ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ И СУММАРИЗАЦИЯ ТЕКСТОВ В ОБЛАСТИ КИБЕРБЕЗОПАСНОСТИ

Васильев В.И.[1], Вульфин А.М[2]., Кучкарова Н.В.[3]

Цель исследования: *повышение качества анализа текстовых документов за счет применения моделей машинного обучения и интеллектуального анализа в задачах реферирования и тематического моделирования, что позволит снизить нагрузку на эксперта, выполняющего анализ и обобщение значительных объемов слабоструктурированных текстовых данных по тематике информационной безопасности из различных источников.*

Метод исследования: *для оперативной обработки и анализа больших объемов разнородной, плохо структурированной информации на естественном языке (ЕЯ) использованы методы машинного обучения. Применены методы тематического моделирования и суммаризации текстов на основе глубоких нейронных сетей, включая нейросетевые языковые модели на базе архитектуры трансформеров.*

Полученные результаты: *выделены основные этапы машинной процедуры тематического моделирования и суммаризации профессиональных текстов в области информационной безопасности. Приводятся результаты сравнительной оценки эффективности применения для этих целей моделей кластеризации, латентно-семантического анализа, языковых моделей FastText, Text Rank и трансформеров BERT. Даны рекомендации относительно перспектив практического применения этих моделей в качестве средств интеллектуальной поддержки профессиональной деятельности специалистов в области кибербезопасности.*

Научная новизна: *предложен комплекс моделей машинного обучения для тематического моделирования и суммаризации профессиональных текстов, основанный на нейросетевых моделях вложений и моделях-трансформерах, отличающийся алгоритмом подготовки корпуса текстов для обучения моделей и применением алгоритма переноса обучения, что позволит повысить эффективность анализ и обобщения предметно-ориентированных корпусов текстов.*

Ключевые слова: *интеллектуальный анализ текстов, векторное вложение слов, трансформер, кластеризация текстов, реферирование, информационная безопасность.*

DOI:10.21681/2311-3456-2023-2-2-22

Введение

Сфера обеспечения кибербезопасности объектов критической информационной инфраструктуры (КИИ) сегодня в значительной степени связана с инженерией знаний (Knowledge Engineering) – направлением искусственного интеллекта (ИИ), обеспечивающим методологическую и практическую поддержку процессов получения, представления, структурирования и использования знаний на этапе принятия решений. Значительную помощь специалистам сегодня оказывают источники дополнительной оперативной информации (бюллетени рассылок, веб-сайты, форумы, научные статьи, материалы конференций), просматривая которые с определенной степенью регулярности можно быть в курсе последних «свежих» новостей – уязвимостей «нулевого дня», новых разновидностей угроз безопасности информации, «нашумевших» компьютерных атак на промышленные предприятия и организации, новых решений в области создания проактивных систем защиты от вредоносного ПО и многое другое. В то же время, зачастую физически специалист по информационной безопасности (ИБ) в силу своей профессиональной занятости не располагает временем для сколько-нибудь детальной проработки указанной дополнительной информации, хотя и

1 Васильев Владимир Иванович, доктор технических наук, профессор, Уфимский университет науки и технологий, г. Уфа, Россия. E-mail:vasilev.vi@ugatu.su

2 Вульфин Алексей Михайлович, кандидат технических наук, доцент, Уфимский университет науки и технологий, г. Уфа, Россия. E-mail:vulfin.am@ugatu.su

3 Кучкарова Наиля Вакилевна, старший преподаватель, Уфимский университет науки и технологий, г. Уфа, Россия. E-mail: kuchkarova.nv@ugatu.su

ощущает настоятельную необходимость в ее использовании. Этим обстоятельством и объясняется повышенный интерес к разработке эффективных методов тематического моделирования и автоматического реферирования, изложения в сжатой, концентрированной форме текстовых сообщений, достаточной для предварительного ознакомления, уяснения сути (смысла) исходного сообщения.

Под тематическим моделированием в общем случае понимается построение модели коллекции (корпуса) текстовых документов, которая определяет, к каким темам относится каждый из документов. Для решения этой задачи обычно используются методы кластеризации, такие как метод k-средних (k-means), латентно-семантического анализа (Latent Semantic Analysis, LSA), вероятностного семантического анализа (probabilistic LSA, pLSA), скрытого распределения Дирихле (Latent Dirichlet Allocation, LDA) и др.[4],[5] [1, 2]. Методы суммаризации текста, как правило, делят на две большие группы – экстрактивные и абстрактивные методы [2-4]. Экстрактивные методы суммаризации базируются на извлечении из исходного текста наиболее значимых информационных блоков (абзацев, предложений, ключевых слов); они обладают интуитивной понятностью и простотой реализации, но невысоким качеством. В отличие от них, абстрактивные методы заключаются в генерации краткого содержания с порождением нового текста, содержательно обобщающего первичный документ; обладают более высоким качеством, хотя и более сложны в реализации. Существенный прогресс в развитии абстрактивных методов суммаризации связан с их реализацией с помощью методов машинного обучения, и в том числе рекуррентных нейронных сетей (НС), сверточных НС, сетей долгой краткосрочной памяти (LSTM), хорошо зарекомендовавших себя при обработке текстов как последовательностей чередующихся взаимосвязанных слов[6] [5, 6].

Дальнейший импульс в развитии абстрактивных методов суммаризации связан с появлением нового поколения глубоких НС – нейросетевых языковых моделей, основанных на архитектуре трансформера, впервые предложенной в 2017 г. в работе [7]. Об уникальных возможностях этого нового направления говорит и обилие работ последних лет, посвященных построению систем автоматической суммаризации текстов с применением технологий трансформеров BERT, GPT-3, T5[7],[8] и др. [8-14].

В последующих разделах приводятся некоторые базовые сведения относительно методов и алгоритмов обработки естественного языка (ЕЯ), а также результаты проведенных экспериментальных исследований, заключающихся в реализации тематического моделирования и суммаризации текстов журнальных статьей в области кибербезопасности с использованием стандартных библиотек машинного обучения.

2. Методы и алгоритмы обработки ЕЯ

Предобработка текста (text pre-processing) – переводит текст на ЕЯ в формат, удобный для дальнейшей работы. Основные этапы предобработки:

- нормализация (приведение символов текста к нижнему регистру, удаление знаков пунктуации, чисел, пробельных символов);
- токенизация (разбиение текста на отдельные слова, более мелкие текстовые единицы – токены);
- удаление стоп-слов (союзов, предлогов);
- стемминг (удаление суффиксов, приставок, окончаний);
- лемматизация (приведение слов к начальной форме – именительный падеж, единственное число существительного, инфинитив глагола и т.п.).

Латентно-семантический анализ (LSA) – находит статистические зависимости множества анализируемых документов (предварительно обработанного текста) и термами (словами, n-граммами), которые они содержат. Результатом применения LSA является матрица «термы-документы», количество строк которой равно количеству различных термов (слов), встречающихся в документах, а количество столбцов – количеству документов. Элементы этой матрицы представляют собой частоту $TF(w_i, d_j)$ – (от англ. слов: term frequency) присутствия i-го терма w_i в j-ом документе d_j.

В качестве элементов матрицы «термы-документы» также могут использоваться показатели TF-IDF (от англ.: term frequency – inverse document frequency) –

4 Threat intelligence: What is it, and How Can it Protect You from Today`s Advanced Cyber-Attacks? URL: https://www.gartner.com/imagersv/media-products/pdf/webroot/issue1_webroot.pdf

5 Mariott M. Vulnerability Intelligence: a Best Practices Guide. URL: https://www.digitalshadows.com/blog-and-research/vulnerability-intelligence-a-best-practice-guide/

6 Goncales L. Automatic Text Summarization with Machine Learning – An overview. URL: https://medium.com/luisfredgs/automatic-text-summarization-with-machine-learning-an-overview-68ded5717a25

7 Automatic text summarization system using Transformers – Are you tired of reading a long paper? URL: https://neurondai.medium.com/automatic-text-summarization-system-using-transformers-are-you-tired-of-reading-a-long-paper-22d2cd9f5260

8 Summarization with Transformers: Setting up for Success. URL: https://www.sicara.fr/blog-technique/summarization-with-transformers-setting-up-for-success

«частота терма – обратная частота документа». Значение TF-IDF определяет вес (важность) каждого терма (слова) в контексте анализируемого набора документов (D). Больший вес в TF-IDF получат термы (слова) с высокой частотой в пределах конкретного документа и с низкой частотой в других документах. И наоборот, если какой-то терм w_i присутствует во всех документах, то $IDF(w_i, D) = 0$.

Модель *скрытого распределения Дирихле* (LDA) отличается от LSA тем, что в ней документы представлены как вероятностная смесь скрытых (латентных) тем, при этом в явном виде моделируется распределение слов в каждой теме, а также априорное распределение тем в документе.

Векторное вложение слов (word embedding) – это преобразование обработанного текста в векторную форму, т.е. каждому слову из текста ставится в соответствие некоторый числовой вектор в векторном семантическом пространстве заданной размерности. Основное условие такого преобразования: близкие по своему смыслу слова должны получить близкие, т.е. рядом расположенные векторы в семантическом пространстве. Одним из наиболее распространённых алгоритмов векторизации слов является алгоритм Word2Vec [10], реализуемый с помощью двух моделей обучения: CBOW (Continuous Bag-of-Words – «непрерывный мешок слов») и Skip-gram. В модели CBOW по контексту подбирается наиболее вероятное слово, а в модели Skip-gram по слову предсказывается слово из его контекста.

