

Вестник Сыктывкарского университета.
Серия 1: Математика. Механика. Информатика. 2023.
Выпуск 3 (48)
Bulletin of Syktyvkar University.
Series 1: Mathematics. Mechanics. Informatics. 2023; 3 (48)

Научная статья

УДК 004.912

https://doi.org/10.34130/1992-2752_2023_3_72

**РАЗРАБОТКА И ОРГАНИЗАЦИЯ БИЗНЕС-ПРОЦЕССОВ
ПОДГОТОВКИ ЧАСТОТНЫХ СЛОВАРЕЙ В ЦЕЛЯХ
АВТОМАТИЗАЦИИ ОБРАБОТКИ ЕСТЕСТВЕННОГО
ЯЗЫКА ПРИ ВЫПОЛНЕНИИ ЗАДАЧ ПО
ЛИНГВИСТИЧЕСКОМУ АНАЛИЗУ ТЕКСТОВ**

**Михаил Сергеевич Крашенинников^{1,2}, Иван Иванович
Лавреш², Владимир Александрович Устюгов²**

¹ГАУ РК «Центр информационных технологий»

²Сыктывкарский государственный университет
имени Питирима Сорокина, kib@syktsu.ru

Аннотация. В данной статье рассматриваются процессы обработки текстовых данных в рамках информационно-аналитического обеспечения деятельности органов власти Республики Коми. Показано взаимодействие государственных информационных систем и органов власти Республики Коми и их различных подразделений. Проведена последовательная декомпозиция процесса информационно-аналитического обеспечения, основной составляющей которого является обработка входящих данных, что в дальнейшем сводится к процессу классификации текстов на основе онтологии. Особое место занимает описание применения метода частотного анализа при составлении словаря синонимов для объектов онтологии. Применение такого метода позволяет автоматизировать задачи классификации текстов без постоянного вмешательства оператора-аналитика.

Ключевые слова: обработка естественного языка, частотный анализ текста, лингвистический анализ текста, частотные словари, информационно-аналитическое обеспечение

Для цитирования: Крашенинников М. С., Лавреш И. И., Устюгов В. А. Разработка и организация бизнес-процессов подготовки частотных словарей в целях автоматизации обработки естественного языка при выполнении задач по лингвистическому анализу текстов // *Вестник Сыктывкарского университета. Сер. 1: Математика. Механика. Информатика*. 2023. Вып. 3 (48). С. 72–89. https://doi.org/10.34130/1992-2752_2023_3_72

Article

Development and organization of business processes for preparing frequency dictionaries in order to automate natural language processing when performing tasks of linguistic text analysis

Mikhail S. Krasheninnikov^{1,2}, Ivan I. Lavresh², Vladimir A. Ustyugov²

¹SAI of the Komi Republic «Information Technology Center»

²Pitirim Sorokin Syktyvkar State University, kib@syktsu.ru

Abstract. This article discusses the processes of processing text data within the framework of information and analytical support for the activities of authorities of the Komi Republic. The interaction of state information systems and authorities of the Komi Republic and their various divisions is shown. A sequential decomposition of the information and analytical support process has been carried out, the main component of which is the processing of incoming data, which further reduces to the process of classifying texts based on ontology. A special place is occupied by the description of the application frequency analysis method in compiling a dictionary of synonyms for ontology objects. The use of this method allows you to automate text classification tasks without the constant intervention of an operator-analyst.

Keywords: natural language processing, frequency analysis of text, linguistic analysis of text, frequency dictionaries, information and analytical support

For citation: Krasheninnikov M. S., Lavresh I. I., Ustyugov V. A. Development and organization of business processes for preparing frequency dictionaries in order to automate natural language processing when performing tasks of linguistic text analysis. *Vestnik*

Syktvykarskogo universiteta. Seriya 1: Matematika. Mekhanika. Informatika [Bulletin of Syktvykar University, Series 1: Mathematics. Mechanics. Informatics], 2023, no 3 (48), pp. 72–89. (In Russ.) https://doi.org/10.34130/1992-2752_2023_3_72

Введение

Информационно-аналитическое обеспечение государственного управления представляет собой совокупность процессов, направленных на создание оптимальных условий для удовлетворения информационных потребностей должностных лиц, снижение неопределенности на этапах выработки и принятия управленческих решений, осуществление адаптивного управления процессами в организации [1].

В связи с увеличением скорости принятия решений и приближением скорости обучения и реагирования к предельно возможной для человека потребовалось автоматизировать процессы сбора, обработки, структурирования, анализа информации и предварительного принятия решений на ее основе, в том числе в сфере государственного управления.

Одним из вариантов усовершенствования процесса на этапах обработки и структурирования данных рассматривается частотный анализ текстов как метод, позволяющий обеспечить более полную, быструю и унифицированную обработку растущих объемов текстовых данных в ходе формирования базы знаний предметной области.

Таким образом, цель статьи — предложить способ применения частотного анализа текстов при выполнении задач лингвистического анализа текстов в процессе информационно-аналитического обеспечения деятельности органов государственной власти Республики Коми.

