Результаты автоматического интеллектуального анализа отдельных полей реестра операторов персональных данных

П.Ю. Пушкин, А.М. Русаков

***Аннотация*—В работе представлены результаты интеллектуального анализа записей, содержащихся в полях реестра операторов персональных данных «перечень действий с персональными данными» и «срок или условие прекращения обработки персональных данных» и оценка их соответствия требованиям законодательства о персональных данных. В качестве исследуемого операторского сообщества выбраны высшие учебные заведения, что позволяет учесть схожие особенности обработки персональных данных при формировании экспертных оценок и интеллектуальном анализе данных. Для цели исследования сформирован корпус текстов, который возможно использовать для анализа методов интеллектуального анализа данных по тематикам обработки и защиты информации. Для интеллектуального анализа текста применялись библиотеки: Scikit-learn, Gensim, PyMystem3, FuzzyWuzzy. Поисковые запросы выполнялись с учетом словарей синонимов и нечеткого расположения слов. Для поиска устойчивых ключевых сочетаний слов вычислялась весовая функция TF-IDF. Проведено сравнение методов лемматизации слов для целей исследования. Полученные результаты показывают верность экспертных оценок по заполнению полей реестра операторов персональных данных: максимальный кластер, определенный по результатам интеллектуального анализа, соответствует экспертному шаблону. Результаты автоматического интеллектуального анализа требуют верификации эксперта в области обработки и защиты персональных данных. Использование методов интеллектуального анализа данных позволяет существенно повысить эффективность деятельности экспертов при работе с большими объемами информации, содержащейся в реестре операторов персональных данных. Работа направлена на формирование отдельных разделов рекомендаций по разработке отраслевого (в сфере высшего образования и науки) кодекса поведения в области защиты прав субъектов персональных данных в целях повышения уровня защищенности такой информации.**

***Ключевые слова*—Персональные данные, защита персональных данных, информационно-аналитические системы, интеллектуальный анализ данных.**

I. Введение

В целях государственного контроля за деятельностью операторов, осуществляющих обработку персональных данных, Роскомнадзор за 2019 год по результатам 804 плановых проверок выдал 799 предписаний об устранении 2580 выявленных нарушений обязательных требований [1]. Систематическими нарушениями, выявленными при проведении плановых проверок, стали:

- представление в уполномоченный орган уведомления об обработке персональных данных (далее – уведомление), содержащего неполные и (или) недостоверные сведения, – 427 нарушений (17% от общего количества нарушений);
- непринятие оператором мер, необходимых и достаточных для обеспечения выполнения обязанностей, предусмотренных Федеральным законом «О персональных данных» и принятыми в соответствии с ним нормативными правовыми актами, – 290 нарушений (11%);
- непредставление в уполномоченный орган сведений о прекращении обработки персональных данных или об изменении информации, содержащейся в уведомлении об обработке персональных данных, – 257 (10%);
- несоблюдение оператором требований по информированию лиц, осуществляющих обработку персональных данных без использования средств автоматизации, – 184 нарушения (7%);
- обработка персональных данных в случаях, не предусмотренных Федеральным законом «О персональных данных», – 121 нарушение (5%).

Из приведенной в отчете [1] статистики следует, что мероприятия по разработке и своевременной подаче уведомления об обработке персональных данных [2] (далее – уведомление) является самым распространённым нарушением в практике Роскомнадзора. Кроме того, непредоставление в Роскомнадзор требуемых сведений, также связано с несоблюдением установленной федеральным законом [3] процедуры уведомления уполномоченного органа по защите прав субъектов персональных данных.

Описанные в отчете [1] проблемные вопросы, могут быть связаны с недостаточной компетенцией операторов в области обработки и защиты персональных данных. Процесс составления уведомления требует от операторов правовых, организационных и технических навыков деятельности в указанной сфере: знания нормативно-методической базы, методик определения информационных потоков в организации с учетом

Статья получена 23 декабря 2020.

Павел Юрьевич Пушкин, MIREA - Russian Technological University, Moscow, Russia (e-mail: is-irk@mail.ru).

Алексей Михайлович Русаков, MIREA - Russian Technological University, Moscow, Russia (e-mail: rusal@bk.ru).

специфики отрасли, применяемых для обработки персональных данных информационных систем и технологий, методов и средств защиты информации, организации делопроизводства и документооборота. Поэтому разработка отраслевых методических рекомендаций для операторских сообществ по организации и обработке персональных данных, безусловно, сможет повысить уровень компетенции сотрудников операторов и, следовательно, общий уровень защиты персональных данных в организациях. Разработка отраслевых кодексов поведения в области защиты прав субъектов персональных данных для оказания методической помощи членам операторских сообществ рекомендована рабочей группой, состоящей из членов Консультативного совета при Уполномоченном органе по защите прав субъектов персональных данных [4].

Таким образом, вопросы, связанные с подготовкой уведомления и дальнейшим взаимодействием операторов персональных данных (далее - операторы) с регулятором, являются актуальными как для самих операторов, так и для уполномоченного органа по защите прав субъектов персональных данных и характеризуют состояние работ по организации и защите персональных данных.

Сведения из поданных операторами уведомлений, после проверки заносятся Роскомнадзором в Реестр операторов, осуществляющих обработку персональных данных (далее – Реестр) [5]. Отдельные поля Реестра являются общедоступными, что позволяет провести анализ представленных в них данных на соответствие требованиям законодательства о персональных данных и их защите, а также сформировать рекомендации по их заполнению для конкретных операторских сообществ.

Целью данной работы является исследование размещенных в Реестре сведений на основе интеллектуального анализа данных для формирования отдельных разделов рекомендаций университетскому операторскому сообществу по их заполнению.

Для достижения указанной цели исследования в работе решались следующие задачи:

- формирование набора данных (корпуса текстов) по тематике в сфере обработки и защиты данных на основе Реестра для проведения исследований с использованием интеллектуальных методов анализа текста;
- оценка возможности автоматического формирования «эталонного» (шаблонного) заполнения отдельных полей Реестра на основе интеллектуального анализа данных из всей исследуемой выборки и последующего его сравнения с полученными экспертными оценками;
- реализация и оценка интеллектуального анализа данных, содержащихся в отдельных полях Реестра, с использованием экспертных данных и оценок.

Научная новизна работы заключается в проведении оценки использования методов интеллектуального анализа данных для решения задачи разработки рекомендаций по заполнению обязательных организационно-распорядительных и отчетных документов в сфере обработки и защиты персональных данных.