В основе, использованной нами в ходе дальнейших экспериментов предиктивной языковой модели формализации текстовых документов BERT лежит архитектура Transformer, использующая механизм внимания (Self-Attention) [7], что позволяет использовать данную модель для анализа длинных контекстных зависимостей между словами в тексте. Модель BERT может работать с текстами без предварительной нормализации (без фильтрации и лемматизации) благодаря внутренней предобработке входных данных с помощью алгоритма Byte Pair Encoding (BPE) [11], что существенно упрощает подготовку набора данных. Модель BERT изначально не предназначена для задач информационного поиска, кластеризации предложений и текстов. В работе [12] предложена модель Sentence-BERT (SBERT) – развитие модели BERT для построения векторов предложений текста, степень семантической близости которых оценивается с помощью косинусного расстояния. Русскоязычные предобученные модели BERT, SBERT (12-слоев, 768-скрытых, 12-голов, 180 миллионов параметров) и FastText доступны в библиотеке DeepPavlov[9] и в репозитории[10].

Для визуализации результатов векторного вложения можно воспользоваться алгоритмами понижения размерности t-SNE (англ.: t-distributed stochastic neighbor embedding – «стохастическое вложение соседей с t-распределением») [13].

При решении задач суммаризации текста важную роль играет выбор критерия оценки качества результатов суммаризации. Для этих целей можно воспользоваться оценкой ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [14], которая показывает, как сильно пересекаются два разных текста (в данном случае, это «правильное» краткое содержание текста и сгенерированное с помощью выбранного метода суммаризации). В общем виде, оценка ROUGE-n вычисляется как среднее гармоническое между оценками точности ROUGE-n-Precision и полноты ROUGE-n-Recall, где:

ROUGE-n-Precision = (количество пересекающихся словосочетаний из n слов) / (количество словосочетаний из n слов в сгенерированном тексте);

ROUGE-n-Recall = (количество пересекающихся словосочетаний из n слов) / (количество словосочетаний из n слов в правильном тексте).

Обычно при оценке эффективности по результатам экспериментов подсчитываются оценки ROUGE-1, ROUGE-2 и ROUGE-L, где L означает, что поиск пересечений идет не по фиксированному размеру словосочетаний (n-грамм), а по их наибольшему размеру.

Еще одной такой метрикой оценки является BLEU (Bilingual Evaluation Understudy) [15]. Алгоритм BLEU сравнивает последовательные фразы машинного перевода текста с последовательными фразами, которые он находит в эталонном переводе и взвешенно подсчитывает количество совпадений.

3. Результаты экспериментальных исследований

Исходными данными для построения проблемно-ориентированного корпуса текстов послужили полнотекстовые статьи выпусков журнала «Вопросы кибербезопасности» за период с 2013 по 2022 гг., размещенные в сети Интернет.

Общий план проведения экспериментов представлен на рис. 1.

9 DeepPavlov. URL: https://github.com/deepmipt/DeepPavlov

10 model-zoo. URL: https://github.com/sberbank-ai/model-zoo

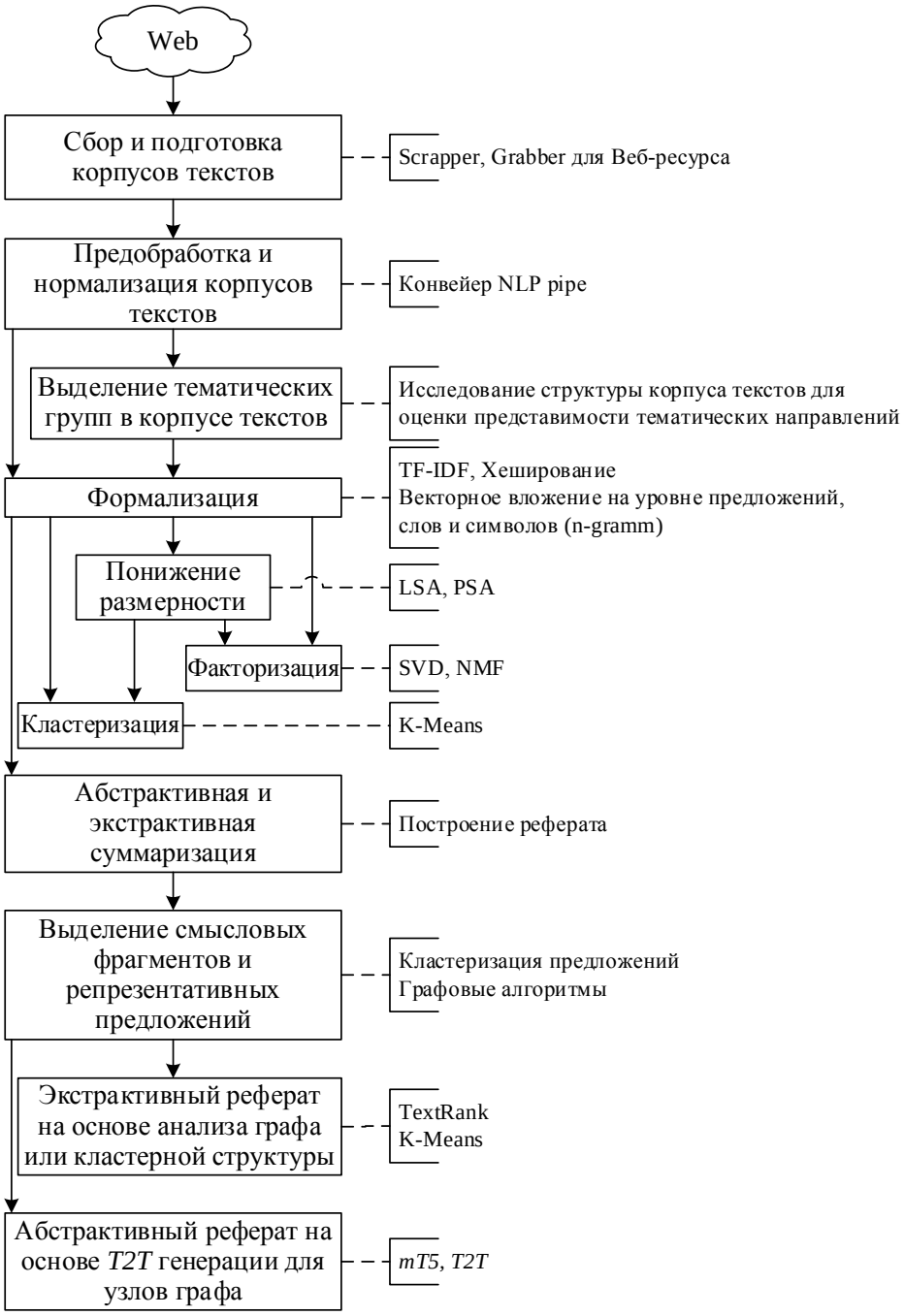


Рис. 1. Общий план построения и исследования проблемно-ориентированного корпуса текстов

3.1 Модуль сбора и подготовки корпуса текстов

Для автоматизации сбора и подготовки искомых данных разработан модуль, изображенный на рис. 2.

Данный модуль работает следующим образом. С помощью поискового робота (2) выполняется анализ стартовой HTML-страницы (1) с целью поиска и извлечения ссылок на документы формата pdf по схеме «обход в ширину». Собранные документы в формате pdf (всего найдено 438 документов) помещаются в хранилище исходных файлов ($БД_1$).

Следующий этап подразумевает получение текстовой информации из собранных pdf-документов с помощью модуля извлечения текстового слоя (3) для корректно созданных файлов или применения инстру-

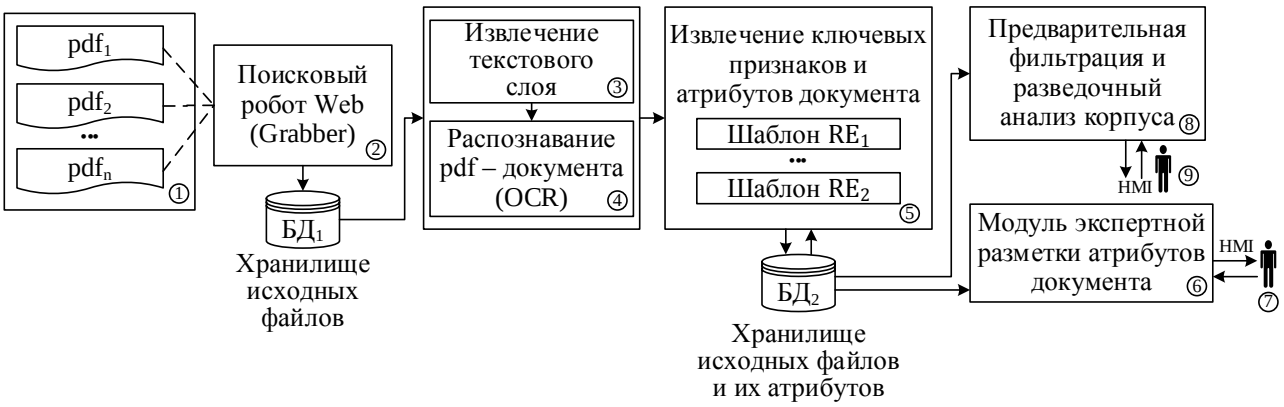


Рис. 2. Структура модуля сбора и подготовки корпуса документов

Таблица 1

Структура текстов и аннотаций

Статистика в разрезе	Параметр	Значение
Текстов	Максимальное количество слов	10488
	Минимальное количество слов	483
	Среднее количество слов	3381
	Среднее количество предложений	128
	Среднее количество уникальных слов	1149
	Среднее количество уникальных лемм	811
Аннотаций	Максимальное количество слов	340
	Минимальное количество слов	1
	Среднее количество слов	98,4
	Среднее количество предложений	4,57
	Среднее количество уникальных слов	67,2
	Среднее количество уникальных лемм	59,8
Пересечение текстов и аннотаций	Среднее по числу уникальных слов	67,8 %
	Среднее по числу уникальных лемм	59,8 %
	Среднее по числу слов	98,96 %
	Среднее по числу лемм	98,95 %
	Среднее по числу биграмм	93,97 %
	Среднее по числу триграмм	80,92 %

