ОБОГАЩЕНИЕ БАЗЫ ЗНАНИЙ С ПОМОЩЬЮ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ СЛОВ И АННОТАЦИЙ

Э. Г. Тунян[1,2,3а], **Р. С. Сазиков**[1,2,3,б], **С. А. Харламов**[1,2в]

[1] *Сургутский государственный университет, г. Сургут, Российская Федерация*

[2] *ООО «ЕДРО», г. Сургут, Российская Федерация*

[3] *Сургутский филиал федерального государственного автономного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Национального исследовательского центра «Курчатовский институт», г. Сургут, Российская Федерация*

[а] *ORCID: https://orcid.org/0009-0003-3260-1310, ✍ tunyan@edro.su*

[б] *ORCID: https://orcid.org/0009-0005-0078-0013, sazikov@edro.su*

[в] *ORCID: https://orcid.org/0009-0000-5605-0531, harlamov_sa@surgu.ru*

Аннотация: в работе рассматриваются разные способы автоматического извлечения терминов: от давно известных и технологически несложных — вроде TF-IDF, RAKE или TextRank — до более современных решений, основанных на трансформерах, включая KeyBERT и модели типа LLM. Применение методов выделения ключевых слов показало, что именно гибридные схемы проявляют себя наиболее эффективно: когда статистика и нейросетевые модели работают не по отдельности, а в связке, удается добиться как формальной релевантности, так и смысловой глубины в отборе терминов. Предложен поэтапный подход: сначала — чистка и разметка текста, потом — параллельный запуск нескольких алгоритмов, позволяющий свести к минимуму случайные отклонения, и уже затем — более тонкое ранжирование на основе нейросетевых моделей. Такой алгоритм, основанный на комплексном подходе к автоматическому построению баз знаний, позволяет существенно улучшить качество автоматического выделения релевантных терминов и значительно повысить как точность, так и полезность извлекаемой информации. Автоматизация анализа больших текстовых массивов без необходимости предварительной разметки и обучения под конкретную предметную область дает возможность быстро масштабировать подход как для исследовательских задач, так и для прикладных решений в сфере цифровизации, например, в медицине, образовании, документообороте.

Ключевые слова: обогащение баз знаний, извлечение терминов, ключевые слова, аннотация, TF-IDF, RAKE, TextRank, KeyBERT, LLM, интеллектуальный поиск, автоматизация обработки текста, семантический анализ.

Для цитирования: Тунян Э. Г., Сазиков Р. С., Харламов С. А. Обогащение базы знаний с помощью автоматического извлечения ключевых слов и аннотаций. *Успехи кибернетики*. 2025;6(2):108–113.

Поступила в редакцию: 01.06.2025. *В окончательном варианте: 16.06.2025.*

AUTOMATIC EXTRACTION OF KEYWORDS AND SUMMARIES FOR KNOWLEDGE BASE POPULATION

E. G. Tunyan[1,2,3,a], **R. S. Sazikov**[1,2,3,b], **S. A. Kharlamov**[1,2,c]

[1] *Surgut State University, Surgut, Russian Federation*

[2] *LLC Edro, Surgut, Russian Federation*

[3] *Surgut Branch of Scientific Research Institute for System Analysis of the National Research Centre "Kurchatov Institute", Surgut, Russian Federation*

[a] *ORCID: https://orcid.org/0009-0003-3260-1310, ✍ tunyan@edro.su*

[b] *ORCID: https://orcid.org/0009-0005-0078-0013, sazikov@edro.su*

[c] *ORCID: https://orcid.org/0009-0000-5605-0531, harlamov_sa@surgu.ru*

Abstract: we studied a range of automatic term extraction methods, from well-established techniques such as TF-IDF, RAKE, and TextRank to recent transformer-based approaches, including KeyBERT and large language models (LLMs). Our findings show that hybrid approaches — combining statistical and neural methods – achieve superior performance by ensuring both formal relevance and semantic depth in term selection. We proposed a multi-stage pipeline: initial text preprocessing and annotation, followed by the parallel application of several extraction algorithms to minimize stochastic variation, and final refinement through neural network-based ranking. This integrated algorithm significantly improves the quality of term

extraction and enhances both the accuracy and practical utility of the retrieved information. The proposed method enables scalable analysis of large unstructured text corpora without the need for manual annotation or domain-specific training, making it suitable for a wide range of research and applied digitalization tasks, including applications in medicine, education, and document management.