Практическая значимость работы заключается в определении требуемого в соответствие с законодательством шаблонного заполнения отдельных полей Реестра для университетского операторского сообщества, возможности использования разработанных алгоритмов и программных модулей для формирования отраслевых рекомендаций по обработке и защите персональных данных, оценке операторскими сообществами соблюдения требований и условий обработки и защиты персональных данных с использованием интеллектуальных методов анализа сведений, размещенных в Реестре, оценке текущей активности вузов по подаче уведомлений, а следовательно, и по организации работ по обработке и защите персональных данных.

II. Методика проведения исследования

В качестве исследуемого операторского сообщества выбраны высшие учебные заведения, имеющие схожие условия обработки персональных данных и требования к их защите. К общим отличительным признакам университетского операторского сообщества можно отнести:

- наличие несовершеннолетних и совершеннолетних субъектов персональных данных;
- наличие общей нормативно-правовой базы, определяющей цели и задачи обработки персональных данных;
- смешанную обработку персональных данных;
- схожие технологические процессы обработки, категории субъектов, категории и объем обрабатываемых персональных данных;
- использование электронной образовательной среды;
- подключение к федеральным государственным информационным системам для передачи персональных данных;
- передача данных обучающихся и сотрудников за пределы Российской Федерации для участия в международных мероприятиях, а также реализация международного обучения (трансграничная передача).

Данные отличительные особенности позволяют сформировать и проверить отдельные записи в Реестре, заполнение которых с большей степенью вероятности должно совпадать для всего операторского сообщества. В качестве таких полей для целей настоящего исследования будут использоваться «Перечень действий с персональными данными» и «Срок или условие прекращения обработки персональных данных».

Поле «Перечень действий с персональными данными», представленное в Реестре, формируется из сведений, указанных в уведомлении в графе «Перечень действий с персональными данными, общее описание используемых оператором способов обработки персональных данных» [2]. Возможные способы обработки однозначно определены в электронной форме уведомления на портале Роскомнадзора [6], а именно:

автоматизированная, неавтоматизированная, смешанная, без передачи по внутренней сети юридического лица, с передачей по внутренней сети юридического лица, без передачи по сети Интернет, с передачей по сети Интернет. Перечень возможных действий с персональными данными определен в Федеральном законе [3] и включает: сбор, запись, систематизацию, накопление, хранение, уточнение (обновление, изменение), извлечение, использование, передачу (распространение, предоставление, доступ), обезличивание, блокирование, удаление, уничтожение персональных данных. Очевидно, что в высшем учебном заведении с персональными данными осуществляют все, указанные в [3] действия, или такие действия должны быть осуществлены при определенных законом условиях. Исключением может являться обезличивание данных, правила проведения которого определены не во всех ВУЗах, но в той или иной степени все-равно используются в деятельности.

Для проведения исследования Реестра использовалась разработанная авторами информационно-аналитическая система мониторинга выполнения операторами персональных данных требований законодательства (далее – информационно-аналитическая система) [7].

Общая методика проведения исследования отдельных полей реестра операторов персональных данных с использованием информационно-аналитической системы представлена на рисунке 1.

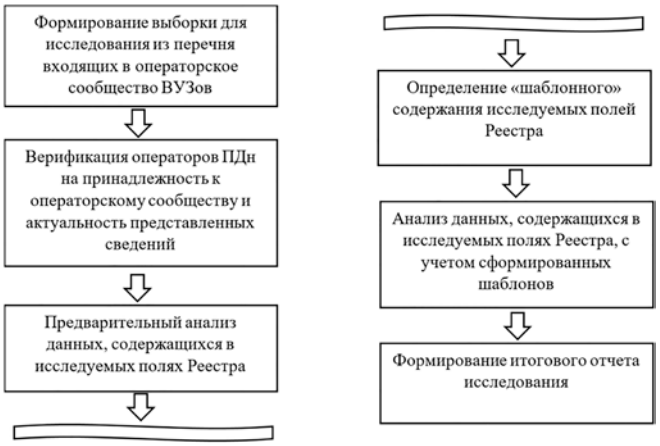

Рис. 1. Методика проведения исследования отдельных полей Реестра

III. Формирование выборки ВУЗов для исследования

Для формирования выборки ВУЗов использовались открытые официальные источники информации в сети Интернет. Получить данные от таких источников возможно следующими способами [8]:

- загрузить данные в машиночитаемом формате с официальных ресурсов открытых данных;
- подключиться к каталогу данных и вручную сконвертировать в нужный формат;
- воспользоваться программными средствами для автоматического извлечения информации с сайта;
- использовать внутренний программный интерфейс (API) каталога данных для выгрузки данных.

Для формирования перечня актуальных данных о ВУЗах России авторами проведен анализ ряда открытых официальных источников информации. Изучены сведения, размещенные на портале открытых данных Российской Федерации https://data.gov.ru/ [9]. Данный портал включает в себя список образовательных учреждений в едином формате, но данные представлены не по всем субъектам Российской федерации. Каждый субъект РФ представлен отдельным источником данных, а дата обновления самого свежего набора данных для региона Москва — 19.02.2020. Таким образом использование только портала открытых данных РФ не позволит сформировать полный и актуальный перечень сведений о ВУЗах.

В связи с чем авторами, в целях обеспечения актуального перечня ВУЗов РФ, принято решение об использовании наборов данных с сайта: «Реестр организаций, осуществляющих образовательную деятельность по имеющим государственную аккредитацию образовательным программам» [10]. Реестр образовательных организаций имеет закрытый программный интерфейс. Для доступа к данным реестра разработан модуль автоматического получения информации с веб-сайта. В реестре образовательных организаций [10] обнаружены проблемы, связанные с неверной классификацией запросов. При отправке статуса лицензии некоторые пункты визуально выглядят одинаково, но дают разную выборку. Например, если открыть вкладку «Текущий статус свидетельства» реестра образовательных организаций [10], то там есть сразу несколько одинаковых параметров, как показано на рис. 2.

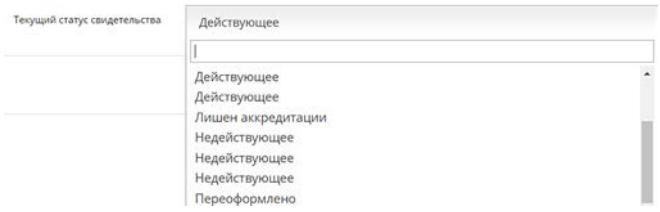

Рис. 2. Скриншот фрагмента реестра образовательных организаций [10]

На рисунке 2. Представлен скриншот фрагмента реестра образовательных организаций [10], на котором отображены два одинаковых слова «Действующее» и три визуально одинаковых слова «Недействующее» Данная проблема решена, путем отсылки всех некорректных запросов по-отдельности.

