Barcelona, Spain: Association for Computational Linguistics, 2004. — С. 404–411.

5 Bougouin, A., Boudin, F., Daille, B. Topicrank: Graph-based topic ranking for keyphrase extraction // Proc. Int. Joint Conf. Natural Lang. Process. (IJCNLP). — Nagoya, Japan: Asian Federation of Natural Language Processing, 2013. —

C. 543-551.

6 Liu, Z., Huang, W., Zheng, Y., Sun, M. Automatic keyphrase extraction via topic decomposition // Proc. Conf. Empirical Methods Natural Lang. Process. — Cambridge, MA: Association for Computational Linguistics, 2010. — C. 366-376.

7 Официальный сайт CRAN для скачивания R для macOS. [Электронный ресурс]. URL: https://cran.r-project.org/bin/macosx/ (дата обращения: 08.05.2025).

8 Официальный сайт библиотеки Shiny. [Электронный ресурс]. URL: https://shiny.posit.co/ (дата обращения: 08.05.2025).

9 Рейтинг устойчивого развития городов и регионов России. [Электронный ресурс]. URL: https://устойчивые-территории.города.рф/ (дата обращения: 15.04.2025).