Keywords: knowledge enrichment, term extraction, keywords, abstract, TF-IDF, RAKE, TextRank, KeyBERT, LLM, intelligent search, automated text processing, semantic analysis.

Cite this article: Tunyan E. G., Sazikov R. S., Kharlamov S. A. Automatic Extraction of Keywords and Summaries for Knowledge Base Population. *Russian Journal of Cybernetics*. 2025;6(2):108–113.

Original article submitted: 01.06.2025. *Revision submitted: 16.06.2025*.

Введение

В современном мире технологии искусственного интеллекта (ИИ) стремительно развиваются и интегрируются во все сферы деятельности человека. Развитие будет набирать еще большие обороты, что подтверждается объемом инвестиций. Инвестиции США, Европы и Израиля в 2024 году на развитие ИИ и облачных технологий увеличились на 27% — с 62,5 млрд до 79,2 млрд долларов. В 2025 году OpenAI, Oracle и SoftBank инвестировали 500 млрд долларов, Национальный фонд Китая — 138,01 млрд долларов.

С развитием ИИ векторные базы данных требуют эффективных методов структурирования и поиска информации. В качестве метаданных документа служат аннотации и ключевые слова. На фоне экспоненциального роста текстовых данных, лишенных четкой структуры, что, впрочем, закономерно в условиях ускоряющейся цифровизации и активного внедрения ИИ-инструментов, постепенно обозначается нехватка средств, способных не просто обрабатывать, но и вычленять из потока существенные смысловые фрагменты. В частности, термины и аннотации, без которых ни поиск, ни генерация, ни построение когнитивных карт не могут считаться полноценными.

Поскольку ручное выделение ключевых слов и индексирование является слишком трудоемким и времязатратным, используется автоматическое извлечение из текста слов и фраз, наиболее подходящих по смыслу (Automatic Keyword Extraction). Подобные данные могут применяться для обогащения баз знаний, к примеру, в качестве тегов для документов или атрибутов для баз знаний.

Классические подходы выделения ключевых фраз используют принцип статистического анализа, то есть берут термины, которые чаще всего встречаются. Принцип Term Frequency — Inverse Document Frequency (FI-IDF) дает каждому термину свой вес пропорционально его частоте в документе и обратно пропорционально его встречаемости в совокупности текстов. Этот метод впервые был описан в работе Karen Spärck Jones в 1972 году [1], которая стала основой для множества поисковых систем.

В последнее время было разработано множество эффективных методов выделения основных фраз без учителя (unsupervised), такими алгоритмами являются Rapid Automatic Keyword Extraction (RAKE) и TextRank. Алгоритм RAKE, представленный S. Rose с соавторами в 2010 году [2], разделяет весь текст на слова, при этом отбрасывает все стоп-слова, выделяет всех кандидатов, строит граф связей между словами, определяет важность, основываясь на частоте встречаемости слова. Алгоритм TextRank был предложен в 2004 году Mihalcea R. и Tarau P. [3]. Это графовый алгоритм, который основывается на алгоритме ссылочного ранжирования PageRank, используемом в том числе Google. TextRank строит граф слов или целых предложений, где вершины — это слова или фразы, а в качестве ребер выступают связи этих слов и фраз. После этого вычисляются веса вершин по принципу: наиболее важные узлы считываются словами или предложениями.