Для автоматической отправки запросов разработан программный модуль, направляющий запросы в формате JSON и осуществляющий парсинг ответов в режиме реального времени с разделением на отельные потоки для повышения производительности. Разработанный программный модуль написан на «Python» с использованием библиотек «BeautifulSoup», «Fake_Useragent», «Requests». Полученные объекты

выборки загружались в базу данных «mongoDB». В результате получилась выборка из 808 ВУЗов. При дальнейшем анализе обнаружилось дублирование данных: один и тот же ВУЗ может иметь несколько действующих лицензий.

После группировки лицензий по ВУЗам выявлены 79 продублированных записей. В итоге получена выборка из 729 оригинальных записей учебных заведений. Однако в выборке по строке записи «Адрес» определены учебные заведения, находящиеся за пределами Российской Федерации, информация о которых в Реестре будет отсутствовать – 10 ВУЗов.

На момент написания статьи (03.11.2020) количество ВУЗов, полученных с использованием информационно-аналитической системы, составило 719, против 724, которые были получены в результате поисковых запросов [11], [12].

IV. Верификация операторов ПДн

Выборка исследуемых ВУЗов верифицировалась с существующими открытыми базами данных для подтверждения актуальности представленной в Реестре информации. Проверялись: статус юридического лица (действующее/недействующее); сведения об основном и дополнительных видах деятельности по общероссийскому классификатору видов экономической деятельности (ОКВЭД); наличие государственной аккредитации образовательной деятельности; сведения об учредителях; адрес учреждения и сведения об учете в налоговом органе. Для чего информационно-аналитической системой использовались сведения из ЕГРЮЛ/ЕГРИП [13].

Полученный список ВУЗов сверен с реестром ЕГРЮЛ/ЕГРИП [13] по коду ОКВЭД и актуальности действия юридического лица. Каждая запись из полученной выборки имеет статус «85.22 Образование высшее», что дополнительно подтверждает верность выборки. Механизм взаимодействия с сайтом «Предоставление сведений из ЕГРЮЛ/ЕГРИП» во многом схож с реализованным авторами доступом к реестру [10]. Программный модуль для доступа к информации ЕГРЮЛ/ЕГРИП реализован через внутренний программный интерфейс сайта. При отправке поискового запроса выдаются токены, которые передаются от запроса к запросу, что позволяет ресурсу блокировать слишком частые запросы. Для более быстрого получения информации использовались списки бесплатных прокси серверов. Разработанный программный модуль получения данных из реестра ЕГРЮЛ/ЕГРИП является важным элементом структуры информационно-аналитической системы и позволяет использовать ряд включенных в него сведений для формирования и проверки выборок по отдельным операторским сообществам. Также формируется список актуальных ИНН операторов персональных данных для их последующего использования при организации поиска в Реестре.

Реестр [5] позволяет сделать выгрузку данных в машинно-читаемом формате XML. Данная выборка сконвертирована в формат JSON и загружена в базу данных «mongoDB». Общий объём записей Реестра на 03.11.2020 составляет 516577 записей.

V. Предварительный анализ данных

Сформированная выборка сверялась с включенными в Реестр ВУЗами. В Реестре на 03.11.2020 содержатся записи только 570 из 719 ВУЗов (рис. 3).

Полученные результаты проверены и подтверждены с помощью непосредственного ввода данных об отсутствующих операторах на портале Реестра. В списках отсутствующих в Реестре учебных заведений имеются государственные и негосударственные образовательные учреждения. Тем не менее, высокий процент, содержащихся в Реестре записей о ВУЗах, позволяет считать такую выборку репрезентационной. Таким образом, сформирован набор данных (корпус текстов) в формате JSON, обладающий необходимыми свойствами для последующего интеллектуального анализа полей Реестра.

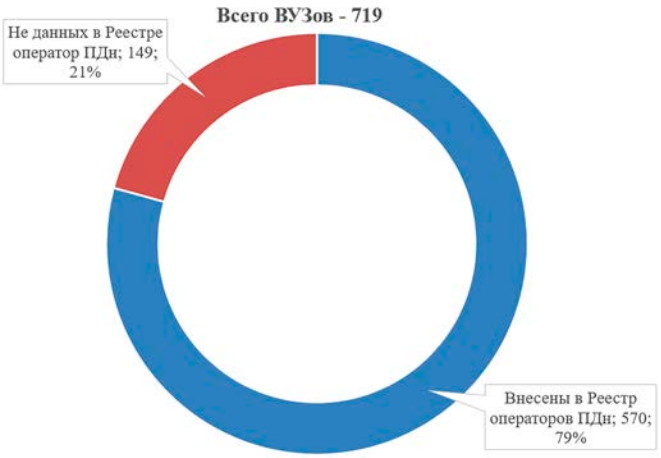

Рис. 3. Количество ВУЗов, включенных в Реестр операторов ПДн

Причины отсутствия записей отдельных учреждений в Реестре определяет Роскомнадзор. Данные, представленные на рисунке 3, подтверждают выводы рабочей группы, состоящей из членов консультативного совета при уполномоченном органе по защите прав субъектов персональных данных, о необходимости оказания операторским сообществом методической помощи своим членам с целью оказания им содействия в исполнении требований Закона о персональных данных [3].

VI. Определение «шаблонного» содержания исследуемых полей Реестра

Пример выборки данных, содержащихся в полях «Перечень действий с персональными данными» (Actions_category) и «Срок или условие прекращения обработки персональных данных» (Stop-condition), приведен в таблице 1.