Информационно-аналитическое обеспечение деятельности органов власти Республики Коми

Государственное автономное учреждение Республики Коми «Центр информационных технологий» (ГАУ РК «ЦИТ») осуществляет разработку, внедрение, развитие и эксплуатацию информационных систем, техническое сопровождение компонентов инфраструктуры центров обработки данных, ведение баз данных и информационно-аналитическое обеспечение деятельности органов государственной власти Республики Коми.

Большинство информационных систем органов власти Республики Коми можно разделить на основные группы: автоматизация отдельных процессов, автоматизация отчетности, сбор и хранение данных. Практически все они имеют отраслевую специализацию и охватывают информационный уровень процесса информационно-аналитического обеспечения. Аналитический уровень вне рамок отраслевой специализации реализуется при использовании информационных систем, находящихся в ведении отдела анализа данных и развития аналитических платформ (АДиРАП). К ним относятся ГИС «Единая информационно-аналитическая система органов власти Республики Коми» (ЕИАС РК) [2; 3], ИАС «Семантический архив» [4] и ИАС «Диссонанс».

Процесс информационно-аналитического обеспечения деятельности органов государственной власти, целью которого является повышение качества принятия управленческих решений, осуществляется отделом АДиРАП во взаимодействии с иными структурными подразделениями ГАУ РК «ЦИТ» и внешними контрагентами. В стандартном представлении процесс информационно-аналитического обеспечения состоит из следующих стадий: формирования технического задания на основе задачи заказчика, сбора первичных данных из указанных источников, их обработки, анализа и представления результатов в формате оговоренного заранее информационного продукта.

Относительно формы представления данных в информационно-аналитическом обеспечении выделяются отдельные направления аналитической работы с числами-показателями, мультимедийным контентом (изображение, видеозапись, звук) и текстовой информацией [1]. Анализ текстовых данных позволяет осуществлять бизнес-разведку и оценку рисков, OSINT-разведку, анализ политической и социальной ситуации, мониторинг общественного мнения, отслеживание показателей, связанных с оценкой информационной открытости.

Сбор большого объема неструктурированной текстовой информации из открытых источников (СМИ, социальные медиа, мессенджеры, онлайн-базы данных), ее обработка, хранение и анализ осуществляются с использованием ИАС «Семантический архив» и ИАС «Диссонанс». Данные процессы прошли частичное улучшение. Для сбора данных совместно с ИАС «Семантический архив» применяется комплекс дополнительных инструментов: распознавание текста с изображений с помощью OCR (Optical Character Recognition), транскрибация опубликованных

в открытом доступе выпусков телевизионных передач, автоматизация внесения текстов в базу данных с помощью технологии RPA (Robotic Process Automation). Обработка текстовых данных в части присвоения тексту значения тональности осуществляется при помощи лингвистического модуля компании PROMT. Представление пользователю результатов обработки и анализа данных из базы данных ИАС «Семантический архив» производится в интерфейсе ИАС «Диссонанс», которая является инициативной разработкой отдела АДиРАП. Архитектура взаимодействия используемых решений представлена на рис. 1.