В последние годы модели на основе трансформеров (BERT) и больших языковых моделей (LLM) начали применять для задач выделения ключевых слов, так как они извлекают более осмысленные фразы. К примеру, метод KeyBERT использует эмбеддинги BERT для формирования словосочетаний по смыслу. Также генеративные модели, такие как GPT4, DeepSeek, могут выделять ключевые слова без дополнительного обучения (zero-shot), просто получив запрос от пользователя на естественном языке. Кроме того, такие модели могут генерировать аннотации, что также полезно для построения баз знаний.

Цель текущего исследования — сравнить разные подходы к автоматическому выделению ключевых слов и оценить их эффективность для построения баз знаний.

Материалы и методы

В данном разделе описаны основные алгоритмы и методы автоматического выделения ключевых слов, фраз, аннотаций из документов для формирования баз: статические (TF-IDF), графовые (TextRank), RAKE — метод на основе правил, с использованием трансформеров (BERT) и методов применения больших языковых моделей (LLM).

TF-IDF — один из самых первых и широко используемых методов для извлечения главных терминов из документа. Задача — оценить значимость слова на основе того, как часто встречается слово в документе (TF) и обратно пропорционально его встрече в совокупности текстов (IDF).

Формула веса:

$$TFIDF\,(t,d,D) = TF\,(t,d) \cdot IDF\,(t,D)\,, \tag{1}$$

где:
t — слово,
d — конкретный документ,
D — совокупность документов,
TF — частота термина в документе,
IDF — обратная частота документа.
Формула частоты термина в документе:

$$TF(t,d) = \frac{f_{t,d}}{\Sigma_{t'\epsilon d} f_{t',d}}, \tag{2}$$

где:
$f_{t,d}$ — количество вхождений термина t в документ d,
$\Sigma_{t'\epsilon d} f_{t',d}$ — общее количество слов в документе.

Обратная частота документа:

$$IDF\,(t,D) = log\left(\frac{|D|}{1 + |d\epsilon D : t\epsilon d|}\right), \tag{3}$$

где:
$|D|$ — общее количество документов в корпусе,
$|d\epsilon D : t\epsilon d|$ — количество документов, содержащих термин t,
1 нужно для того, чтобы избавиться от деления на 0.

К примеру, для научной статьи TF-IDF выделит термины, относящиеся к определенной области науки, игнорируя простые слова. Несмотря на относительную простоту реализации, TF-IDF остается одним из эффективных методов быстрой индексации текстов [4], но, так как не учитывается порядок слов, ключевые слова могут быть пропущены, также не учитывается семантика и контекст употребления слова, что негативно влияет на качество результата.

Алгоритм RAKE создан для быстрого автоматического, то есть без предварительной обработки текста, выделения ключевых слов из конкретного текста. RAKE использует тот факт, что ключевые слова, в основном, это существительные, прилагательные и редко являются стоп-словами. По этой причине текст разбивается на слова с использованием знаков препинания и стоп-слов как разделителей. Полученные слова выступают в качестве кандидатов и оцениваются по алгоритму, учитывающему частоту слов. В первоначальной реализации каждому уникальному слову задается оценка, которая равна степени, деленной на частоту. Степень слова — это количество других слов, с которыми идет сравнение. Частота — это общее количество, сколько раз встречается кандидат в тексте. Затем оценка фразы считается суммой оценок слов, которые в нее входят.

RAKE обладает рядом преимуществ, одним из которых является его непривязанность к конкретной предметной области. Не требуется заранее подготовленный текст для обучения, и не нужны внешние словари, легко настраивается под предметную область, обладает высокой скоростью вычисления, которая растет пропорционально размеру текста и равна O(n).

Алгоритм TextRank — это графовый алгоритм, созданный для ранжирования важности отрывков текста, предложенный исследователями Mihalcea R. и Tarau P. в 2004 году. Он применяется как для сортировки текста, так и для извлечения ключевых слов или фраз. Для выделения фраз берется k слов

с самыми высокими весами, после чего по ним, находя в тексте смежные слова, восстанавливаются фразы. Разработчики показали, что TextRank выдавал результат, сопоставимый с лучшими на то время системами на таких тестах, как Inspec и SEMEval.