Таблица 1. Пример выборки данных

Название организации	**Поле «Actions _category»**	**Поле «Stop_condition»**
Название организации №1	сбор, запись, систематизация, накопление, хранение, уточнение (обновление, изменение), извлечение, использование, передача (распространение, предоставление, доступ), удаление, уничтожение	реорганизация или ликвидация юридического лица
Название организации №2	Сбор, систематизация, хранение, использование, уточнение (обновление, изменение), накопление, обезличивание, блокирование, уничтожение, передача в казначейство, банк, налоговую службу, федеральные фонды.	При достижении целей обработки данных и в случае отзыва субъектом своего согласия на обработку; прекращение деятельности университета.
Название организации №3	сбор, запись, систематизация, накопление, хранение, уточнение (обновление, изменение), извлечение, использование, передача (распространение, предоставление, доступ), обезличивание, блокирование, удаление, уничтожение персональных данных.	достижение целей обработки персональных данных или утрата необходимости в их достижении, истечение срока действия договора/согласия или отзыв согласия субъекта персональных данных на обработку его персональных данных, а также выявление неправомерной обработки персональных данных, прекращение деятельности Оператора как юридического лица (ликвидация, реорганизация).
Название организации №4	Сбор, систематизация, накопление, хранение, уточнение, распространение, обезличивание, уничтожение	Прекращение {Название организации № 4} образовательной деятельности, в т.ч. в связи с прекращением деятельности как юридического лица
Название организации №5	автоматизированная обработка – внесение персональных данных в информационные системы операции с персональными данными: сбор, запись, систематизация, накопление, хранение, уточнение (обновление, изменение), извлечение, использование, передача (распространение, предоставление, доступ), обезличивание, блокирование, удаление, уничтожение персональных данных	истечение установленного срока хранения документов, достижение целей обработки персональных данных, отзыв согласия субъекта персональных данных, ликвидация

Под шаблоном в настоящей работе понимается конечный набор устойчивых ключевых слов или N-грамм — словосочетаний, которые по предположению должны все встречаться с максимальной частотой в каждом поле Реестра. Такой шаблон записи можно получить с помощью экспертов на основе анализа требований законодательства о персональных данных, а также автоматически с помощью интеллектуального анализа текста записей при предположении о едином верном заполнении полей Реестра операторами. Возможен и комбинированный вариант, когда с помощью интеллектуальных методов анализа данных эксперту предлагаются варианты устойчивых ключевых словосочетаний и их параметры для составления рекомендаций по заполнению соответствующих полей Реестра.

В результате экспертной оценки, проведенной авторами на основе анализа требований законодательства о персональных данных, сформированный шаблон для поля «Перечень действий с персональными данными» (Actions_category) содержит следующую запись: «сбор, запись, систематизация, накопление, хранение, уточнение, извлечение, использование, передача, обезличивание, блокирование, удаление, уничтожение». Для поля «Срок или условие прекращения обработки персональных данных» (Stop-condition) – «ликвидация, **прекращение** действия договоров, **достижение** целей обработки, **истечение** сроков хранения, отзыв согласия». Поиск выделенных слов может осуществляться с использованием словарей их синонимов. Поиск фразы «прекращение действия договоров» производится с учетом возможного наличия уточняющих слов, например, для включения в результаты поиска фразы: «прекращение действия трудовых и иных договоров».

Для оценки возможности автоматического интеллектуального формирования шаблонов записей, содержащихся в полях «Перечень действий с персональными данными» (Actions_category) и «Срок или условие прекращения обработки персональных данных» (Stop-condition) воспользуемся векторной моделью (Vector Space Model, VSM) [14, 15, 16]. В данной модели каждому слову, содержащемуся в записи, сопоставляется вес в соответствии с выбранной весовой функцией. Рассмотрим весовую функцию TF-IDF, которая есть статистическая мера, используемая для оценки важности слова в контексте документа, входящего в некоторый текстовый корпус.

Весовую функцию частоты появления слова в тексте — TF-IDF можно рассчитать следующим образом [14, 15, 16]:

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D), \quad (1)$$

где TF (Term Frequency) — частота слова t в документе d (сколько раз слово встретилось в документе делить на количество слов в документе), которая рассчитывается по формуле:

$$TF(t,d)=\frac{freq(t,d)}{|D|},$$

здесь $freq(t,d)$ — число вхождений слова t в документе d; $|D|$ — количество d документов в наборе документов D

IDF (Inverse Document Frequency) — обратная частота слов в документах:

$$IDF(t,D)=\log\frac{|D|}{|\{d\in D: t\in d\}|},$$

здесь в числителе $|D|$ — количество d документов в наборе документов D, а в знаменателе $|\{d\in D: t\in d\}|$ — количество документов $d\in D$, в которых встречается слово t.

Обозначим $d\in D$ — одну запись в одном из анализируемых полей, например: «реорганизация или ликвидация юридического лица» для поля «Stop_condition» — первая строка в Табл. 1. Всё множество записей анализируемых текстовых полей обозначим D. Обозначим $t\in d$ —одно слово (терм) из записи поля, например: слово «реорганизация» или «ликвидация юридического лица» из строки «реорганизация или ликвидация юридического лица».

Отметим, что в математической лингвистике принято называть D — набором документов (записей), $d\in D$ — документом (одна запись) из набора документов, $t\in d$ — термом (term) из документа [14, 16]. Термами могут выступать как слова, так и их комбинации, так называемые *N*-граммы. *N*-грамма — это просто последовательность из *N* элементов (звуков, слогов, слов или символов), идущих в каком-то тексте подряд. В нашем случае имеют в виду последовательность слов. Последовательность из двух слов называют биграмма, из трёх элементов — триграмма, из *N* элементов — *N*–грамма [16].

Проанализируем данные (таблица 1) для полей «Actions_category» и «Stop_conditions». В зависимости от решаемой задачи информационного поиска используются различные модификации TF-IDF. При использовании метода расчёта TF-IDF с помощью библиотеки «Scikit-learn» с параметрами по умолчанию не получается выделить устойчивые ключевые словосочетания – N-граммы. Необходимо задать следующий параметр: ngram_range=(1, 4), которые в явном виде задают принудительное использование *N*-грамм (униграмм, биграмм, триграмм и квадрограмм). Построим весовую функцию TD-IDF для полей «Actions_category» и «Stop-condition», используя библиотеку Scikit-learn с параметром ngram_range=(1, 4).