ментов распознавания (OCR)[11] pdf-файлов как набора изображений (4) для документов со сложной структурой и верстkой.

Далее осуществляется извлечение с помощью набора разработанных регулярных выражений (5) основных атрибутов и ключевых признаков документа: название, авторы, аннотация, ключевые слова, DOI, номер журнала (локальный и глобальный), год выпуска, текст статьи, список литературы. Текстовый документ и его атрибуты размещаются в документ-ориентированном хранилище ($БД_2$). Модуль экспертной разметки атрибутов документа (6) с участием человека (7) позволяет откорректировать ошибочно выделенные значения, например разрыв строки в названии, пропуск блока литературы и т.п.

С помощью[12] категоризации n-грамм, выделяемых из текста, и их частотного анализа определяется язык документа: русский или английский. Дальнейший анализ выполнен для 321 статьи на русском языке с непустой аннотацией.

11 EasyOCR. URL: https://github.com/JaidedAI/EasyOCR

12 language-detection. URL: https://github.com/shuyo/language-detection

Таблица 2

Структура конвейера обработки текстовых данных

Этап	Шаги	Действия	Инструменты
Предобработка	Символьная фильтрация	Удаление нерелевантных символов, HTML-тегов	Набор регулярных выражений
	Токенизация	Разбивка текста на токены с помощью предобученной для русского языка нейросетевой модели	Razdel (фреймворк Natasha), Spacy, Stanza, nltk
	Фильтрация нерелевантных токенов	Удаление ссылок, нерелевантных сокращений	Регулярные выражения
Нормализация	Лемматизация	Приведение слов в исходную форму с помощью предобученной нейросетевой модели	Morph (фреймворк Natasha), pymorphy2, spacy
Постобработка	Частеречная фильтрация	Остаются только существительные, глаголы, прилагательные, наречия, местоимения	Morph (фреймворк Natasha)
	Извлечение именованных сущностей	Разметка тегами выделенных типов именованных сущностей	Natasha, spacy
	Фильтрация на основе стоп-словарей	Фильтрация нерелевантных лемм с помощью составного стоп-словаря, включающего наиболее часто встречающиеся слова корпуса текстов	NLTK-russian, english
	Формирование документа-строки	Объединение лемм в нормализованную строку-документ	

Далее выполняется предварительная фильтрация и разведочный анализ (8) корпуса текстов, позволяющий оценить предварительное соотношение текстов документов и их аннотаций (таблица 1).

Особенностью построенного корпуса документов является следующее:

1) и аннотации, и тексты документов достаточно объемны;

2) вариабельность (длина, количество предложений) текстов – средняя;

3) аннотации не являются экстрактивными.

3.2 Структура конвейера обработки текстовых данных

Методы и инструменты обработки текстовых документов, извлечения внутренней структуры на уровне частей документа (разделы, абзацы, предложения), частеречной разметки, семантической разметки (извлечение именованных сущностей, терминов), нормализации и формализации на основе алгоритмов векторных вложений различного уровня с привлечением нейросетевых предобученных моделей описаны в таблице 2 [16].

На рис. 3а показана диаграмма встречаемости слов в корпусе документов, а на рис. 3б – ключевые слова, полученные с портала elibrary.ru.

Анализ частоты встречаемости слов показывает достаточно полное соответствие основным ключевым словам, определяемым авторами журнала.

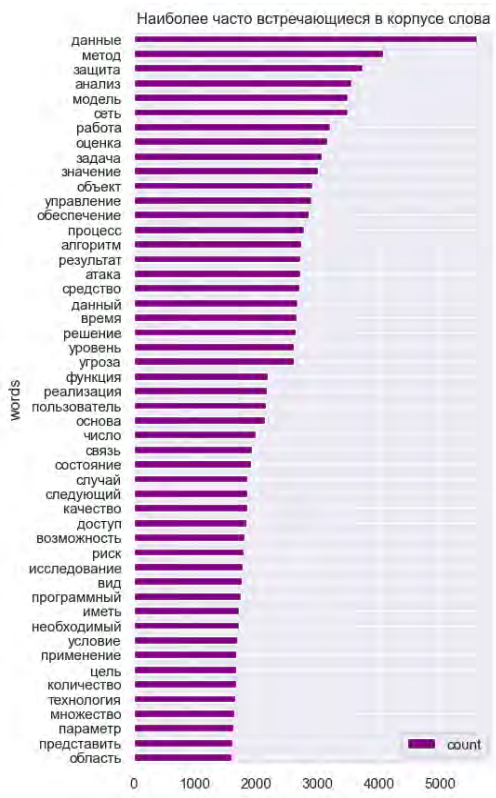
Расширенный стоп-словарь включает 248 токенов и объединяет не только стандартные русскоязычные и англоязычные токены, но и выделенные по результатам анализа собранного корпуса термы: ‘являться’, ‘вопрос’, ‘уязвимость’, ‘использовать’, ‘использование’, ‘нарушитель’, ‘эксплуатация’, ‘рис’, ‘также’, ‘однако’ и т.п. Итеративный анализ частотного словаря позволяет корректировать словарь и повысить общее качество моделей анализа для акцентирования внимания на терминологии и тезаурусе предметной области кибербезопасности.

Пример исходного и нормализованного фрагмента текста после применения описанной схемы обработки приведен в таблице 3.

3.3 Выделение тематических групп в корпусе текстов

Последующий анализ корпуса текстов направлен на решение задачи тематического моделирования, т.е. обнаружение структуры текстовых документов путем выделения и группировки схожих по смыслу текстов и их фрагментов с помощью алгоритмов кластеризации и оценки семантической близости (таблица 4).

а)



б)

№	Ключевое слово	Публикаций
1.	INFORMATION SECURITY	96
2.	ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ	81
3.	КИБЕРБЕЗОПАСНОСТЬ	46
4.	ЗАЩИТА ИНФОРМАЦИИ	41
5.	CYBERSECURITY	37
6.	БЕЗОПАСНОСТЬ ИНФОРМАЦИИ	20
7.	КРИПТОГРАФИЯ	15
8.	MACHINE LEARNING	14
9.	SECURITY	12
10.	КИБЕРПРОСТРАНСТВО	12
11.	МАШИННОЕ ОБУЧЕНИЕ	12
12.	CRYPTOGRAPHY	11
13.	CISSP	10
14.	CLOUD COMPUTING	10
15.	CYBERSPACE	10
16.	HASH FUNCTION	10
17.	VULNERABILITY	10

Рис. 3. а) диаграмма встречаемости слов в корпусе текстов; б) ключевые слова по частоте встречаемости на основе данных портала elibrary.ru

Таблица 3

Фрагмент исходного и нормализованного текста

Исходный текст	Нормализованный текст
Основной задачей настоящей статьи является определение основных угроз безопасности информации, содержащейся в информационных системах, связанных с руткитами. Указанные угрозы определяются впервые. Для решения указанной задачи описываются основные методы реализации главной функции руткитов, перечисляются их дополнительные функциональные возможности, рассматриваются способы их распространения и определяются уязвимости программного обеспечения, эксплуатируемые для внедрения руткитов.	основной задача настоящий статья являться определение основный угроза безопасность информация содержаться информационный система связать руткит', 'указанный угроза определяться впервые', 'решение указать задача описываться основной метод реализация главный функция руткит перечисляться дополнительный функциональный возможность рассматриваться способ распространение определяться уязвимость программный обеспечение эксплуатировать внедрение руткит'

Таблица 4

Основные шаги алгоритмов тематического моделирования

№	Шаг алгоритма	Модель (инструмент)	Примечания
1	Предварительная обработка и нормализация текстов	конвейер обработки текстовых данных	см. п. 3.2
2	Извлечение признаков из текста (формализация)	TF-IDF – построение взвешенного частотного словаря (TF) и разреженной матрицы встречаемости слов с последующим взвешиванием с помощью вектора обратной частоты документа (IDF) – матрица «термы-документ»	-
		хеширование вхождений слов в пространстве фиксированного размера, нормализация векторов слов с помощью L2-нормы	
		предобученная дистрибутивно-семантическая языковая модель FastText*– вектор предложения строится путем нормализации и усреднения векторов токенов	[17]
		предобученная модель-трансформер SBERT – вектор предложения и/или документа строится путем усреднения выходного слоя модели	[18] – нормализация не выполняется
3	Понижение размерности признакового пространства	сингулярное разложение матрицы (NMF) «термы-документ» TF-IDF согласно алгоритму скрытого семантического анализа (LSA)	[19, 20]
4	Тематическое моделирование	кластеризация документов с помощью алгоритма K-Means с оценкой оптимального количества кластеров с помощью метрики «ширина силуэта»	[21, 22]
		неотрицательная матричная факторизация (пакетная реализация с различными целевыми функциями: нормой Фробениуса и обобщенной дивергенцией Кульбака-Лейблера)	-
		скрытое распределение Дирихле (LDA)	[23]
		представление на уровне документа, на уровне n-грамм ключевых слов и словосочетаний с помощью предобученной модели-трансформера на основе оценки семантической близости их векторов вложений	- модель Top2Vec [24] - модель KeyBert** - модель BERTopic [25]

* Repository of pretrained models. URL: http://docs.deeppavlov.ai/en/master/intro/pretrained_vectors.html.