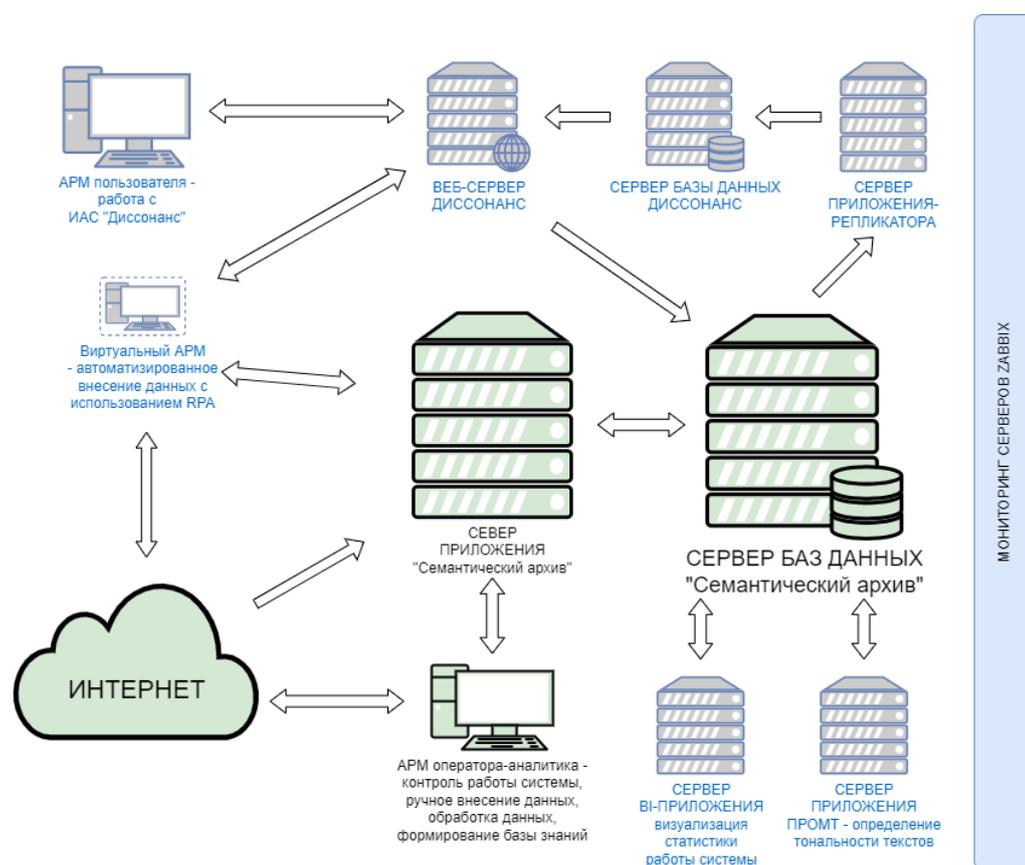


Рис. 1. Архитектура взаимодействия ИАС «Семантический архив», ИАС «Диссонанс» и дополнительных сервисов

Однако лингвистическая обработка текста, направленная на формирование базы знаний в формате древовидной онтологии и последующую классификацию текстовых документов в зависимости от упоминаемости объектов онтологии, продолжает выполняться на основе экспертно-

го мнения операторов-аналитиков в ручном режиме через клиентское приложение ИАС «Семантический архив».

Обработка текстовых данных в рамках процесса информационно-аналитического обеспечения

Последовательная декомпозиция процесса информационно-аналитического обеспечения, основной составляющей которого является обработка входящих данных, в детализированном виде отображенная на рис. 2, приводит нас к процессу классификации текстов на основе онтологии, построенной оператором-аналитиком. ИАС «Семантический архив» дает возможность аналитикам формировать базу знаний предметной области в форме древовидной онтологии, хранить досье на объекты онтологии (персоны, организации, документы стратегического планирования, источники данных, события и факты, тематики и другие сущности), описывать их взаимоотношения, осуществлять автоматический поиск и выделение объектов в текстах публикаций [4].

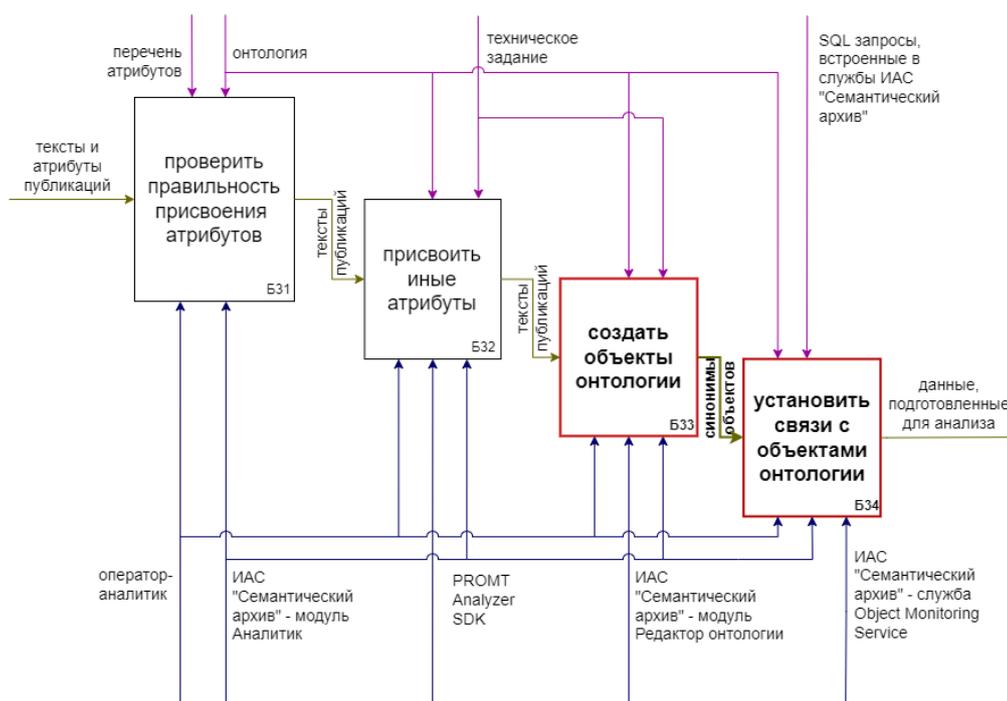


Рис. 2. Процесс обработки текстовых данных

Процесс классификации текстов на основе онтологии объединяет в себе создание оператором-аналитиком классов и объектов онтологии,

составление словаря синонимов для объектов и последующее установление связи объектов с текстами публикаций за счет автоматического поиска синонимов в текстах, сохраненных в базе данных ИАС «Семантический архив». Под синонимом объекта в ИАС «Семантический архив» подразумевается слово, словосочетание или полнотекстовый запрос, позволяющие определить семантическую связанность между объектом и текстом.

Наиболее трудоемкая часть процесса создания объектов онтологии выделена прямоугольной областью на рис. 3. Она заключается в осуществлении оператором-аналитиком лингвистического анализа текстов публикаций на предмет выявления синонимов для создаваемых объектов.