Таблица 2. Значение TF-IDF для «Actions_category»

Набор устойчивых ключевых слов	TF-IDF
хранение	0,1948
сбор	0,1830
данных	0,1715
уточнение	0,1715
накопление	0,1703
уничтожение	0,1652
использование	0,1620
накопление хранение	0,1569
хранение уточнение	0,1557
изменение	0,1533
систематизация	0,1518
обновление	0,1506
персональных	0,1498
персональных данных	0,1498
передача	0,1458
накопление хранение уточнение	0,1458
обновление изменение	0,1439
уточнение обновление	0,1435
уточнение обновление изменение	0,1411
систематизация накопление	0,1407
систематизация накопление хранение	0,1356
хранение уточнение обновление	0,1348
хранение уточнение обновление изменение	0,1328
накопление хранение уточнение обновление	0,1296
систематизация накопление хранение уточнение	0,1257
запись	0,1213
блокирование	0,1201
удаление	0,1178
сбор запись	0,1166
извлечение	0,1083
извлечение использование	0,1031
распространение	0,1031
запись систематизация	0,1012
удаление уничтожение	0,1000
сбор запись систематизация	0,0988
изменение извлечение	0,0988
обезличивание	0,0980
обновление изменение извлечение	0,0980
уточнение обновление изменение извлечение	0,0976
запись систематизация накопление	0,0976

Таблица 3. Значение TF-IDF для «Stop_conditions»

Набор устойчивых ключевых слов	TF-IDF
ликвидация	0,4117
деятельности	0,3445
прекращение	0,3347
прекращение деятельности	0,2932
реорганизация	0,2859
лица	0,2077
юридического	0,2040
юридического лица	0,2040
ликвидация реорганизация	0,1991
данных	0,1368
учреждения	0,1356
персональных	0,1332
персональных данных	0,1332
деятельности юридического лица	0,1112
деятельности юридического	0,1112
прекращение деятельности юридического лица	0,1075
прекращение деятельности юридического	0,1075
лица ликвидация	0,0855
юридического лица ликвидация	0,0843

Рассмотрим данные, представленные в таблице 2. В этой таблице все слова упорядочены по убыванию весовой функции TF-IDF. Из таблицы 2 видно, что слово «хранение» встречается также и в строке с биграммой «накопление хранение». Предлагается убрать повторяющееся слова и тем самым сократить объём выборки для набора устойчивых ключевых выражений. Уберем из таблиц 2 и 3 все повторяющиеся слова, использующиеся в биграммах и триграммах. Результат представим в таблицах 4 и 5.

Таблица 4. Значение TF-IDF для полей «Actions_category» без повторяющихся слов

Набор устойчивых ключевых слов	TF-IDF
персональных данных	0,1498
передача	0,1458
хранение уточнение обновление изменение	0,1328
накопление хранение уточнение обновление	0,1296
систематизация накопление хранение уточнение	0,1257
блокирование	0,1201
сбор запись	0,1166
извлечение использование	0,1031
распространение	0,1031
запись систематизация	0,1012
удаление уничтожение	0,1000
обезличивание	0,0980
запись систематизация накопление	0,0976

Таблица 5. Значение TF-IDF для полей «Stop_conditions» без повторяющихся слов

Набор устойчивых ключевых слов	TF-IDF
ликвидация реорганизация	0,1991
учреждения	0,1356
персональных данных	0,1332
прекращение деятельности юридического лица	0,1075
юридического лица ликвидация	0,0843

В результате интеллектуального анализа текста исследуемых полей Реестра сгенерирован шаблон для поля «Перечень действий с персональными данными» (Actions_category), который содержит следующую запись с учетом удаления повторяющихся слов: «сбор, запись, систематизация, накопление, хранение, уточнение, извлечение, использование, передача, обезличивание, блокирование, удаление, уничтожение, обновление, изменение, распространение, персональных данных». Принимая во внимание, что слова «обновление», «изменение» и «распространение» в понятиях Федерального закона [3] носят уточняющий характер основного термина (стоят после него в скобках) и могут не применяться или выборочно использоваться в описании действий операторов с персональными данными, а биграмма «персональных данных» к отношению перечисляемых действий может не добавляться и не является действием, авторами в экспертный поисковый шаблон данные слова включены не были. В остальном сформированный с помощью интеллектуальных методов шаблон соответствует экспертному. Тем не менее, наличие уточняющих действий с данными и их разный вес TF-IDF, представляет интерес для дальнейшего анализа экспертами и возможной корректировки шаблона.

Для поля «Срок или условие прекращения обработки персональных данных» (Stop-condition) шаблон, сформированный с использованием методов интеллектуального анализа данных, с учетом удаления повторяющихся слов (N-грамм) содержит следующую запись: «ликвидация, реорганизация, прекращение деятельности юридического лица, персональных данных, учреждения». Совпадение с экспертным шаблоном только в одном слове делает такой шаблон не состоятельным и требует дополнительной экспертной оценки. Отсутствие высоких показателей TF-IDF у N-грамм «достижение целей обработки», «истечение сроков хранения», «отзыв согласия», «действия договоров», а также у отдельных входящих в них терм позволяет сделать заключение о неиспользовании значительной частью операторского сообщества таких условий для прекращения обработки персональных данных. Однако они реализуют принципы и соответствуют условиям обработки персональных данных, предусмотренными Федеральным законом [3], а значит должны применяться операторами, входящими в университетское операторское сообщество. Не выполнение данных требований ставит под угрозу обрабатываемые персональные данные.

Отметим, что несмотря на существенное развитие современных инструментов интеллектуального анализа данных составить точный словарь устойчивых выражений и ключевых слов для формирования поискового шаблона без привлечения экспертов невозможно. Методы интеллектуального анализа данных эффективно используются для обработки больших массивов информации [17]. Применение весовой функции TF-IDF существенно упрощает задачу экспертной оценки, содержащихся в полях Реестра данных.

VII. Анализ данных с учетом сформированных шаблонов

Для анализа данных полей Реестра воспользуемся алгоритмом прямого информационного поиска (алгоритм последовательного поиска) [16]. Полученные шаблоны представлены в виде списка ключевых слов и устойчивых словосочетаний (N-грамм), которые будут являться поисковым запросом к имеющимся данным. Отметим, что в исходном тексте могут встретиться семантически связные слова, стоящие в разном падеже, роде, числе. Используемые в работе методы определения TF-IDF, в силу своей специфики, будут воспринимать их как семантически не связанные друг с другом, что приведёт к искажению результатов, полученных в ходе исследования. Например, слово «сбор» в исследуемом тексте может быть записано как: сбор, сборов, собирать, собирают и в других словоформах. Поэтому предлагается использовать преобразование всех слов в единую базовую форму. Это можно сделать, используя стемминг. Стемминг – это грубый эвристический процесс, который отрезает «лишнее» от корня слов, часто это приводит к потере словообразовательных суффиксов [14, 16]. На исходных данных (таблица 1) стемминг не показал хороших результатов. Более корректный результат получен при использовании лемматизации — специальной функции, приводящей слова к нормальной форме с помощью морфологического анализа, заключающегося в определении части речи и последовательном изменении её морфологических характеристик до нормальной формы [14, 16].