** MaartenGr/KeyBERT. URL: https://zenodo.org/record/4461265

Таблица 5

Схема кластеризации корпуса текстов

Исходное признаковое пространство	Алгоритм понижения размерности	Алгоритм кластеризации	Оценка коэффициента «ширина силуэта» для выделенной структуры кластеров («больше – лучше»)
TF-IDF	-	Алгоритм k-средних (k-means, 5 проходов, 1000 итераций, 10 центров)	0,009 ± 0,012
	LSA		0,064 ± 0,007
		Алгоритм k-средних для наборов (MiniBatchKMeans, 10 центров)	0,034 ± 0,028
Хешированное векторное представление	LSA	Алгоритм k-средних (k-means, 5 проходов, 1000 итераций, 10 центров)	0,059 ± 0,004
		Алгоритм k-средних для наборов (MiniBatchKMeans, 10 центров)	0,010 ± 0,021

Тематическое моделирование на основе кластеризации документов с предварительным понижением размерности признакового пространства (таблица 5). Понижение размерности признакового пространства выполнено с помощью сингулярного разложения матрицы «термы-документы» TF-IDF согласно алгоритму латентно-семантического анализа (LSA).

Как показали эксперименты, кластеризация в LSA-представлении матрицы признаков выполняется значительно быстрее ввиду существенно меньшего их количества, а оценка качества кластеризации, наоборот, улучшилась. При инверсии матрицы признаков становится возможным идентификация центров кластеров как списка наиболее важных слов для каждого кластера:

```
Cluster 0: национальный российский
защита год федерация организация
технология международный стандарт
требование
Cluster 1: вероятность злоумышленник
проникновение кабинет взлом канал утечка
функция схема блок
Cluster 2: модель защита атака объект
воздействие управление сеть оценка
угроза процесс
Cluster 3: сеть пользователь код
изображение алгоритм программный
устройство сигнал защита мобильный
Cluster 4: риск оценка угроза управление
ущерб реализация защита нарушение
уровень этап
Cluster 5: ключ подпись схема алгоритм
квантовый шифрование криптографический
шифр функция блочный
Cluster 6: метаданные аис эда элд
документ электронный целостность запись
криптографический рекурсивный
Cluster 7: текст признак корпус словарь
противоправный личностный именной группа
лексический психологический
Cluster 8: полином булев уравнение
число матрица моном комбинаторный
алгебраический линейный нуль
Cluster 9: автомат клеточный ячейка
обобщенный функция состояние
параллельный кольцо граф криптография
```

Автоматически выявленная структура кластеров частично отражает основные тематики, определяемые редакцией журнала. Например: Безопасность в сетях общего доступа, Анализ рисков информационной безопасности, Методы и средства кодирования информации, Методы и средства стеганографии, Методы анализа программ и верификации, Теоретические основы информатики. В то же время достаточно сложно автоматически декомпозировать документы «сложных» тематических групп, таких как *Концептуальные вопросы кибербезопасности, Киберпротивоборство и операции, Безопасность критической инфраструктуры* и т.п.

Проецирование матрицы TF-IDF с установленными метками кластеров с помощью метода главных компонент (PCA) в двумерное признаковое пространство позволяет оценить взаиморасположение и плотность полученного разбиения на кластеры (рис. 4). В качестве наложенных подписей выделены наиболее близкие к центрам кластеров слова (термы).

Анализ выделенной структуры позволяет оценить взаиморасположение тематических кластеров, их плотность и обособленность, а также принять решение о необходимости слияния или, наоборот, разделения тематических групп.

Тематическое моделирование с помощью скрытого распределения Дирихле (LDA) и неотрицательной матричной факторизации (NMF) (пакетная реализация с различными целевыми функциями: нормой Фробениуса и обобщенной дивергенцией Кульбака-Лейблера) позволяет построить альтернативный вариант аддитивных моделей тематической структуры корпуса и сравнить с построенной ранее кластерной моделью.

Анализ выделенных тем (таблица 6) показывает, что пакетная реализация с целевой функцией, определяемой обобщенной дивергенцией Кульбака-Лейблера, позволяет выделить тематики документов с более сбалансированным количеством документов, а при использовании в качестве целевой функции нормы Фробениуса явно выделяет основные ключевые слова для каждой из тематик. Аддитивная модель на основе скрытого распределения Дирихле, напротив, акцентирует внимание на трех основных тематиках.

Тематическое моделирование на основе моделей векторных вложений

Для формирования гетерогенных моделей вложений использованы модель BERTopic (рис. 5) и модель Top2Vec.

В сочетании с моделью вложений на уровне документа предлагается использовать дистрибутивно-семантическую языковую модель FastText [17], которая рассматривает каждое слово как композицию *n*-грамм символов и позволяет строить более качественные

Рис. 4. Визуализация выделенных кластеров и наиболее важных слов для каждого из них

Таблица 6

Сравнение выделенных тематик

Модель Тема	MiniBatchNMF (Frobenius norm)	MiniBatchNMF (Kullback-Leibler norm)	LDA
1	Модель защиты	Модель защиты	Аппаратные и программные закладки
2	Криптография	Алгоритмы криптографии	Цифровая подпись
3	Криптография + управление	Национальная кибербезопасность	Моделирование сетей
4	Оценка рисков	Оценка рисков	Оценка рисков
5	Клеточные автоматы	Моделирование сетей	Модель объекта защиты + угроз
6	Цифровая подпись	Цифровая подпись	Не определена
7	Мобильные устройства	Мобильные и веб-приложения	Модель канала
8	Социальные сети	Социальные сети	Не определена
9	Сертификация специалистов ИБ	Модель доступа + Сертификация специалистов ИБ	Стеганография
10	Стеганография	Стеганография	Хеширование

Рис. 5. Общая схема применения моделей-трансформеров для извлечения ключевых слов

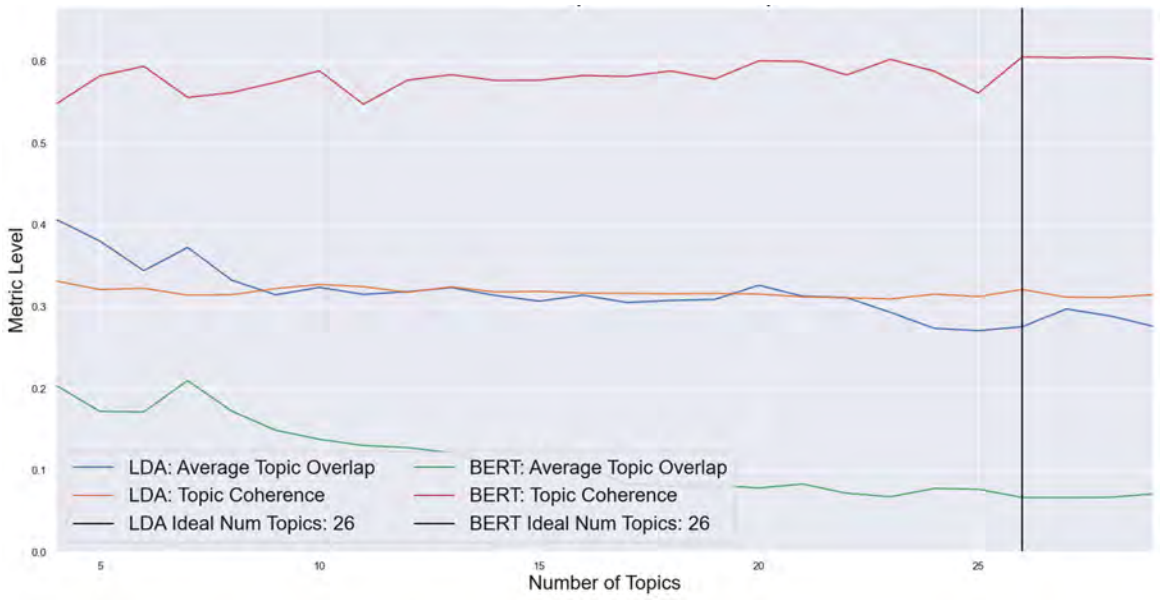


Рис. 6. Коэффициенты согласованности и перекрытия тем для моделей SBERT и LDA в зависимости от количества выбираемых тем (количество тем варьировалось от 4 до 35).