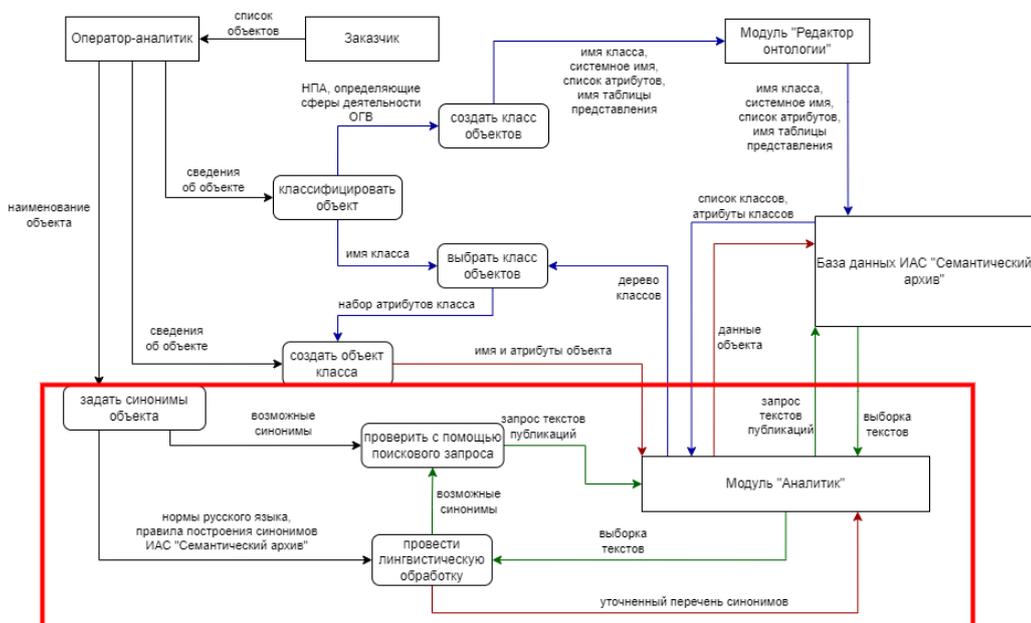


Рис. 3. Процесс создания объектов онтологии

Детализация действий, выполняемых оператором-аналитиком при составлении в базе данных словаря синонимов, приведена на рис. 4.

По мере роста объемов информации, которую следует прочитать, проанализировать, изучить, для специалистов отдела АДРАП острее встает вопрос автоматизации процесса лингвистического анализа текстов при создании объектов онтологии. Так, за 2022 год в базу данных ИАС «Семантический архив» внесено 3 836 347 публикаций СМИ, новостных и официальных сайтов, 4 935 113 постов и комментариев

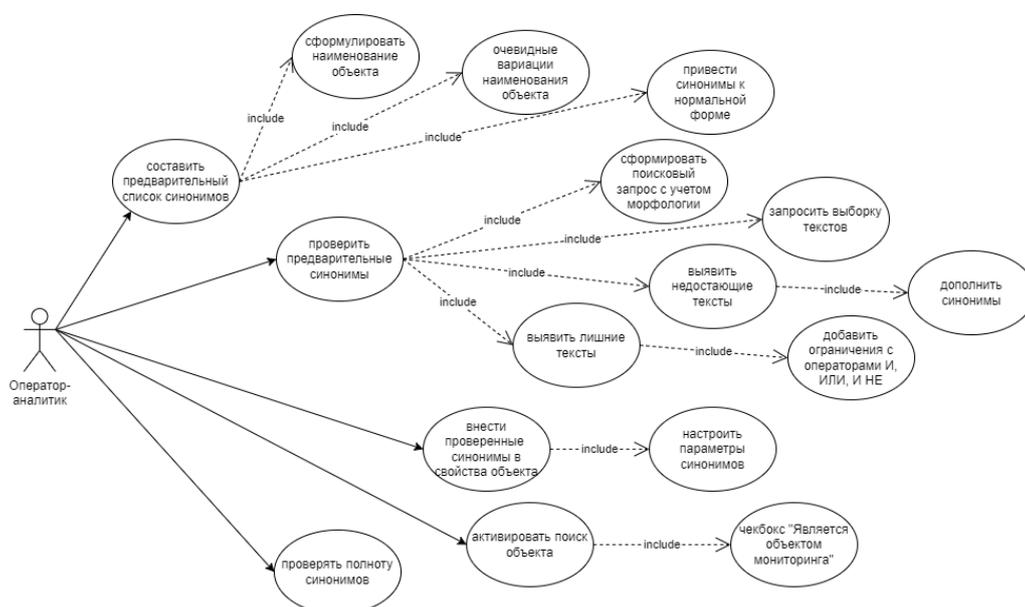


Рис. 4. Действия, выполняемые оператором-аналитиком в процессе создания синонимов объектов онтологии

социальных сетей и мессенджеров. Ручной режим требует значительных трудозатрат и не позволяет охватить весь объем текстов. Кроме того, выделение синонимов объектов выполняется экспертом на основе его знаний, интуиции и опыта, что ставит эффективность выполнения процесса в зависимость от конкретного специалиста [3]. При наличии спорных ситуаций экспертное мнение является трудно подтверждаемым обоснованием и может вступать в противоречие с другим экспертным мнением. В этих условиях выбор технологических решений, комбинация которых позволит достичь целей и будет соответствовать критериям эффективного использования финансовых и человеческих ресурсов, становится нетривиальной задачей для должностных лиц.