Для лемматизации русскоязычных текстов в настоящее время широко используются библиотеки Pymorphy2, PyMystem3, Snowball [14, 16]. Проанализируем эти библиотеки для целей использования в нашем исследовании. Проводить сравнение результатов будем относительно двух

характеристик: качество нормализации слов и скорость работы анализаторов [14]. Для решения данной задачи использованы данные из поля «Actions_category», разметка проводилась вручную. Данные анализа представлены в таблице 6.

Таблица 6. Сравнение методов лемматизации слов

	Качество (%)	Скорость (сек)
PyMystem3	89,6	44.36
Pymorphy2	81,5	0.21
Snowball	76,6	0.19

Из представленных в таблице 6 результатов видно, что при решении задачи приведения к нормальной форме по качеству нормализации эффективнее PyMystem3, которую и будем использовать в наших исследованиях. Следует отметить, что PyMystem3 работает намного дольше, но выдает более качественные результаты.

Также при выполнении запросов необходимо воспользоваться словарями синонимов, которые предлагается формировать для всех ключевых сочетаний, используя полученные ранее значения весовой функции TF-IDF. Для получения списков синонимов использовался «Открытый словарь синонимов» [18]. Например, создадим словарь синонимов для слова «ликвидация». Список, полученный из словаря синонимов [18], будет иметь вид: «оборка», «окончание», «разгон», «глушение», «погашение», «расторжение», «завершение», «прекращение», «уничтожение», «истребление», «искоренение», «упразднение», «вытравление», «вытравливание», «тушение», «устранение», «раскассирование», «отмена», «ликвидирование», «выкорчевывание», «слом», «изжитие», «сокрушение», «гекатомба», «изничтожение». Найдем встречающиеся синонимы в исследуемых полях Реестра. Полученный список синонимов будет иметь вид: «окончание», «расторжение», «завершение», «прекращение».

Представим конвейер (общий алгоритм) обработки текста полей «Actions_category» и «Stop_conditions» следующим образом:

Рис.4. Конвейер (общий алгоритм) обработки текста

Для реализации алгоритма поиска, по ключевым словам, применялась следующая последовательность действий:

- предварительная фильтрация (удаление знаков препинания и лишних символов, перевод слов в нижний регистр) [14,16];
- удаление стоп-слов: использовались списки стоп-слов из NLTK и Scikit-learn3 [14,15];
- токенизация — разбиение на слова или текстовые единицы [14] (использовалась библиотека NLTK (Natural Language Toolkit);
- лемматизация — процесс приведения слова к нормальной (словарной) форме [16] (использовалась библиотека PyMystem3);
- генерация словарей синонимов (использовался открытый словарь синонимов [18]);
- вычисление числа вхождения ключевых слов для каждой записи (использовалась библиотека FuzzyWuzzy);
- кластеризация (получение сходных между собой записей) [14,19].

В текущей версии информационно-аналитической системы использовалась поисковая система FuzzyWuzzy, так как конвейер (рис. 4) большей частью реализован на Python. Поисковая система настроена для поиска устойчивых сочетаний слов с учетом нахождения между ними 1 – 4 слов. Например, поисковый запрос «прекращение деятельности юридического лица» найдет также и следующую запись: «прекращение деятельности юридического (физического) лица».

Проанализируем размеры полученных кластеров. Для этого построим кольцевую диаграмму вхождений ключевых слов шаблона в результаты поиска для поля «Actions_category» (рисунок 5).

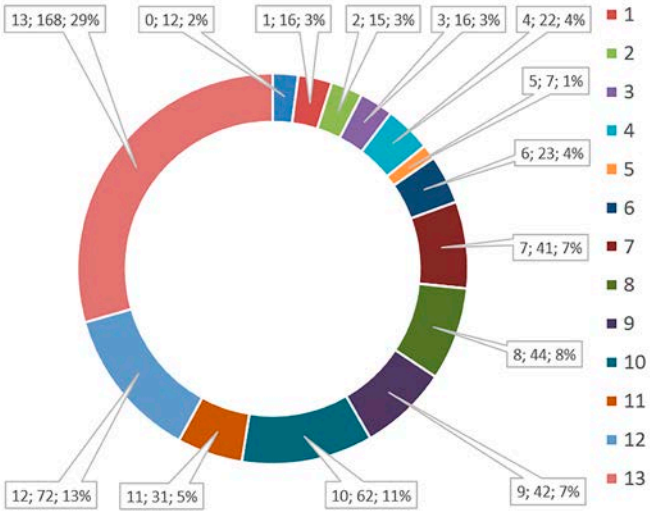

Рис.5. Количество вхождений слов шаблона для поля «Actions_category»

Как видно из рисунка 5 максимальный размер определенного кластера (29%) соответствует попаданию всех 13 ключевых слов (в 168 случаях), что подтверждает экспертные выводы и оценки. В кластере, содержащем 12 слов из поискового запроса, с максимальной частотой отсутствует слово «обезличивание», что также совпадает с предположением экспертов. Итоговое совпадение в таком случае составляет 42%. В кластерах 5% и 11% в большинстве случаев кроме вышеописанных отсутствуют слова «блокирование» и «извлечение» соответственно, что может говорить о неправильной трактовке операторами таких действий с персональными данными. Так статьей 21 Федерального закона [3] прямо предусмотрена обязанность оператора по блокированию персональных данных, если они обрабатываются неправомерно, содержат неточности, или отсутствует возможность уничтожения персональных данных в течение установленного срока. Невключение действия «извлечение» в уведомления может быть связана с отсутствием такого понятия в Федеральном законе [3] и подзаконных актах и, вследствие чего, неправильной трактовкой этого термина операторами.

Результаты информационного поиска для поля «Stop_conditions» представлены на рисунке 6.