векторные вложения для редких и отсутствующих в обучающем корпусе слов при работе с флективными языками (в подобных языках доминирует словоизменение при помощи флексий – формантов, сочетающих сразу несколько значений)13. Для получения представления на уровне документа модель BERTopic14 формирует вложение (embedding) с помощью предобученного трансформера KeyBERT. Затем строятся вложения слов и слов/словосочетаний в виде n-грамм. Строится набор моделей векторизации, которые извлекают ключевые фразы с шаблонами частей речи из коллекции текстовых документов и преобразуют их в матрицу ключевых фраз документа [27] («ключевая фраза-документ»), позволяют проводить фильтрацию по «стоп-словарю» и адаптивное взвешивание векторов вложений ключевых фраз. Далее оценивается косинусное расстояние для поиска слов и словосочетаний, наиболее семантически близких к документу, понижение размерности с помощью алгоритма UMAP [28] и кластеризации (HDBSCAN) [29]. Метод максимальной граничной значимости (MMR) информационного поиска [30] позволяет повысить разнообразие получаемых ключевых слов, отсортированных в порядке убывания меры близости к исходному документу.

Одной из важнейших задач тематического моделирования является подбор количества выделяемых тем. Для собранного корпуса текстов подбор количества тем выполнен на основе анализа коэффициента согласованности [31] (coherence) и коэффициента перекрытия кластеров для моделей SBERT и LDA (рис. 6).

Автоматическая оценка количества выделяемых тем показала, что для обеих моделей оптимальное количество тем составляет 26. Подобная оценка применима, если предварительно не заданы ключевые слова/словосочетания, используемые в качестве центров предполагаемых тематических групп. Ввиду ограниченного объема анализируемого корпуса такая детализация является излишней, т.к. отдельные темы содержат недостаточное количество документов либо являются слишком близкими по содержанию. При увеличении количества документов автоматический выбор количества тем становится более обоснованным.

3.4 Основные результаты тематического моделирования в выявлении новых направлений исследований и применяемых инструментов в области кибербезопасности

В таблице 7 представлены извлеченные ключевые слова и словосочетания для всего корпуса документов и для подмножеств документов по годам.

С помощью выделенных наиболее релевантных ключевых слов можно проследить динамику распре-

[13] Бабина О.И., Дюмин Н.Ю. Корпусный метод автоматического морфологического анализа флективных языков // Вестник Южно-Уральского государственного университета. Серия: Лингвистика. 2012. №. 25. С. 38-44.

[14] BERTopic. URL: https://maartengr.github.io/BERTopic/index.html

Таблица 7

Ключевые слова и словосочетания для корпуса документов

Год	Ключевые словосочетания	Оценка близости к документам за год
2014	функционирование инфраструктуры управления противника, схема объект воздействия эффекты воздействия технологии защиты, числе применение кибероружия, ряд направлений толкования определения кибербезопасность, ходе исследования перечень угроз, перевод кибероружия,	0,5549 0,5321 0,5181 0,5138 0,5118 0,4893
2015	описание метода криптоанализа, реализации угроз безопасности информации содержание, внимание описание структуры канала передачи данных, основе известности видов кодирования сигналов, хранение базы данных подсистемы кэширования запросов, области безопасности информации взаимодействие,	0,6460 0,6132 0,6081 0,6026 0,5985 0,5972
2016	конкретизация порядка импортозамещения, описание этапов планирования состава средств защиты информации, механизмов реализации концепции, условиях усложнения проблем обеспечения кибербезопасности государства, концепция развития энергетики, решении проблемы импортозамещения,	0,6243 0,5700 0,5332 0,5326 0,5246 0,5198
2017	условии задания критерия вероятность нарушения конфиденциальности информации, случае представления авторами результатов анализа, пример представления вычислений уравнениями подобия, учете влияния множественности центров управления, примеры методики анализа рисков, вероятностей ошибок обоих типов,	0,5924 0,5780 0,5424 0,5338 0,5308 0,5197
2018	расчет коэффициента ущерба объекта информатизации органа, множество оценок эффективности использования средств защиты, пределы значимости уязвимости элементы защиты объекта информатизации, периодичности регистрации пакетов скорость динамики показателя тревоги, зашумление каналов передачи данных сигналами помех, исследовании задача классификации систем управ- ления инцидентами безопасности,	0,5797 0,5782 0,5693 0,5607 0,5525 0,5516
2019	конкретизация моделей угроз безопасности информации, степени соответствия результатов защиты информации цели защиты информации, основе сравнения возможностей реализации угроз, результатом применения модели распределения механизмов защиты, базиса решетки решение задач теории решеток, практике исследования угроз безопасности информации,	0,6401 0,6044 0,5982 0,5907 0,5559 0,5490
2020	разработка методики оценки степени влияния системы защиты, каналы влияния источников угроз безопасности информации, гарантий корректности реализаций механиз- мов обеспечения защиты информации, алгоритмы адаптации интерфейсов пользователя, информации предотвращение передачи инфор- мации предотвращение применения, принцип декомпозиции процесса деятельности,	0,6177 0,6042 0,5974 0,5780 0,5743 0,5742

Год	Ключевые словосочетания	Оценка близости к документам за год
2021	определение матрицы вероятностей перехода, этапе разработка метода сравнения двух битов, уравнений метод решения систем, методах препроцессинга вектора, дефрагментация расшифрование обработка двух заголовков, отношении достижения задач тестирования,	0,6110 0,5977 0,5770 0,5553 0,5450 0,5355
2022	имя способа реализации угрозы копирования информации, угроз безопасности информации сопоставления описаний уязвимостей, выводов результатов криптоанализа схем, статье метод контроля корректности функционирования программ, совершенствование методов определения пригодности криптосистем, суть алгоритма факторизации,	0,6374 0,6344 0,6203 0,6148 0,6079 0,5873

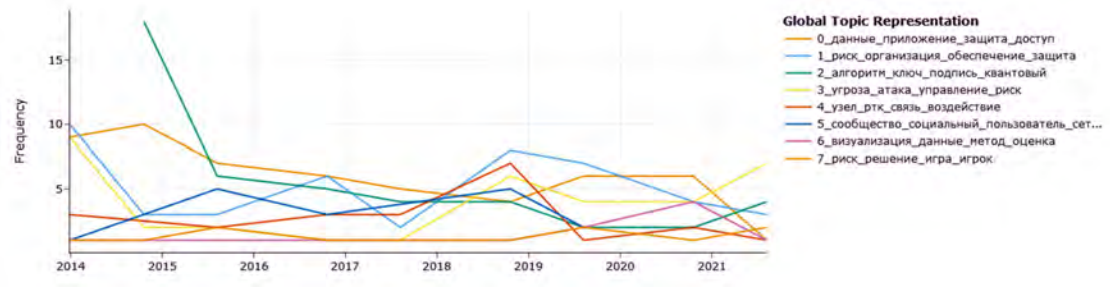


Рис.7. Количество документов в выделенных темах по годам выхода журнала

деления статей по тематикам, отражающим основные тренды в области кибербезопасности, затрагиваемой авторами.

Для выделения ключевых тем использована модель BERTopic, которая с помощью предобученной модели трансформера (построение вложений для отдельных предложений документов с взвешиванием – SBERT) и матрицы «термы-документ» TF-IDF с последующей кластеризацией позволяет формировать компактные кластеры – темы документов. С помощью модели SBERBANK_AI[15] и алгоритма кластеризации HDBSCAN [29] автоматически выделено 22 группы документов с последующим объединением в 8 ключевых тем, построен график зависимости количества документов каждой из тем по годам выхода журнала (рис. 7).

Более информативным является анализ количества публикаций каждой из тематик по годам публикации. Можно отметить преобладание в количественном соотношении документов по тематикам «алгоритм_ключ_подпись_квантовый» («Криптография») и «данные_приложения_защита_доступ» («Обеспечение защиты данных в информационных системах») в период 2013-2016 гг. с последующим выравниванием по количеству публикаций других тематик.

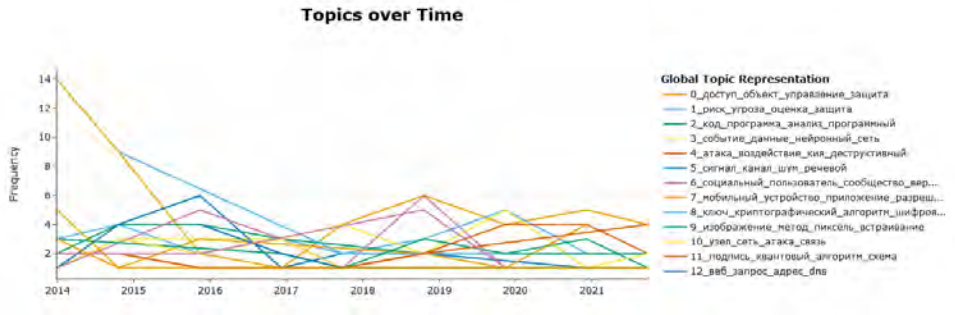
При использовании трансформера SBERT для создания вложений документов с векторизацией по ключевым фразам и модели FastText для создания вложений слов набор тематик более детализирован и включает 19 тем (рис. 8).