Основным преимуществом TextRank является учет структуры текста. Если два слова часто встречаются рядом или связаны через стоп-слова, то алгоритм рассматривает это как усиление связи. Так находятся качественные словосочетания, даже если слова часто встречаются. Если фраза состоит из двух слов, то оба слова могут выйти в топ. Чтобы решить эту проблему, используется постобработка и группировка смежных слов. Несмотря на имеющиеся недостатки, такие как чувствительность к параметрам, высокая трудоемкость, TextRank стал основой для многих алгоритмов.

Контекстуальные методы, основанные на BERT, и появление его модификаций дали новые возможности для извлечения ключевых слов с учетом контекста. Одним из самых эффективных инструментов является KeyBERT, который был придуман M. Grootendorst в 2020 году [5]. Модель создает вектор всего документа при помощи BERT, а далее находит слова и словосочетания, векторы которых максимально близки к эмбеддингу документа. В отличие от TF-IDF, BERT находит слова, наиболее точно отражающие смысл, учитывает синонимы и контекст поиска. KeyBERT, кроме того, позволяет записывать выражения разными словами, задавать желаемую длину фраз и применять механизм Maximal Marginal Relevance (MMR) для повышения разнообразия ключевых слов, чтобы они не были слишком похожи друг на друга. Эксперименты показывают, что даже без обучения на заранее подготовленных данных метод BERT обходит традиционные алгоритмы.

Методы, основанные на LLM, такие как GPT, DeepSeek, Mistral, используют самые последние достижения для работы с текстом. Данные модели показали возможность решать новые задачи без предварительного обучения и дообучения, применяя текстовый запрос (prompt). За счет ответов на естественном языке на уровне специализированных моделей LLM дают возможность выделять ключевые слова достаточно качественно. В отличие от предыдущих алгоритмов, LLM не опираются просто на правила, а воспринимают содержимое по смыслу, даже если тема указана неявно. Основным недостатком данных моделей является то, что они работают по принципу черного ящика, то есть ответ могут дополнить лишними словами и фразами. У каждой модели есть свои особенности, и в совокупности они дают максимальную гибкость, могут выделять ключевые термины, писать аннотации, даже делать поиск по заданному тексту, с дополнением и обогащением баз знаний при помощи больших языковых моделей, могут служить универсальным интерфейсом для необработанного текста, из которого будет получена структурированная база знаний, однако имеется проблема с оценкой результата этого метода из-за возможных галлюцинаций.

Результаты и их обсуждение

Многоэтапный подход для обогащения базы знаний, который при помощи LLM автоматически извлекает ключевые слова с последующей аннотацией, будет показывать хорошие результаты. Идея в том, чтобы не просто выделять из документа термины, а пополнять их информацией, которая связана по смыслу, и затем добавлять эти данные в базу. Такой подход значительно улучшит поиск по базе знаний. Предлагаемый метод (рис.) представляет собой многоэтапный алгоритмический конвейер, основанный на интеграции классических статистических подходов с современными нейросетевыми технологиями.

На первом этапе исходный текст подвергается глубокой предварительной обработке. В рамках данной стадии осуществляется очистка документа от лишних символов и HTML-тегов, разбиение текста на предложения, токенизация, лемматизация (приведение слов к их базовой форме) и выделение именованных сущностей (NER). Такой комплексный анализ обеспечивает создание корректного представления исходных данных для дальнейшей обработки.

Далее, используя обработанный текст, параллельно запускаются несколько алгоритмов для выделения кандидатов в ключевые слова. Среди них классический метод RAKE, основанный на анализе стоп-слов и частотном анализе, графовый алгоритм TextRank, применяющий принципы ранжирования по аналогии с PageRank, а также современные подходы на базе трансформеров — KeyBERT и LLM-модуль, способный в режиме zero-shot извлекать ключевые термины посредством генеративных моделей. Полученные списки кандидатов объединяются с последующим удалением дублирующих элементов и уточняются посредством финального ранжирования, выполняемого с помощью LLM, что позволяет оценить релевантность каждого термина с учетом контекстуальных связей.