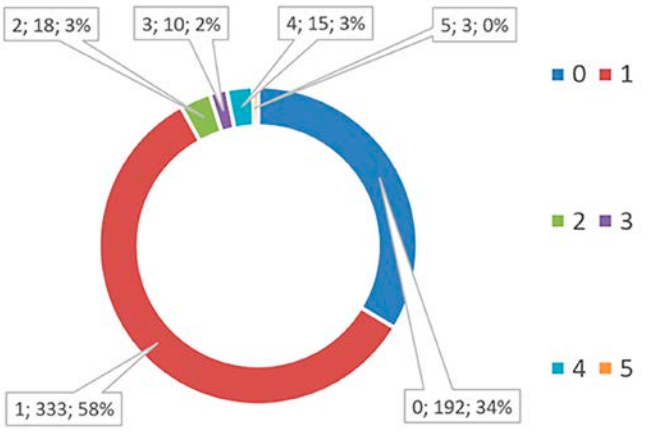

Рис.6. Количество вхождений ключевых слов и *N*-грамм шаблона для поля «Stop_conditions»

Как видно из рисунка 6 максимальный размер определенного кластера (58%) соответствует попаданию лишь одного элемента из сформированного поискового шаблона в результаты поиска: с максимальной частотой в данный кластер вошло ключевое слово «ликвидация». Это соответствует одному из условий прекращения обработки персональных данных – ликвидации юридического лица. Второй по размеру кластер (34%) вообще не имеет вхождений ключевых слов или *N*-грамм из экспертного шаблона. Анализ данных этого кластера показал, что с максимальной частотой в нем присутствует *N*-грамма «прекращение деятельности …» как синоним понятия «ликвидация юридического лица». Вместе с тем понятие «ликвидация юридического лица» предусмотрено статьей 61 ГК РФ [20], в соответствии с которым влечет его прекращение, а понятие «прекращение деятельности юридического лица» в гражданском кодексе отсутствует. Это подтверждает верность сформированного экспертами шаблона в части формулировок. Однако, если не придерживаться четких юридических терминов, то для целей заполнения уведомлений можно считать эти понятия синонимами. Таким образом, понятия «ликвидация или прекращение деятельности юридического лица» в сумме встречаются в 92% поданных уведомлений. Также целесообразно будет добавить в поисковый шаблон *N*-грамму «прекращение деятельности», в том числе для проведения исследований других операторских сообществ.

Остальные *N*-граммы из экспертного шаблона в результатах поиска представлены следующими процентными соотношениями: прекращение действия договоров (6%), отзыв согласия (5%), истечение сроков хранения (4%), достижение целей обработки (3%). Все перечисленные выше *N*-граммы являются условиями прекращения обработки персональных данных в ВУЗах, однако все эти условия встречаются в уведомлениях только 3 учебных заведений или в 0,5% от общей выборки. Не знание и не соблюдение условий прекращения обработки персональных данных влечет за собой возможность их неправомерного использования и создает угрозу их конфиденциальности. Таким образом еще раз подтверждена необходимость разработки отраслевых методических рекомендаций для операторских сообществ, направленных на создание условий для безопасной обработки персональных данных.

VIII. Заключение

Проведенные исследования позволили получить следующие основные результаты.

Определено, что в Реестре операторов персональных данных на 03.11.2020 содержатся записи только 570 из 719 ВУЗов России. Примененный подход к формированию выборки ВУЗов, можно использовать и для ряда подобных задач, связанных с получением

данных из открытых источников в сети Интернет.

Сформирован корпус текстов по тематике в области обработки и защиты данных, который может использоваться для проведения исследований интеллектуальных методов анализа текста в указанной сфере. Определено, что ВУЗы имеют схожую структуру при заполнении исследуемых полей Реестра.

Выполнена проверка возможности автоматического формирования «эталонного» (шаблонного) заполнения отдельных полей Реестра на основе интеллектуального анализа данных из всей исследуемой выборки. Показана возможность автоматического формирования шаблона для поля «Actions_category», когда для этой задачи используются отдельные ключевые слова. Для поля «Stop_conditions» автоматически сформировать шаблон из устойчивых ключевых слов и словосочетаний не получилось из-за отсутствия необходимого количества данных в анализируемых полях Реестра и разных используемых операторами формулировок при описании условий прекращения обработки персональных данных. Для верификации полученных данных необходимо привлекать экспертов, однако их работа при использовании интеллектуального анализа данных существенно упрощается.

Проведен поиск и интеллектуальный анализ данных на основе сформированных экспертных шаблонов. Результаты подтвердили верность сформированных поисковых шаблонов, выявили необходимость корректировки направленных операторами уведомлений и процентное соотношение верно заполненных исследуемых полей Реестра. Наиболее проблемным для заполнения полем у операторов явилось «срок или условие прекращения обработки персональных данных» - только 3 ВУЗа использовали все определенные экспертами условия.

Результаты работы позволяют рекомендовать университетскому операторскому сообществу использовать сформированные и проверенные экспертные шаблоны при заполнении уведомлений об обработке персональных данных. Рассмотренные в работе подходы к формированию экспертных оценок с использованием методов интеллектуального анализа данных могут использоваться для разработки отраслевых рекомендаций по организации обработки и защите персональных данных для иных операторских сообществ.

БИБЛИОГРАФИЯ

[1] Отчет о деятельности Уполномоченного органа по защите прав субъектов персональных данных за 2019 год, https://rkn.gov.ru/docs/Otchet_UO-2019_new.pdf

[2] Методические рекомендации по уведомлению уполномоченного органа о начале обработки персональных данных и о внесении изменений в ранее представленные сведения», приложение к приказу Роскомнадзора от 30.05.2017 № 94, https://24.rkn.gov.ru/directions/p5987/p4245/

[3] Федеральный закон от 27.07.2006 №152-ФЗ «О персональных данных», http://pravo.gov.ru/proxy/ips/?docbody&nd=102108261

[4] Методические рекомендации по разработке отраслевого кодекса поведения в области защиты прав субъектов персональных данных, https://pd.rkn.gov.ru/docs/Metodicheskie_rekomendacii_rabochej_gruppy_KS.pdf

[5] Реестр операторов, осуществляющих обработку персональных данных, https://pd.rkn.gov.ru/operators-registry/operators-list/

[6] Форма уведомления. Портал персональных данных уполномоченного органа по защите прав субъектов персональных данных, https://pd.rkn.gov.ru/operators-registry/notification/form

[7] Лось В.П., Никульчев Е.В., Пушкин П.Ю., Русаков А.М. Информационно-аналитическая система мониторинга выполнения операторами персональных данных требований законодательства // Проблемы информационной безопасности. Компьютерные системы. – 2020. – №3. – С. 16-23.

[8] Москаленко А. А., Лапонина О. Р., Сухомлин В. А. Разработка приложения веб-скрапинга с возможностями обхода блокировок // Современные информационные технологии и ИТ-образование. – 2019. – №2. – С. 413-420.

