на правах рукописи

Беляева Оксана Владимировна

Автоматическое восстановление структуры текстовых документов

Специальность 2.3.5 —
«Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Автореферат

диссертации на соискание учёной степени

кандидата технических наук

Москва — 2025

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте системного программирования им. В.П. Иванникова Российской Академии Наук.

Научный руководитель:	**Турдаков Денис Юрьевич,** кандидат физико-математических наук
Официальные оппоненты:	**Котельников Евгений Вячеславович**, доктор технических наук, доцент, профессор Автономной некоммерческой образовательной организации высшего образования «Европейский университет в Санкт-Петербурге». **Дородных Никита Олегович,** кандидат технических наук, старший научный сотрудник Федерального государственного бюджетного учреждения науки Институт динамики систем и теории управления имени В.М. Матросова Сибирского отделения Российской академии наук (ИДСТУ СО РАН).
Ведущая организация:	Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»

Защита состоится 17 апреля 2025 г. в 14:30 на заседании диссертационного совета 24.1.120.01 при Федеральном государственном бюджетном учреждении науки Институт системного программирования им. В. П. Иванникова Российской академии наук по адресу: 109004, г. Москва, ул. А. Солженицына, дом 25.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Институт системного программирования им. В. П. Иванникова Российской академиии наук.

Автореферат разослан “____” _____ 2025 г.

Ученый секретарь
диссертационного совета 24.1.120.01,
кандидат физико-математических наук

Зеленов С. В.

Общая характеристика работы

Актуальность темы. В условиях стремительного роста объема электронных документов, создаваемых в различных сферах деятельности, возникает острая потребность в их автоматической обработке с целью экономии человеческих ресурсов. Большинство документов представлены в неструктурированном виде, что требует применения интеллектуальных методов обработки документов для их структуризации.

Автоматический анализ информации из социальных сетей, интернета, сайтов, структурирование как открытых, так и закрытых баз знаний невозможно выполнить качественно без автоматического извлечения содержимого и структуры электронных текстовых документов в данных источниках.

Под анализом информации в данной работе понимается извлечение фрагментов текстовой и графической информации с последующей её структуризацией для целей хранения, организации поиска, вычисления статистических данных и обобщения результатов. В обработке электронных документов анализ информации сводится к анализу содержимого документов, а он в свою очередь невозможен без первоначального этапа - извлечения содержимого и восстановления структуры из документа.

Под структурой текстового документа понимается иерархическая структура, то есть иерархическое представление совокупности частей документа таким образом, чтобы его части располагались на соответствующих уровнях иерархии и была обеспечена возможность навигации с использованием разбивки документа на главы, секции, разделы и тому подобное.

Наличие информации о содержимом и структуре электронных документов облегчает их цифровую обработку. В первую очередь это требуется для информационно-аналитических систем, обеспечивающих сбор и поиск информации с последующей интеллектуальной обработкой содержимого документов. Знания об иерархической структуре документа способствуют решению таких задач, как обеспечение навигации по документу (восстановление оглавления), суммаризации (составление краткого описания к тексту) и фрагментации документа, проверки правильности составления документа, поиск по заголовкам, поиск связей внутри одного или нескольких документов.

Автоматическая обработка электронных текстовых документов является трудной задачей, поскольку документы могут быть представлены в различных форматах, таких как PDF, DOCX, HTML, изображений, а их структура и виды фрагментов содержимого могут существенно различаться в разных предметных областях и принятых там типах документов

(например, технические задания, законы, рекламные брошюры или исследовательские работы). Поэтому для качественного анализа информации необходимо учитывать особенности предметной области документа и его формат, задающий спецификацию хранения текстовой, графической, табличной информации и работы с ней.

К особенностям предметной области документа (свойствам типов документа) относят совокупность правил составления документа: правила составления структуры содержимого документа, правила визуального оформления (форматирования) содержимого, тематика текстового содержимого. Качество разработанных методов в диссертационной работе демонстрируется для трех разных типов документов “Техническое задание”, “Выпускная квалификационная работа” и “Нормативно-правовой акт”, но спектр применимости методов гораздо шире и не ограничивается только данными типами.

Среди широкого набора разнообразных форматов электронных документов можно выделить две основные группы. Документы могут быть представлены форматами, такими как PDF с текстовым слоем, HTML, DOCX, и т.д. Такие форматы являются ***структурированными***, то есть в них содержатся структурные теги, позволяющие выделить в документах заголовки разного уровня, списки, таблицы, данные о форматировании. При этом в каждом формате внутренняя разметка (теги) и их виды определены по-разному.