Заключительным этапом является автоматическая генерация аннотаций для выделенных ключевых слов. Для каждого термина формируется краткое определение, описание и при необходимости указываются ссылки на внешние источники или соответствующие элементы онтологии. Полученные метаданные интегрируются в базу знаний, что не только обеспечивает ускоренное индексирование, но и повышает семантическую связанность информационных систем. Примечательно, что этап аннотирования здесь рассматривается не как формальное приложение, а, скорее, как звено, связывающее смысловую обработку с онтологическим представлением.

Таким образом, представленный метод, сочетающий традиционные алгоритмы (например, TF-IDF, RAKE и TextRank) с современными нейросетевыми подходами (KeyBERT, LLM), позволяет су-

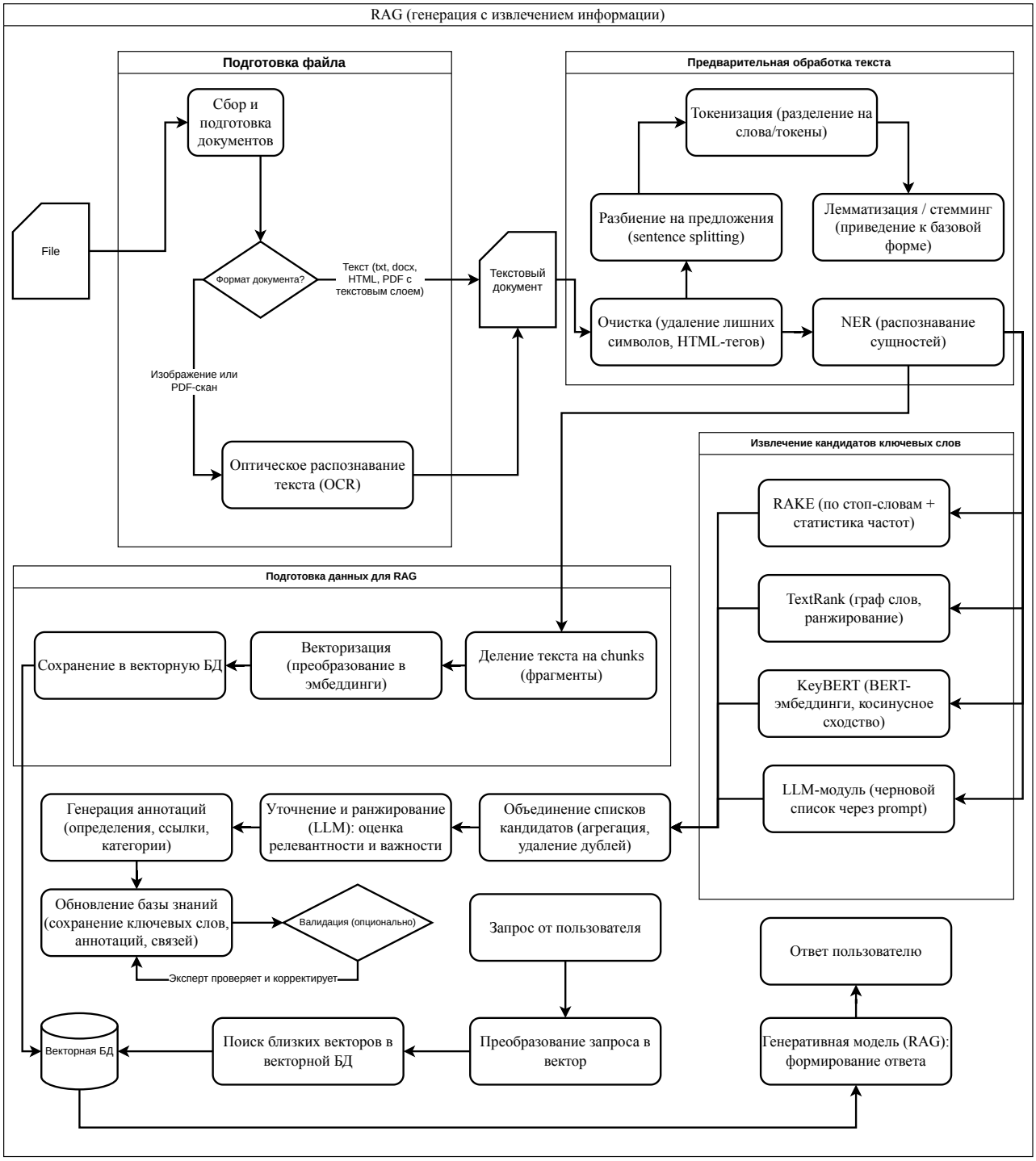

Рис. *Схема комплексного автоматического выделения ключевых слов, аннотирования и поддержки RAG*

щественно улучшить качество автоматического выделения релевантных терминов, повысить точность поиска и обеспечить оперативное обновление баз знаний. Если говорить о применимости, то метод оказывается достаточно гибким — особенно в тех случаях, когда требуется подстроиться под конкретную задачу или корпус. В определенных условиях — особенно при наличии предварительной адаптации под специфику задачи — метод способен обеспечить более точное смысловое считывание данных, что, в свою очередь, может оказаться полезным при разработке или эксплуатации систем, ориентированных на аналитическую поддержку принятия решений.

Заключение

Проведенное исследование подтвердило, что комплексный подход к автоматическому построению баз знаний, действительно, может значительно повысить как точность, так и полезность извлекаемой информации. Применение методов выделения ключевых слов — от классических алгоритмов вроде TF-IDF, RAKE и TextRank до современных решений, построенных на трансформерах (KeyBERT, LLM), — показало, что именно гибридные схемы работают наиболее эффективно. Когда статистика и нейросетевые модели работают не по отдельности, а в связке, удается добиться как формальной релевантности, так и смысловой глубины в отборе терминов.