[9] Портал открытых данных Российской Федерации, https://data.gov.ru/

[10] Реестр организаций, осуществляющих образовательную деятельность по имеющим государственную аккредитацию образовательным программам, http://isga.obrnadzor.gov.ru/accredreestr/

[11] Официальный сайт о высшем образовании в России для иностранных студентов, https://studyinrussia.ru

[12] АО «Аргументы и Факты», https://aif.ru

[13] Предоставление сведений из ЕГРЮЛ/ЕГРИП, https://egrul.nalog.ru

[14] Jurafsky D., Martin J. H. Speech and language processing (August 2020), https://web.stanford.edu/~jurafsky/slp3/

[15] Митренина О. В., Николаев И. С., Ландо Т. М. Прикладная и компьютерная лингвистика. – 2016. – 320 с.

[16] Бенгфорт Б., Билбро Р., Охеда Т. Прикладной анализ текстовых данных на Python //Машинное обучение и создание приложений обработки естественного языка. СПб.: Питер. – 2019. – 368 с.

[17] Силаева А. Э., Никульчев Е. В., Ильин Д. Ю., Малых С. Б. Обработка открытых вопросов веб-опросов в системе образования на основе методов искусственного интеллекта. // Тринадцатая международная конференция «Управление развитием крупномасштабных систем» (MLSD'2020). Россия, Москва, ИПУ РАН, 28-30 сентября 2020 г. – С. 1692-1697

[18] Бесплатный онлайн-словарь русских синонимов, https://synonymonline.ru

[19] Пархоменко П.А., Григорьев А.А., Астраханцев Н.А. Обзор и экспериментальное сравнение методов кластеризации текстов // Труды ИСП РАН. – 2017. – №2. – С. 161-200.

[20] Гражданский кодекс Российской Федерации: Часть первая – четвертая: [Принят Гос. Думой 23 апреля 1994 года, с изменениями и дополнениями по состоянию на 03 декабря 2020 г.] // Собрание законодательства РФ. – 1994. – № 22. Ст. 2457.

Results of automatic mining individual fields of personal data operators register

P. Yu. Pushkin, A.M. Rusakov

Abstract—The work presents the results of mining the records contained in the fields of the register personal data operators "list of actions with personal data" and "period or condition of termination personal data processing" and assessment of their compliance with the requirements of the legislation on personal data. Higher educational institutions were chosen as the research operator community, which allows taking into account similar features of personal data processing when forming expert assessments and intelligent data analysis. For the purpose of the study, a body of texts has been formed that can be used to analyze data mining methods by information processing and protection topics. For text mining, the following libraries were used: Scikit-learn, Gensim, PyMystem3, FuzzyWuzzy. Search queries were performed taking into account synonym dictionaries and the fuzzy location of words. To find stable keyword combinations, the TF-IDF weight function was calculated. Comparison of methods lemmatization of words for research purposes was made. The obtained results show the fidelity of expert assessments on filling in the fields of the register of personal data operators: the maximum cluster determined by the results of mining analysis corresponds to the expert template. The results of automatic mining require the verification of an expert in the field of personal data processing and protection. The use of data mining methods makes it possible to significantly increase the efficiency of experts when working with large volumes of information contained in the register of personal data operators. The work is aimed at forming separate sections of recommendations for the development of a sectoral (in the field of higher education and science) code of conduct in the field of protection of the rights of personal data subjects in order to increase the level of security of such information.

***Keywords*—Personal data, personal data protection, information and analytical systems, intelligent data analysis.**

REFERENCES

[1] Report on the activities of the Authorized Body for the Protection of the Rights of Personal Data Subjects for 2019, https://rkn.gov.ru/docs/Otchet_UO-2019_new.pdf

[2] Methodological recommendations on notification of the authorized body on the beginning of personal data processing and on amendments to previously submitted information, "appendix to the order of Roskomnadzor dated 30.05.2017 No. 94, https://24.rkn.gov.ru/directions/p5987/p4245/

[3] Federal law of 27.07.2006 No. 152-FZ "About Personal Data", http://pravo.gov.ru/proxy/ips/?docbody&nd=102108261

[4] Methodological recommendations for the development of an industry code of conduct in the field of protection of the rights of personal data subjects, https://pd.rkn.gov.ru/docs/Metodicheskie_rekomendacii_rabochej_gruppy_KS.pdf

[5] Register of operators processing personal data, https://pd.rkn.gov.ru/operators-registry/operators-list/

[6] Notification form. Personal data portal of the authorized body for protection of rights of personal data subjects, https://pd.rkn.gov.ru/operators-registry/notification/form

[7] Los V.P., Nikulchev E.V., Pushkin P.Yu., Rusakov A.M. Information and analytical system for monitoring the compliance of personal data operators with the requirements of the legislation//Problems of information security. Computer systems. 2020. No. 3. P. 16-23.

[8] Moskalenko A. A., Laponina O. R., Sukhomlin V. A. Development of a web-scraping application with blocking bypass capabilities//Modern information technologies and IT education. 2019. No. 2. P. 413-420.

[9] Open data portal of the Russian Federation, https://data.gov.ru/

[10] Register of organizations carrying out educational activities under state accredited educational programs, http://isga.obrnadzor.gov.ru/accredreestr/

[11] Official website on higher education in Russia for foreign students, https://studyinrussia.ru

[12] LCC «Argumenty i Fakty», https://aif.ru

[13] Provision of information from the EGRUL/EGRIP, https://egrul.nalog.ru

[14] Jurafsky D., Martin J. H. Speech and language processing (August 2020), https://web.stanford.edu/~jurafsky/slp3/

[15] Mitrena O. V., Nikolaev I. S., Lando T. M. Applied and computer linguistics, 2016. P.360.

[16] Bengfort B., Bilbro R., Ojeda T. Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning. – " O'Reilly Media, Inc.", 2018.

[17] Silaeva A. E., Nikulchev E. V., Ilyin D. Yu., Maly S. B. Processing open questions of web surveys in the education system based on artificial intelligence methods. // Conference: 13 International Conference "Managing the Development of Large-Scale Systems" – (MLSD'2020) At: Moscow, Russia Volume: Pp. 1692-1697

[18] Parkhomenko P.A., Grigoryev A.A., Astrakhantsev N.A. Review and experimental comparison of clustering methods of texts//Proceedings of the ISP RAS. 2017. No. 2. P. 161-200.

[19] Free online dictionary of Russian synonyms, https://synonymonline.ru

[20] Civil Code of the Russian Federation: Part One - Fourth: [Adopted by the State. The Duma on April 23, 1994, with amendments and additions as of December 03, 2020]//Assembly of Legislation of the Russian Federation. 1994. No. 22. Article. 2457.