Рассмотренные инструменты позволяют отслеживать количество документов определенной тематики и оценивать интересы исследователей к отдельным направлениям и задачам обеспечения кибербезопасности.

Отметим, что тематическое моделирование с использованием инструментов семантического анализа позволяет осуществлять гибкий поиск (семантический поиск) и оценку степени сходства документов с заданной темой (ключевым словом или словосочетанием), а не опираться только на указанные авторами ключевые слова. Например, на рис. 9 показана динамика обсуждения терминов «Машинное обучение» и «Искусственный интеллект» по годам выхода журнала на основе указанных авторами ключевых слов и семан-

15 ruT5, ruRoBERTa, ruBERT: как мы обучили серию моделей для русского языка. URL: https://habr.com/ru/company/sberbank/blog/567776/

а)



б)

Рис. 8. а) Кластерная структура документов, полученная с помощью гетерогенных моделей вложений; б) динамика тематик по годам выхода журнала (количество документов выделенных тематик по годам)

тически близких к заданным терминам документах.

Как видно из рисунка 9, количество близких к названным темам документов существенно больше, чем при использовании строгой фильтрации по ключевым словам, что расширяет возможности поиска и обработки текстовых данных как с учетом возможности варьирования словоформ, так и путем включения семантически близких терминов.

3.5 Суммаризация корпуса текстов

Предложенный алгоритм суммаризации основан на применении комплекса моделей [32] предобработки, нормализации и формализации текстовых документов и включает в себя основные шаги, представленные в таблице 8.

Для выделения наиболее важных предложений, раскрывающих смысловые фрагменты документа, предлагается использовать подходы на основе графового алгоритма Text Rank [34]:

1. Вершины взвешенного неориентированного графа сопоставляются с предложениями текста. Весовые коэффициенты ребер графа отражают степень сходства двух предложений, которая определяется как количество совпадающих в них слов, нормированное их суммарной длиной ($TextRank_1$), либо оценкой семантической близости на основе косинус-меры сходства формальных векторов предложений ($TextRank_1$).

2. С помощью алгоритма PageRank [35] каждой вершине сопоставляется коэффициент, определяемый рекуррентным выражением.

3. После стабилизации коэффициенты вершин упорядочиваются по убыванию значения веса и в реферат включаются первые *n* предложений.

Альтернативным вариантом определения наиболее значимых предложений для включения в реферат является алгоритм кластеризации [36]:

1. Для каждого предложения текстового документа строится вектор формальных признаков.

Рис. 9. Анализ количества документов (по оси абсцисс), включающих термины «Машинное обучение» (МО и МО_СБ) и «Искусственный интеллект» (ИИ и ИИ_СБ) по годам выхода журнала

Таблица 8

Основные шаги алгоритма суммаризации

№	Шаг алгоритма	Модель (инструмент)	Примечания
1	Предварительная обработка и нормализация текстов	Конвейер обработки текстовых данных	см. п.3.2
2	Извлечение признаков из текста (формализация)	TF-IDF – каждое предложение входного текста рассматривается независимо, строится матрица «термы-документ»	-
		предобученная модель FastText – вектор предложения строится путем нормализаци и усреднения векторов токенов	[17]
		предобученная модель RuBERT и SBERT – вектор предложения строится путем усреднения выходного слоя модели	[18] – предобработка, фильтрация и нормализация не выполняется
3	Выделение наиболее важных предложений, раскрывающих смысловые фрагменты документа	$TextRank_1$ – алгоритм PageRank с оценкой близости предложений как количество совпадающих слов в них, нормированное их суммарной длиной	[17]
		$TextRank_2$ – алгоритм PageRank с оценкой семантической близости предложений на основе косинус-меры	[18] формальный вектор предложения строится с помощью модели SBERT
		Кластеризация предложений с помощью алгоритма K-Means с оценкой оптимального количества кластеров с помощью метрики «ширина силуэта»	[18, 21, 22]

№	Шаг алгоритма	Модель (инструмент)	Примечания
4	построение итогового реферата	Выбор наиболее важных узлов графа TextRank	-
		Включение в реферат предложений, наиболее близких к выделенным центроидам кластеров	-
		Генерация текста с помощью модели трансформера T5 на основе выделенных с помощью TextRank наиболее важных узлов графа	[33] Модель-трансформер mT5, обученная на датасете XL-Sum (более 1 млн. аннотированных статей на 44 языках)

Таблица 9

Результаты суммаризации корпуса текстов

Модель	Экстрактивные методы						Абстрактивные методы
	TF-IDF + KMeans	SBERT + KMeans	Fast Text + KMeans	SBERT + Page Rank	Fast Text + Page Rank	Text Rank	Text Rank + T5
Метрика							
BLEU	0,115	0,110	0,069	0,022	0,002	0,082	0,107
ROUGE-1-F	0,250	0,234	0,185	0,092	0,037	0,254	0,257
ROUGE-2-F	0,126	0,110	0,068	0,017	0,003	0,141	0,130
ROUGE-L-F	0,206	0,194	0,143	0,074	0,027	0,194	0,204
Совпадение по леммам	0,356	0,320	0,338	0,354	0,522	0,512	0,501
Совпадение по парам лемм	0,092	0,068	0,091	0,129	0,302	0,202	0,267
Совпадения по тройкам лемм	0,042	0,029	0,047	0,082	0,217	0,116	0,154
Средняя длина	1066,84	1017,57	896,22	457,5	146,51	3043,99	768,28

2. Рациональное количество кластеров определяется по методу «локтя» на основе оценки метрики качества кластеризации «ширина силуэта».

3. Реферат строится из предложений, наиболее близких к центроидам выделенных кластеров.

Вариант алгоритма абстрактивной суммаризации основан на последовательном применении алгоритма $TextRank_2$ с последующим выделенного набора предложений для преобразования «text-to-text» с помощью модели трансформера mT5, обученной на наборе данных XL-Sum (более 1 млн аннотированных статей на 44 языках).

Итоговое сравнение для рассмотренных выше алгоритмов суммаризации представлено в таблице 9.

Согласно полученным оценкам, наилучшими экстрактивными моделями реферирования являются TextRank, PageRank и TF-IDF+KMeans. Комплексная модель на основе трансформера T5 показала по отдельным метрикам сравнимый и в целом превосходящий результат. Однако, как отмечается в [34], автоматические оценки работы алгоритмов реферирования на научных текстах имеют существенные ограничения, необходима экспертная оценка полученных рефератов. В таблице 10 представлены фрагменты рефератов, полученные с помощью каждой из рассмотренных моделей (тексты приведены без корректировки, фильтрации символов и корректировки знаков препинания).

3.6 Анализ полученных результатов

Анализ полученных результатов показал следующее:

1. Формализация текстовых описаний с помощью моделей-трансформеров и SBERT не является удачным вариантом для последующего применения графовых алгоритмов реферирования, но хорошо подходит для тематического моделирования и выделения ключевых слов, словосочетаний и тематик с возможностью оценки семантической близости документов предметной области.

2. Наилучшие результаты реферирования продемонстрировал традиционный графовый алгоритм на основе Text Rank, который не требует сложных лингвистических моделей, знаний предметной области или языка. Сочетание графовых алгоритмов и возможностей построения векторных вложений на уровне документа, предложений, слов и символов позволяет получить гибкий инструмент для выделения ключевой информации из анализируемых документов.

3. Для формализации признаков на уровне предложений и их дальнейшей кластеризации наиболее подходящей оказалась традиционная модель TF-IDF в задаче суммаризации и модель SBERT в задачах тематического моделирования.

4. Ручной анализ рефератов показал, что наилучшие результаты получены с помощью модели TextRank для текстовых документов, посвященных одной тематике. Модель SBERT + KMeans позволяет выделить в тексте несколько тематик и выбрать из них наиболее важные предложения для включения в реферат.

5. Модель формализации FastText продемонстрировала наихудшие результаты по автоматически определяемым метрикам в задаче реферирования, но позволяет повысить качество векторных вложений в сочетании с моделями-трансформерами в задаче тематического моделирования.

6. Для большинства нейросетевых моделей абстрактивного реферирования необходим большой предметно-ориентированный корпус для обучения (технология «переноса обучения»), скромные результаты работы моделей трансформеров «text-to-text» обусловлены, в первую очередь, малым размером собранного корпуса.