Кроме того, существуют форматы ***неструктурированных*** данных, например, изображения или PDF-документы, содержащие страницы, являющиеся сканированными копиями напечатанных на бумаге или написанных от руки документов. Такие документы легко воспринимаются человеком, но плохо поддаются автоматическому анализу, поскольку не содержат ни текстовой (копируемый текст), ни структурной (встроенные в формат теги/разметка) информации о содержимом документа.

Область автоматического извлечения содержимого и восстановления структуры документов различных форматов, в частности неструктурированных форматов изображений и PDF остается по сей день вызовом для систем автоматического интеллектуального анализа текстовых электронных документов.

Объектом исследования являются текстовые электронные документы различных предметных областей в виде структурированных и неструктурированных форматов документов.

Предметом исследования выступают методы автоматического извлечения содержимого и восстановления структуры из исследуемых текстовых документов.

Степень разработанности темы. Исследования в области обработки электронных документов неструктурированных форматов активно ведутся уже более двадцати лет.

Важные результаты были получены в работах Mao S., Namboodiri A., где авторы отметили важность проблемы извлечения содержимого и восстановления структуры документов. В последнее время чаще появляются работы для обработки изображений сканированных документов с использованием нейронных сетей для разного рода задач, например для сегментирования страницы документа (Binmakhashen G, Eskenazi S.), табличной обработки (Schreiber S., Zhong X., Gao L.).

Несколько авторов (Михайлов А.) исследует автоматическую обработку некорректных PDF-документов, а также причин, приводящих к нарушению корректности. Некоторые работы авторов, такие как Gonesh C, James H исследуют вопросы около данной темы, например классификация текстов на корректность.

Множество последних исследований в области восстановления иерархической структуры документов формата PDF стало возможным благодаря международному соревнованию FINTOC по обработке финансовых документов, где участники применяют современные методы и подходы на основе машинного обучения.

Целью исследования является разработка методов и расширяемого программного средства для автоматического извлечения содержимого и восстановления структуры из электронных текстовых документов. Разрабатываемые методы и программные средства должны удовлетворять следующим требованиям:

1. Обеспечение автоматического анализа информации для текстовых документов технических заданий, нормативно-правовых актов, выпускных квалификационных работ;
2. Возможность расширения средств извлечения содержимого и восстановления структуры для новых форматов документов и документов новых предметных областей.

Для достижения поставленной цели решаются следующие **задачи**:

1. Разработать метод автоматического извлечения содержимого PDF документов с использованием проверки текстового слоя, обеспечивающий достоверность извлечения и скорость обработки документов;
2. Разработать метод автоматического восстановления структуры из содержимого текстовых документов;
3. Реализовать предложенные методы в виде программного комплекса, обладающего возможностью расширения новыми форматами и типами текстовых документов.

Научная новизна заключается в следующих результатах работы:

1. Метод автоматического извлечения содержимого PDF документов с использованием проверки текстового слоя, обеспечивающий достоверность извлечения и скорость обработки документов на русском и английском языках;
2. Метод автоматического восстановления иерархической структуры из содержимого документов. Метод показывает более высокое качество восстановления структуры, по сравнению с другими методами на наборе данных соревнования FINTOC2022;

Теоретическая и практическая значимость. Теоретическая значимость диссертации заключается в разработке и усовершенствовании методов извлечения содержимого и восстановления структуры документов неструктурированных форматов в автоматическом режиме. В рамках диссертации разработан метод, позволяющий с большей точностью восстанавливать иерархическую структуру, что подтверждено измерениями на наборе данных FINTOC2022. Важными результатами диссертации являются новые методы автоматической обработки документов в формате PDF. Увеличивает теоретическую ценность работы рассмотрение и использование нейросетевых методов в построенной обработке изображений сканированных документов, что позволяет достигать высокого качества извлечения текстовой информации.

В плане практической значимости важным результатом является открытый программный комплекс для автоматического извлечения содержимого и восстановления иерархической структуры текстовых электронных документов различных форматов и предметных областей, который может быть использован в качестве первоначального этапа для систем автоматической интеллектуальной обработки электронных документов. Внедрения и ПО с открытым доступом:

- интеграция в открытую библиотеку LangChain[1];
- внедрение в платформу Талисман, предназначенную для построения интеллектуальных информационно-аналитических систем сбора и обработки данных;
- внедрение в систему анализа выпускных квалифицированных работ.

Результаты диссертации применимы для разработчиков информационно-аналитических систем, предназначенных для структуризации и анализа сырых необработанных данных, в том числе электронных документов.

Методология и методы исследования. В диссертационной работе применялись методы обработки изображений, машинного обучения, теории вероятностей и оптимизации.