Предложенный многоэтапный метод, в котором на каждом шаге происходит фильтрация, параллельное извлечение терминов через разные алгоритмы и финальное ранжирование с помощью LLM, позволяет получить не просто набор ключевых слов, а по-настоящему осмысленную структуру знаний. Генерация аннотаций на основе этих слов усиливает связность информации: каждый термин получает краткое пояснение, контекст, что делает такие базы особенно пригодными для задач вопрос-ответ, семантического поиска или тематической навигации.

Практика показала, что именно подключение языковых моделей — в том числе zero-shot генерации через GPT и аналоги — открывает принципиально новые горизонты. Автоматизация анализа больших текстовых массивов без необходимости предварительной разметки и обучения под конкретную предметную область дает возможность быстро масштабировать подход под разные сферы — от научных архивов до внутренних корпоративных баз.

Разработанная архитектура обладает высокой гибкостью: алгоритмы можно комбинировать, заменять, подстраивать под особенности входных данных. Она подходит как для исследовательских задач, так и для прикладных решений в сфере цифровизации, например, в медицине, образовании, документообороте. В условиях растущих объемов данных и быстрого внедрения ИИ подобные инструменты становятся не вспомогательными, а ключевыми элементами в новых информационных экосистемах.

ЛИТЕРАТУРА

1. Sparck Jones K. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*. 1972;28:11–21.
2. Rose S., Engel D., Cramer N., Cowley W. Automatic Keyword Extraction from Individual Documents. *Text Mining: Applications and Theory*. 2010. DOI: 10.1002/9780470689646.ch1.
3. 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; 2004:404–411.
4. Маннинг К. Д., Рагхаван П., Шютце Х. *Введение в информационный поиск*. Вильямс; 2020. 528 с. ISBN 978-5-907203-20-4.
5. Grootendorst M. *KeyBERT: Minimal Keyword Extraction with BERT*. Режим доступа: https://www.maartengrootendorst.com/blog/keybert/.