4. Заключение

Результаты анализа проблемы тематического моделирования и реферирования (суммаризации) проблемно-ориентированных текстов для специалистов в области кибербезопасности показали следующее:

Таблица 10

Фрагменты созданных рефератов

Модель суммаризации	Фрагмент реферата, созданный моделью («как есть» без корректировки)
TF-IDF + KMeans	*«Главными тенденциями развития угроз являются следующие* : рост числа атак, многие из которых ведут к большим потерям; возрастание сложности атак, которые могут включать несколько этапов и применять специальные методы защиты от возможных методов противодействия; воздействие практически на все электронные (цифровые) устройства, в числе которых в последнее время все большую значимость приобретают мобильные устройства, а они в наибольшей степени подвержены рискам в области информационной безопасности; ...»*
SBERT + KMeans	*«В NUM г. Масштабные нарушения, затрагивающие все стороны жизни общества, в основе которых лежат новейшие методы осуществления атак на компьютерные сети, а также управление общественным сознанием требуют системного подхода к созданию комплексной системы безопасности, способной противостоять этим угрозам. ...»*
Text Rank	*«Главными тенденциями развития угроз являются следующие : рост числа атак , многие из которых ведут к большим потерям ; возрастание сложности атак , которые могут включать несколько этапов и применять специальные методы защиты от возможных методов противодействия ; воздействие практически на все электронные (цифровые) устройства , в числе которых в последнее время все большую значимость приобретают мобильные устройства , а они в наибольшей степени подвержены рискам в области информационной безопасности ; ...»*

* Пробелы перед знаками препинания – это не ошибки авторов статьи. Это фрагменты рефератов, полученные с помощью каждой из рассмотренных моделей.

— применение моделей машинного обучения и интеллектуального анализа позволит на основе тонкой настройки существующих нейросетевых языковых моделей, построенных на обобщенных корпусах текстов, в сочетании с классическими графовыми моделями, существенно повысить качество анализа исходных документов и снизить нагрузку на эксперта, в профессиональные обязанности которого входит анализ текстовых документов из различных источников в виде «сырых» слабоструктурированных данных, за счет их реферирования и автоматического выделения тем;
— для русскоязычного сегмента сегодня представлены лишь отдельные, достаточно скромные по размеру корпусы проблемно-ориентированных текстов по тематике кибербезопасности, что объясняется высокой трудоемкостью их сбора и отсутствием общепринятого протокола разметки, учитывающего структуру и номенклатуру отечественной документации;
— имеет место недостаточное количество полилингвальных языковых предметно-ориентированных моделей семейства BERT, T5 и др.;
— автоматически выявленная структура тематик (кластеров) частично отражает основные тематики, определяемые редакцией журнала, и является перспективным инструментом для анализа специализированных документов предметной области.
— тематическое моделирование с помощью гетерогенных моделей вложений позволяет выделять основные тематики документов и оценивать интересы исследователей к отдельным направлениям и задачам обеспечения кибербезопасности, расширяя возможности интеллектуального поиска и обработки текстовых данных с учетом семантики.

Исследование выполнено при финансовой поддержке Гранта ИБ МТУСИ, предоставленного Минцифры России, Соглашение № 40469-18/2022-к и финансовой поддержке РФФИ в рамках научного проекта № 20-08-00668.

Литература

1. Liu X., Xiong H., Shen N. A hybrid model of VSM and LDA for text clusteing // 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA). IEEE, 2017, pp. 230-233.
2. Gambhir M., Gupta V. Recent automatic text summarization techniques: a survey // Artificial Intelligence Review. 2017, vol. 47, no. 1, pp. 1-66. DOI:10.1007/s10462-016-9475-9
3. Белякова А.Ю., Беляков Ю.Д. Обзор задачи автоматической суммаризации текста // Инженерный вестник Дона. 2020. № 10 (70). С. 142-159.
4. Sri S.H.B., Dutta S.R. A Survey on Automatic Text Summarization Techniques // Journal of Physics: Conference Series. IOP Publishing, 2021, vol. 2040, no. 1, pp. 012044. DOI: 10.1088/1742-6596/2040/1/012044
5. Liang Z. et al. Gated graph neural attention networks for abstractive summarization // Neurocomputing. 2021, vol. 431, pp. 128-136.
6. Masum A.K.M. et al. Abstractive method of text summarization with sequence to sequence RNNs // 2019 10th international conference on computing, communication and networking technologies (ICCCNT). IEEE, 2019, pp. 1-5.
7. Vaswani A. et al. Attantion is All You Need // 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. 2017, vol. 30, pp. 1-11.
8. Jonsson F. Evaluation of the Transformer Model for Abstractive Text Summarization: Degree Project in Computer Science and Engineering. Master's in computer science dissertation. Stockholm, Sweden. 2019. URL: https://www.diva-portal.org/smash/get/diva2:1368180/FULLTEXT01.pdf (дата обращения: 28.10.2022).
9. Gupta A. et al. Automated news summarization using transformers // Sustainable Advanced Computing. Springer, Singapore, 2022. pp. 249-259. DOI: 10.1007/978-981-16-9012-9_21
10. Jatnika D., Bijaksana M.A., Suryani A.A. Word2vec model analysis for semantic similarities in english words // Procedia Computer Science. 2019, vol. 157, pp. 160-167.
11. Yang M. et al. A hierarchical clustering approach to fuzzy semantic representation of rare words in neural machine translation // IEEE Transactions on Fuzzy Systems. 2020, vol. 28, no. 5, pp. 992-1002.
12. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019, pp. 3982-3992. DOI:10.18653/v1/D19-1410
13. Arora S., Hu W., Kothari P.K. An analysis of the t-sne algorithm for data visualization // Conference on learning theory. PMLR, 2018, pp. 1455-1462.
14. Dey A., Jenamani M., Thakkar J.J. Senti-N-Gram: An n-gram lexicon for sentiment analysis // Expert Systems with Applications. 2018, vol. 103, pp. 92-105.
15. Reiter E. A structured review of the validity of BLEU // Computational Linguistics. 2018, vol. 44, no. 3, pp. 393-401.
16. Васильев В.И., Вульфин А.М., Кучкарова Н.В. Оценка актуальных угроз безопасности информации с помощью технологии трансформеров // Вопросы кибербезопасности. 2022. № 2(48). С. 27-38. DOI 10.21681/2311-3456-2022-2-27-38

17. Bojanowski P. et al. Enriching Word Vectors with Subword Information // Transactions of the association for computational linguistics. 2017, vol. 5, pp. 135-146.
18. Miller D. Leveraging BERT for Extractive Text Summarization on Lectures // arXiv preprint arXiv:1906.04165. 2019. doi.org/10.48550/arXiv.1906.04165
19. Lee D.D., Seung H.S. Learning the Parts of Objects by Non-Negative Matrix Factorization // Nature. 1999, vol. 401, no. 6755, pp.788-791. DOI: 10.1038/44565
20. Williams T., Betak J. A Comparison of LSA and LDA for the Analysis of Railroad Accident Text // Procedia computer science. 2018, vol. 130, pp. 98-102.
21. See A., Liu P.J., Manning C.D. Get to the Point: Summarization with Pointer-Generator Networks // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017, pp. 1073-1083.
22. Dinh D.T., Fujinami T., Huynh V.N. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient // Knowledge and Systems Sciences: 20th International Symposium, Da Nang, Vietnam, November 29–December 1, 2019. Springer Singapore, 2019, pp. 1-17.
23. Jelodar H. et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey // Multimedia Tools and Applications. 2019, vol. 78, pp. 15169-15211.
24. Angelov D. Top2vec: Distributed representations of topics // arXiv preprint arXiv:2008.09470. 2020. doi.org/10.48550/arXiv.2008.09470
25. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure // arXiv preprint arXiv:2203.05794. 2022. doi.org/10.48550/arXiv.2203.05794
26. Шереметьева С.О., Бабина О.И. Платформа для концептуального аннотирования многоязычных текстов // Вестник Южно-Уральского государственного университета. Серия: Лингвистика. – 2020. Т. 17. №. 4. С. 53-60.
27. Schopf T., Klimek S., Matthes F. PatternRank: Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase // Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management KDIR. 2022, pp. 243-248. DOI:10.20944/PREPRINTS201908.0073.V1
28. McInnes L., Healy J., Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction // The Journal of Open Source Software. 2018, vol. 3, no. 29, pp. 861. DOI: 10.21105/joss.00861
29. McInnes L., Healy J., Astels S. hdbscan: Hierarchical Density Based Clustering // J. Open Source Softw. 2017, vol. 2, no. 11, pp. 205. DOI:10.21105/JOSS.00205
30. Carbonell J., Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries // Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998. pp. 335-336. DOI:10.1145/290941.291025
31. Краснов Ф.В., Баскакова Е.Н., Смазневич И.С. Оценка прикладного качества тематических моделей для задач кластеризации // Вестник ТГУ. УВТиИ. 2021. № 56. С. 100-111. DOI: 10.17223/ 19988605/56/11
32. Gusev I. Dataset for Automatic Summarization of Russian News // Conference on Artificial Intelligence and Natural Language. Springer, Cham, 2020. pp. 122-134. DOI:10.1007/978-3-030-59082-6_9
33. Hasan T. et al. XL-Sum: large-scale multilingual abstractive summarization for 44 languages // Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing 2021. Association for Computational Linguistics (ACL), 2021, pp. 4693-4703. DOI:10.18653/v1/2021.findings-acl.413
34. Mihalcea R., Tarau P. TextRank: Bringing Order into Texts // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004. pp. 404-411.
35. Bar-Yossef Z., Mashiach L.T. Local Approximation of Pagerank and Reverse Pagerank // Proceedings of the 17th ACM conference on Information and knowledge management. 2008, pp. 279-288. DOI:10.1145/1458082.1458122
36. García-Hernández R. A. et al. Text Summarization by Sentence Extraction Using Unsupervised Learning // Mexican International Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2008, pp. 133-143. DOI:10.1007/978-3-540-88636-5_12

THEMATIC MODELING AND SUMMARIZATION OF TEXTS IN THE FIELD OF CYBER SECURITY (BASED ON JOURNAL PUBLICATIONS)

Vasilyev V.I.[16], Vulfin A.M.[17], Kuchkarova N.V.[18]

Purpose: *improving the quality of text document analysis through the use of machine learning and intellectual analysis models in summarizing and topic modeling tasks, which will reduce the load on an expert who analyzes*

16 Vladimir I. Vasilyev, Dr.Sc.(Eng.), Professor, Ufa University of Science and Technology, Ufa, Russia, E-mail: vasilyev@ugatu.ac.ru
17 Alexey M. Vulfin, Ph.D., Associate Professor, Ufa University of Science and Technology, Ufa, Russia, E-mail: vulfin.am@ugatu.su
18 Nailya V. Kuchkarova, M.Sc., Senior Lecturer, Ufa University of Science and Technology, Ufa, Russia, E-mail: kuchkarova.nv@ugatu.su

and generalizes significant volumes of semi-structured text data on information security topics from various sources.