[1] https://github.com/langchain-ai/langchain/releases/tag/langchain-community%3D%3D0.2.10

Основные методы исследования включают анализ существующих решений, разработку и экспериментальное исследование алгоритмов.

Основные положения выносимые на защиту:

1. Метод автоматического извлечения содержимого PDF документов с использованием проверки текстового слоя, обеспечивающий достоверность извлечения и скорость обработки документов;
2. Метод автоматического восстановления иерархической структуры из содержимого текстовых документов;
3. Архитектура и реализация расширяемого программного комплекса в виде открытой библиотеки DEDOC[2] для автоматического извлечения содержимого и восстановления иерархической структуры из электронных текстовых документов структурированных и неструктурированных форматов.

Апробация работы. Результаты работы докладывались на конференциях, форумах:

1. IVMEM2019 Международная конференция ”Иванниковские чтения 2019”, Великий Новгород, 2019, РФ;
2. FNP 2021 The 3rd Financial Narrative Processing Workshop, 2021, Marseille, France; LREC;
3. IVMEM2022 Международная конференция ”Иванниковские чтения 2022”, 2022, Казань, РФ;
4. ISPRAS OPEN 2022 Открытая конференция ИСП РАН им. В.П. Иванникова, Москва, РФ;
5. AINL: Artificial Intelligence and Natural Language Conference, 2023, Ереван, РА;
6. ISPRAS OPEN 2023 Открытая конференция ИСП РАН им. В.П. Иванникова, 2023, Москва, РФ;
7. IVMEM2024 Международная конференция ”Иванниковские чтения 2024”, 2024, Великий Новгород, РФ;
8. DataFest 2024, в гостях у VK, 2024, Москва, РФ;
9. Гравитация. Международная университетская премия в области искусственного интеллекта и больших данных, 2024, Москва, РФ.

Публикации и личный вклад автора. Автор имеет 10 научных публикаций по теме диссертации. Работы [8, 9, 10] индексируются в Scopus и Web of science. Основные результаты по теме диссертации изложены в 8 печатных изданиях, 5 из которых [5-8, 10]

[2] https://github.com/ispras/dedoc

изданы в журналах, рекомендованных ВАК. Остальные 5 работ опубликованы по результатам конференций. В работах [1-8] автором проведено исследование предметной области, выполнен основной объем теоретических и экспериментальных исследований. В работах [2, 3, 6] Беляевой О.В. и Козлову И. принадлежит постановка задачи, разработка подхода и анализ экспериментов. Работы [1, 4, 5, 7, 9, 10] выполнены под непосредственным руководством Беляевой О.В. В работе [5] автором разработан подход и метод исправления ориентации, разработка экспериментов и анализ результатов проводилась совместно с соавторами. Работа [7] выполнена полностью автором, редакторские правки и анализ результатов выполнялись совместно соавторами. По теме диссертации имеется 3 свидетельства о государственной регистрации программы для ЭВМ [11-12].

Предлагаемые в диссертации инструменты, текстовые наборы данных и исследования разработаны и выполнены автором или при его непосредственном участии.

Внедрение результатов. Результаты, полученные в рамках данной работы, внедрены в следующих организациях:

1. Внедрены в систему “Киберпрофессор” анализа выпускных квалификационных работ студентов (акт о внедрении № 20/01-6 от 06.02.2025) ;
2. Внедрены в состав платформы Талисман, которая используется в ООО “Интерпроком” (акт о внедрении № 18/25 от 28.01.2025);
3. Внедрены в состав платформы Талисман, которая используется в федеральном государственном автономном образовательном учреждении высшего образования «Московский государственный институт международных отношений (университет) Министерства иностранных дел Российской Федерации» (акт о внедрении от 7.02.2025);
4. Внедрены в сервис распознавания изображений документов в ЗАО “ЕС Лизинг”, что подтверждает официальное письмо № ЕСЛ-36 от 10.02.2025).

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках диссертационной работы, формулируются цель и основные задачи, перечисляются основные положения выносимые на защиту, формулируется научная новизна и практическая значимость диссертационной работы.

Первая глава посвящена обзору области автоматического извлечения содержимого и восстановления структуры документов разных типов и форматов. Особое внимание

уделяется рассмотрению существующих методов и систем для автоматического восстановления структуры из текстовых документов и автоматической обработке изображений сканированных документов. В данном разделе рассматриваются основные проблемы обработки PDF-документов в автоматическом анализе, а также представлено подробное исследование причин, влияющих на извлечение некорректного текста из копируемых PDF-документов.