Применение метода частотного анализа при составлении словаря синонимов для объектов онтологии

Статистический подход в виде частотного анализа является одним из сравнительно простых методов NLP, который используется для определения частоты встречаемости слов или фраз в тексте и анализа полученных результатов [5]. Он может быть применен для любых типов текстов и базируется на том, что существует нестандартное статистическое распределение символов внутри текстовых массивов [6].

Метод частотного анализа текстов в литературе рассматривается в двух аспектах: как самостоятельный метод, результаты которого могут быть использованы в других методах, или как составляющая метода машинного обучения. На данном этапе мы рассматриваем встраивание частотного анализа в общую схему процесса создания объектов онтологии в качестве самостоятельного метода для улучшения выполнения отдельных операций. Извлечение из текстов ключевых слов с помощью частотного анализа является альтернативным вариантом формирования словаря синонимов для объектов онтологии. Это упрощает работу операторов-аналитиков и помогает быстрее и полнее анализировать тексты, находящиеся в базе данных ИАС «Семантический архив».

В общем виде процесс частотного анализа текстов представлен на рис. 5.

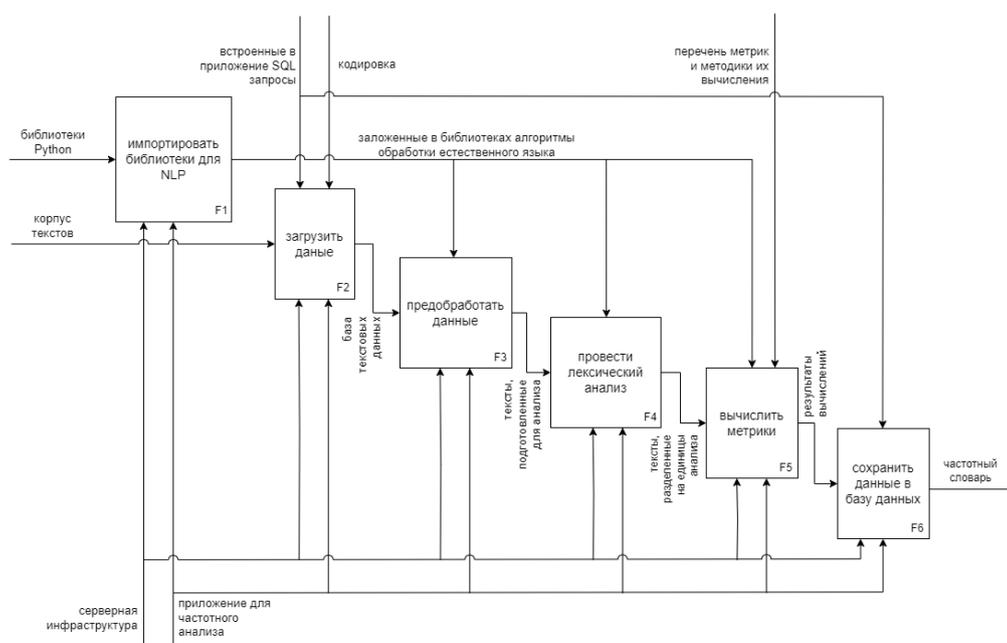


Рис. 5. Процесс частотного анализа текста

Предварительная обработка текста может включать удаление специальных символов, переносов строк, гиперссылок и т. д. Данный этап заключается в приведении текста к виду, позволяющему в большинстве случаев точно определять границы слов и предложений. Лексический анализ включает в себя несколько подпроцессов. Токенизация – разделение на текстовые единицы. В некоторых библиотеках Python токе-

низация проходит в два уровня: предложения, слова. Удаление стоп-слов – часто встречающихся слов, не несущих смысловой нагрузки (союзы, предлоги, местоимения). Лемматизация – приведение слов к их базовой форме. Разметка частей речи (POS). В некоторых случаях используется стемминг – выделение основы слова.

Основной метрикой для частотного анализа текста является частота вхождения слова – количество раз, которое слово встречается в тексте или корпусе текстов. На ее основе вычисляются ранговое распределение, относительная частота, индекс частотности. Для определения важности слова для конкретного текста вычисляется метрика TF-IDF (term frequency-inverse document frequency). Расчет частоты слов и других метрик может осуществляться как для каждого текста в отдельности, так и для всего корпуса текстов в целом. Кроме того, результаты расчетов будут различаться в случаях ориентации на леммы (словарная форма) или на стеммы (основа слова). Более сложным вариантом являются словари, учитывающие последовательность из двух (биграммы) или трех (триграммы) слов, которые в контексте лингвистического анализа текста используются для более детального анализа частотности и семантической связи между словами. На выходе мы получаем несколько вариантов частотного словаря: частотный словарь лемм, частотный словарь стемм, биграмм и триграмм в разрезе каждого проанализированного текста, а также их вариации суммарно для всего корпуса текстов.