Methods: *machine learning methods were used for the operational processing and analysis of large volumes of heterogeneous ill-structured information in natural language (NL). Methods of thematic modeling and summarization of texts based on deep neural networks, including neural network language models based on the architecture of transformers, are applied.*

Practical relevance: *the main stages of the machine procedure of thematic modeling and summarization of professional texts in the field of information security are highlighted. The results of a comparative evaluation of the effectiveness of using clustering models, latent semantic analysis, Fast Text, Text Rank language models and BERT transformers for these purposes are presented. Recommendations are given regarding the prospects for the practical application of these models as a means of intellectual support for the professional activities of cybersecurity specialists.*

Scientific novelty: *a complex of machine learning models for thematic modeling and summarization of professional texts is proposed, based on neural network models of attachments and transformer models, characterized by an algorithm for preparing a corpus of texts for training models and the use of a learning transfer algorithm, which will increase the efficiency of analysis and generalization of domain-specific corpora of texts.*

Keywords: *Text Mining, vector word embedding, transformer, text clustering, summarization, information security, cybersecurity.*

References

1. Liu X., Xiong H., Shen N. A hybrid model of VSM and LDA for text clusteing // 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA). IEEE, 2017, pp. 230-233.
2. Gambhir M., Gupta V. Recent automatic text summarization techniques: a survey // Artificial Intelligence Review. 2017, vol. 47, no. 1, pp. 1-66. DOI:10.1007/s10462-016-9475-9
3. Belyakova A.Yu., Belyakov Yu.D. Obzor zadachi avtomaticheskoj summarizacii teksta // Inzhenerny`j vestnik Dona. 2020, no. 10 (70), pp. 142-159
4. Sri S.H.B., Dutta S.R. A Survey on Automatic Text Summarization Techniques // Journal of Physics: Conference Series. IOP Publishing, 2021, vol. 2040, no. 1, pp. 012044. DOI: 10.1088/1742-6596/2040/1/012044
5. Liang Z. et al. Gated graph neural attention networks for abstractive summarization // Neurocomputing. 2021, vol. 431, pp. 128-136.
6. Masum A.K.M. et al. Abstractive method of text summarization with sequence to sequence RNNs // 2019 10th international conference on computing, communication and networking technologies (ICCCNT). IEEE, 2019, pp. 1-5.
7. Vaswani A. et al. Attantion is All You Need // 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. 2017, vol. 30, pp. 1-11.
8. Jonsson F. Evaluation of the Transformer Model for Abstractive Text Summarization: Degree Project in Computer Science and Engineering. Master's in Computer Science dissertation. Stockholm, Sweden. 2019. URL: https://www.diva-portal.org/smash/get/diva2:1368180/FULLTEXT01.pdf (дата обращения: 28.10.2022).
9. Gupta A. et al. Automated news summarization using transformers // Sustainable Advanced Computing. Springer, Singapore, 2022. pp. 249-259. DOI: 10.1007/978-981-16-9012-9_21
10. Jatnika D., Bijaksana M.A., Suryani A.A. Word2vec model analysis for semantic similarities in english words // Procedia Computer Science. 2019, vol. 157, pp. 160-167.
11. Yang M. et al. A hierarchical clustering approach to fuzzy semantic representation of rare words in neural machine translation // IEEE Transactions on Fuzzy Systems. 2020, vol. 28, no. 5, pp. 992-1002.
12. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019, pp. 3982-3992. DOI:10.18653/v1/D19-1410
13. Arora S., Hu W., Kothari P.K. An analysis of the t-sne algorithm for data visualization // Conference on learning theory. PMLR, 2018, pp. 1455-1462.
14. Dey A., Jenamani M., Thakkar J.J. Senti-N-Gram: An n-gram lexicon for sentiment analysis // Expert Systems with Applications. 2018, vol. 103, pp. 92-105.
15. Reiter E. A structured review of the validity of BLEU // Computational Linguistics. 2018, vol. 44, no. 3, pp. 393-401.
16. Vasilyev, V.I., Vulfin A.M., Kuchkarova N.V. Ocenka aktualnyx ugroz bezopasnosti informacii s pomoshhyu texnologii transformerov // Voprosy kiberbezopasnosti. 2022, no. 2(48), pp. 27-38. DOI 10.21681/2311-3456-2022-2-27-38.
17. Bojanowski P. et al. Enriching Word Vectors with Subword Information // Transactions of the association for computational linguistics. 2017, vol. 5, pp. 135-146.
18. Miller D. Leveraging BERT for Extractive Text Summarization on Lectures // arXiv preprint arXiv:1906.04165. 2019. doi.org/10.48550/arXiv.1906.04165
19. Lee D.D., Seung H.S. Learning the Parts of Objects by Non-Negative Matrix Factorization // Nature. 1999, vol. 401, no. 6755, pp.788-791. DOI: 10.1038/44565

20. Williams T., Betak J. A Comparison of LSA and LDA for the Analysis of Railroad Accident Text // Procedia computer science. 2018, vol. 130, pp. 98-102.
21. See A., Liu P.J., Manning C.D. Get to the Point: Summarization with Pointer-Generator Networks // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017, pp. 1073-1083.
22. Dinh D.T., Fujinami T., Huynh V.N. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient // Knowledge and Systems Sciences: 20th International Symposium, Da Nang, Vietnam, November 29–December 1, 2019. Springer Singapore, 2019, pp. 1-17.
23. Jelodar H. et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey // Multimedia Tools and Applications. 2019, vol. 78, pp. 15169-15211.
24. Angelov D. Top2vec: Distributed representations of topics // arXiv preprint arXiv:2008.09470. 2020. doi.org/10.48550/arXiv.2008.09470
25. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure // arXiv preprint arXiv:2203.05794. 2022. doi.org/10.48550/arXiv.2203.05794
26. Sheremetyeva S.O., Babina O.I. A platform for knowledge assisted conceptual annotation of multilingual texts // Vestnik YUzhno-Uralskogo gosudarstvennogo universiteta. Seriya: Lingvistika. 2020, vol. 17, no. 4, pp. 53-60.
27. Schopf T., Klimek S., Matthes F. PatternRank: Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase // Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management KDIR. 2022, pp. 243-248. DOI:10.20944/PREPRINTS201908.0073.V1
28. McInnes L., Healy J., Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction // The Journal of Open Source Software. 2018, vol. 3, no. 29, pp. 861. DOI: 10.21105/joss.00861
29. McInnes L., Healy J., Astels S. hdbscan: Hierarchical Density Based Clustering // J. Open Source Softw. 2017, vol. 2, no. 11, pp. 205. DOI:10.21105/JOSS.00205
30. Carbonell J., Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries // Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998. pp. 335-336. DOI:10.1145/290941.291025
31. Krasnov F.V., Baskakova E.N., Smaznevich I.S. Ocenka prikladnogo kachestva tematicheskix modelej dlya zadach klasterizacii // Vestnik TGU. UVTil. 2021, no.5, pp. 100-111. DOI: 10.17223/ 19988605/56/11
32. Gusev I. Dataset for Automatic Summarization of Russian News // Conference on Artificial Intelligence and Natural Language. Springer, Cham, 2020. pp. 122-134. DOI:10.1007/978-3-030-59082-6_9
33. Hasan T. et al. XL-Sum: large-scale multilingual abstractive summarization for 44 languages // Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing 2021. Association for Computational Linguistics (ACL), 2021, pp. 4693-4703. DOI:10.18653/v1/2021.findings-acl.413
34. Mihalcea R., Tarau P. TextRank: Bringing Order into Texts // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004. pp. 404-411.
35. Bar-Yossef Z., Mashiach L.T. Local Approximation of Pagerank and Reverse Pagerank // Proceedings of the 17th ACM conference on Information and knowledge management. 2008, pp. 279-288. DOI:10.1145/1458082.1458122
36. García-Hernández R. A. et al. Text Summarization by Sentence Extraction Using Unsupervised Learning // Mexican International Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2008, pp. 133-143. DOI:10.1007/978-3-540-88636-5_12