Частотный анализ встраивается в общую схему процесса создания объекта мониторинга так, как это показано на рис. 6. При этом происходит изменение последовательности шагов в рамках процесса. Выявление ключевых слов осуществляется в начале, после чего выделенные слова получают привязку к объектам определенных классов. Словарь синонимов при таком подходе является результатом классификации элементов частотного словаря за счет установления связей с объектами мониторинга из онтологии ИАС «Семантический архив». Классификация осуществляется оператором-аналитиком в интерфейсе ИАС «Диссонанс» (рис. 7).

Оператор-аналитик осуществляет выбор потенциальных синонимов из частотного словаря и связывает их с доступными объектами онтологии или создает новый объект, соответствующий выбранному синониму. Кроме того, предусматривается возможность для дополнительной разметки словаря, например, с точки зрения эмоциональной окраски.

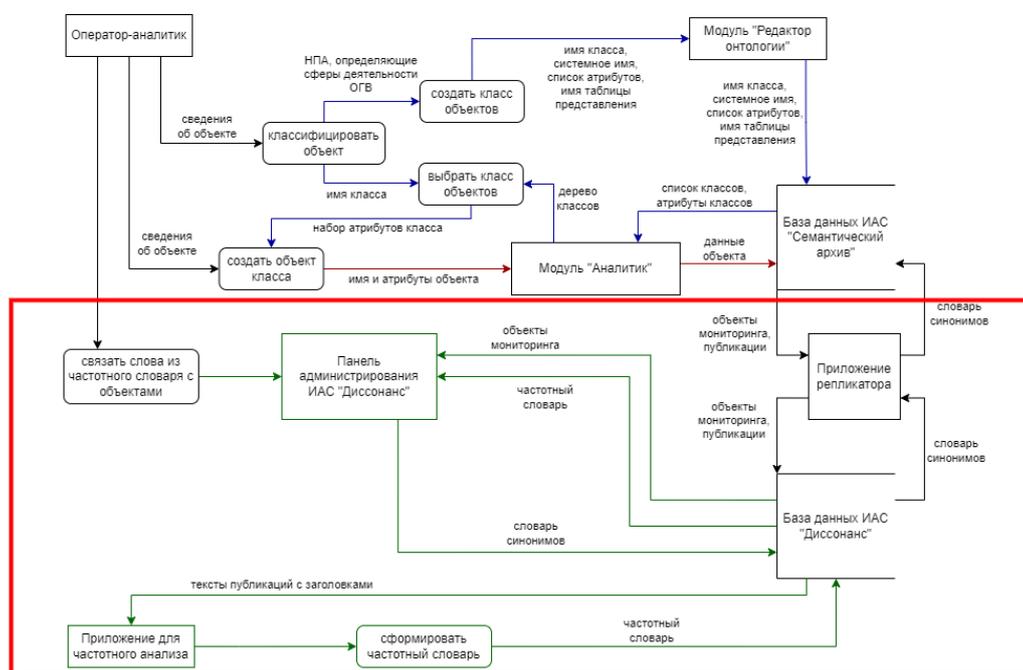


Рис. 6. Целевой вариант процесса создания объектов онтологии с включением частотного анализа

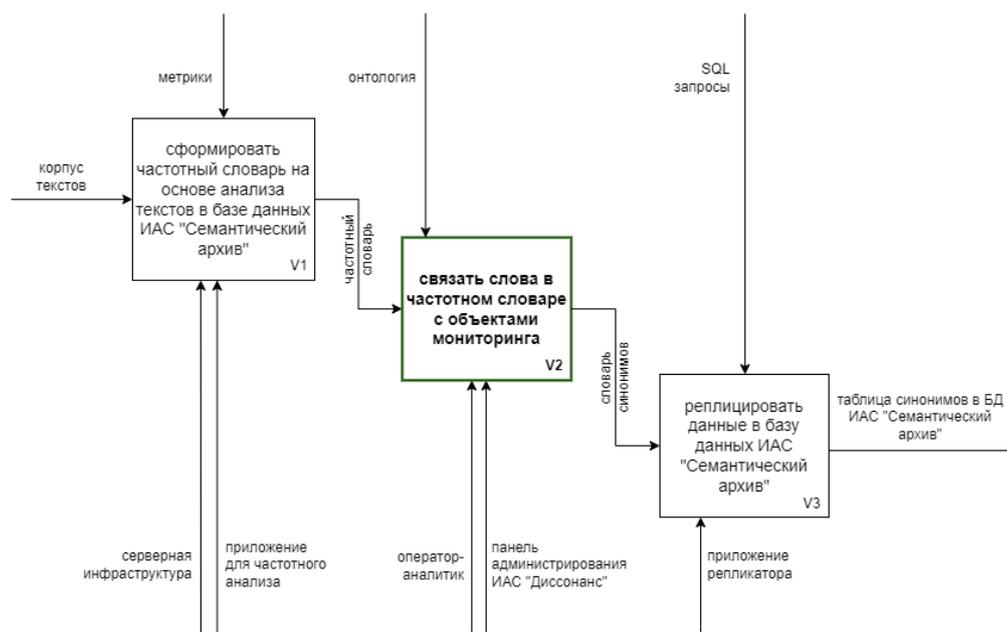


Рис. 7. Процесс создания словаря синонимов на основе частотного словаря

При этом количество действий, выполняемых оператором-аналитиком, существенно сокращается, что наглядно видно при сравнении рис. 4 и рис. 8. Изменения в процессе создания объектов сократят временные издержки и в конечном счете сведут их к минимуму, так как с увеличением базы знаний появление в ней новых сущностей становится все менее вероятным.



Рис. 8. Действия, выполняемые оператором-аналитиком в процессе создания синонимов объектов онтологии после внедрения механизма частотного анализа текста

Перспективы использования методов машинного обучения

Автоматизация оставшейся части процесса создания классов и объектов онтологии, которая выделена на рис. 9, может быть выполнена с помощью методов обработки естественного языка (NLP), которые в основном связывают с использованием технологий искусственного интеллекта. Частотный анализ как составляющая метода машинного обучения закладывает базу для последующего составления и обогащения онтологической модели предметной области на основе извлеченных признаков и сущностей, классификации текстов в рамках онтологии, анализа тональности, анализа авторства, кластеризации текстов для поиска скрытых зависимостей и других задач, выполняемых при работе с текстовыми данными. Выбор конкретного алгоритма зависит от задачи и доступных данных.

Теоретическую основу автоматической обработки текстов составляет компьютерная лингвистика, наиболее востребованы в которой мето-

ненных для моделирования тематик в тексте. Существуют алгоритмы и методы, которые могут решать одновременно задачи определения тематик, определения тональности и извлечения сущностей в тексте. Один из таких подходов – это использование предварительно обученных моделей, таких как BERT (Bidirectional Encoder Representations from Transformers) или GPT (Generative Pre-trained Transformer), которые могут быть дообучены на конкретной задаче. Методы семантического анализа, такие как Latent Semantic Analysis (LSA) или Word2Vec, могут быть использованы для определения семантической близости слов и текстов. Это может быть полезно, например, при поиске похожих документов или при кластеризации текстов по их семантическому содержанию.

Таким образом, применение методов машинного обучения позволит в автоматическом режиме выполнять вышеперечисленные задачи без постоянного вмешательства оператора-аналитика. При этом важно понимать, что автоматизация не может заменить полностью человеческое восприятие и анализ текстов, но позволит производить более полное извлечение знаний из большого объема собранных данных и сэкономить время и силы специалистов.

Заключение

Предварительная проработка модели бизнес-процессов и проведение обработки тестовой выборки документов с помощью наиболее популярных библиотек Python для частотного анализа текстов продемонстрировали перспективность разработки приложения для частотного анализа текстовой информации, агрегируемой в ходе осуществления информационно-аналитического обеспечения деятельности органов власти Республики Коми.

В целом, применение методов частотного анализа позволит повысить эффективность и точность лингвистической обработки текстовой информации, обеспечить автоматическую обработку всего корпуса текстов, накопленных в базе данных, с наиболее полным охватом объектов, упоминаемых в текстах, в том числе с учетом особенностей социальных сетей. Данные улучшения дают возможность снизить зависимость от экспертного уровня оператора-аналитика, сократить время, затрачиваемое операторами-аналитиками на выполнение рутинных задач при составлении базы знаний предметной области, выйти за пределы одной предметной области и, возможно, предложить услуги по анали-

зу текстовых данных коммерческим организациям и выйти за пределы предметной области государственного управления и в перспективе предложить услуги по анализу текстовых данных коммерческим организациям. Кроме того, снижаются риски, связанные с использованием сторонних разработок и санкционными ограничениями.

СПИСОК ИСТОЧНИКОВ

1. **Ширшов Е. В.** Информационно-аналитическое обеспечение менеджмента : учебное пособие по направлению подготовки бакалавров 38.03.02. «Менеджмент». М.: ИД «Академия Естествознания», 2022. 156 с.
2. **Гавердовский В. С.** Практическая эволюция информационно-аналитических систем управления регионом, создаваемых в ГАУ РК «Центр информационных технологий» в 2009–2016 годы // *ИТ Арктика*. 2017. № 1. С. 12–27.
3. Лучшие практики региональной информатизации «ПРОФ-ИТ.2014» : сборник [Электронный ресурс]. URL: <https://d-russia.ru/wp-content/uploads/2015/03/prof-it-2014.pdf> (дата обращения: 15.01.2023).
4. **Епифанцев Б. Н.** Информационно-аналитические системы безопасности: возможности использования ресурсов других специальностей для формирования лабораторной базы // *Информационное противодействие угрозам терроризма*. 2015. Т. 1. № 25. С. 159–166.
5. **Мухаметов М. Р.** Частотный анализ текста в Python // *Мавлютовские чтения : материалы XVI Всероссийской молодежной научной конференции : в 6 т. Уфа, 25–27 октября 2022 года*. Уфа: Уфимский государственный авиационный технический университет, 2022. Т. 5. С. 1054–1056.
6. **Преображенский А. П., Чопорова Е. И., Меняйлов Д. В.** Тематический анализ текстовой информации на основе частотных характеристик // *Цифровая обработка сигналов и её применение*

(DSPА-2022): 24-я Международная конференция, Москва, 30 марта – 01 апреля 2022 года. М.: Российское научно-техническое общество радиотехники, электроники и связи им. А. С. Попова, 2022. Вып. XXIV. С. 136–140.

References

1. **Shirshov E. V.** *Informatsionno-analiticheskoye obespecheniye menedzhmenta : uchebnoye posobiye po napravleniyu podgotovki bakalavrov 38.03.02. «Menedzhment»* [Information and analytical support for management : Textbook for bachelor’s training 03.38.02. “Management”]. Moscow: Publishing House “Academy of Natural Sciences”, 2022. 156 p. (In Russ.)
2. **Gavardovsky V. S.** Practical evolution of information and analytical systems for regional management created at the State Autonomous Institution of the Republic of Kazakhstan "Information Technology Center" in 2009–2016. *IT Arktika* [IT Arctic]. 2017. No 1. Pp. 12–27. (In Russ.)
3. *Luchshiyе praktiki regional’noy informatizatsii «PROF-IT.2014» : sbornik* [Best practices of regional informatization “PROF- IT.2014”: Collection] [Electronic resource]. Available at: <https://d-russia.ru/wp-content/uploads/2015/03/prof-it-2014.pdf> (accessed: 01.15.2023) (In Russ.)
4. **Epifantsev B. N.** Information and analytical security systems: possibilities of using resources of other specialties to form a laboratory base. *Informatsionnoye protivodeystviye ugrozam terrorizma* [Information counteraction to the threats of terrorism]. 2015. Vol. 1. No 25. Pp. 159–166. (In Russ.)
5. **Mukhametov M. R.** Frequency analysis of text in Python. *Mavlyutovskiye chteniya: Materialy XVI Vserossiyskoy molodezhnoy nauchnoy konferentsii : v 6 t. Ufa, 25–27 oktyabrya 2022 goda* [Mavlyutov readings: Materials of the XVI All-Russian Youth Scientific Conference. In 6 volumes, Ufa, October 25–27]. Ufa: Ufa State Aviation Technical University, 2022. Vol. 5. Pp. 1054–1056. (In Russ.)

6. **Preobrazhensky A. P., Choporova E. I., Menyailov D. V.** Thematic analysis of text information based on frequency characteristics. *Tsifrovaya obrabotka signalov i yeyo primeneniye (DSPА-2022): 24-ya Mezhdunarodnaya konferentsiya, Moskva, 30 marta – 01 aprelya 2022 goda* [Digital signal processing and its application (DSPА-2022): 24th International conference, Moscow, March 30 – April 1, 2022]. Moscow: Russian Scientific and Technical Society of Radio Engineering, Electronics and Communications named after A. S. Popova, 2022. Issue XXIV. Pp. 136–140. (In Russ.)

Сведения об авторах / Information about authors

Крашенинников Михаил Сергеевич / Mikhail S. Krasheninnikov
консультант / consultant

Государственное автономное учреждение Республики Коми «Центр информационных технологий» / State autonomous institution of the Komi Republic «Information Technology Center»

167000, Россия, г. Сыктывкар, ул. Интернациональная д.108 а / 167000, Russia, Syktyvkar, Internatsionalnaya str., 108 а

обучающийся магистратуры / master's student

Сыктывкарский государственный университет имени Питирима Сорокина / Pitirim Sorokin Syktyvkar State University

167001, Россия, г. Сыктывкар, Октябрьский пр., 55 / 167001, Russia, Syktyvkar, Oktyabrsky Ave., 55

Лавреш Иван Иванович / Ivan I. Lavresh

к.т.н., доцент кафедры информационной безопасности / Ph.D. in Technics, Associate Professor of the Information Security Department

Сыктывкарский государственный университет имени Питирима Сорокина / Pitirim Sorokin Syktyvkar State University

167001, Россия, г. Сыктывкар, Октябрьский пр., 55 / 167001, Russia, Syktyvkar, Oktyabrsky Ave., 55

Устюгов Владимир Александрович / Vladimir A. Ustyugov

к.ф.-м.н., доцент, заведующий кафедрой информационной безопасности / Ph.D. in Physics and Mathematics, Associate Professor, Head of the Information Security Department

Сыктывкарский государственный университет имени Питирима Сорокина / Pitirim Sorokin Syktyvkar State University

167001, Россия, г. Сыктывкар, Октябрьский пр., 55 / 167001, Russia,
Syktyvkar, Oktyabrsky Ave., 55

Статья поступила в редакцию / The article was submitted 29.08.2023

Одобрено после рецензирования / Approved after reviewing 12.09.2023

Принято к публикации / Accepted for publication 27.09.2023