В большинстве существующих алгоритмов и систем автоматического восстановления структуры текстовых электронных документов используется двухэтапная обработка. Первый этап - это извлечение содержимого из документов согласно специфике обрабатываемого формата. Второй этап - это восстановление структуры документа из содержимого с форматированием, полученного на первом этапе.

В разделе 1.1 рассмотрена исследуемая в диссертации предметная область документов, состоящая из таких типов документов как, “Техническое задание” (ТЗ), “Нормативно-правовой акт” (НПА), “Выпускная квалификационная работа” (ВКР). Также выделены особенности исследуемой предметной области изображений сканированных документов.

В разделе 1.2 работы уделяется внимание обзору существующих методов и систем извлечения содержимого из текстовых документов различных форматов такой как текст, табличная информация и стилевое форматирование текста. При этом полнота извлекаемой информации ограничивается либо сложностью обрабатываемого формата, либо функциональностью внешних используемых доступных библиотек. Под сложностью обрабатываемого формата понимается насколько формат способен содержать в себе разметку (теги). Отсюда в работе представлена классификация форматов документов от сложных до простых в задачах автоматической обработки: не-структурированные (форматы изображений, PDF) и структурированные форматы документов (офисные форматы DOC/DOCX, HTML). В области некоторые исследователи выделяют также категорию слабо-структурированного формата, куда относят копируемые PDF, полученные путем конвертации из структурированного формата (например DOCX) и тем самым хранящие разметку текстового содержимого. В разделе также рассматриваются такие проблемы, как *автоматическая работа с PDF* документами с некорректным текстовым слоем, а также PDF документами не содержащими текстовый слой. Автоматическая обработка такого вида неструктурированных данных представляет собой сложный процесс, требующий использования методов машинного обучения и применения отдельных решений

искусственного интеллекта для извлечения текстовой, табличной информации и набора стилевого форматирования.

В разделе 1.3 представлен обзор методов автоматического восстановления иерархической структуры из полученного содержимого со стилевым форматированием с предыдущего этапа. В данном разделе рассматривается проблема разнообразия предметных областей (типов) электронных текстовых документов. Здесь рассмотрены различные научные работы, применяющих как классические эвристические подходы, так и методы машинного обучения.

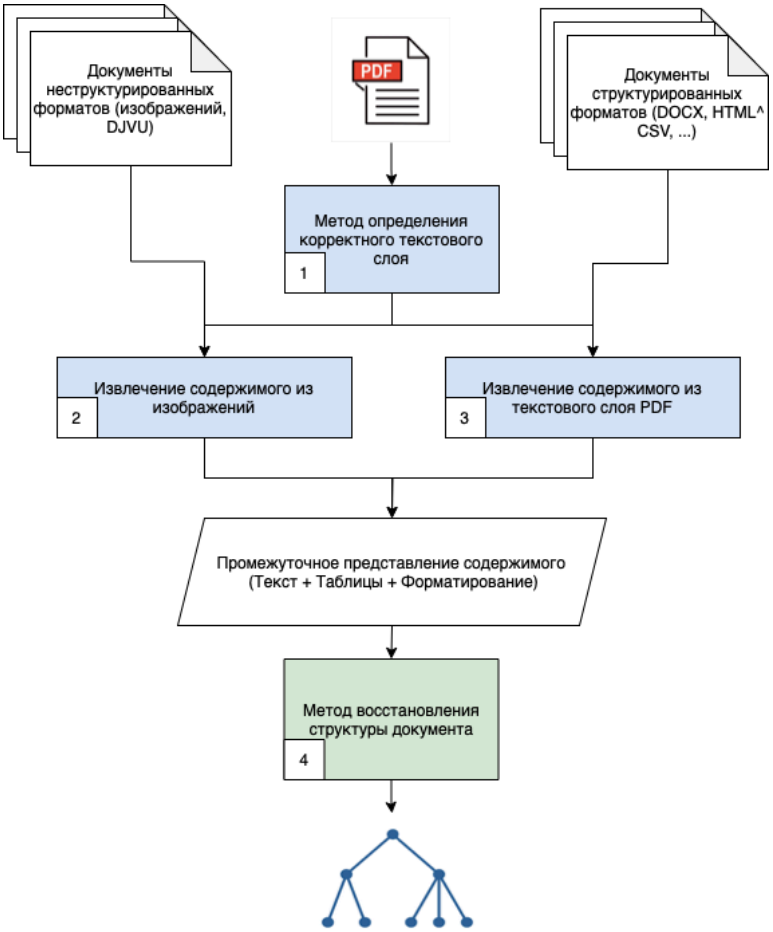

Рисунок 1 - Общая схема обработки электронных текстовых документов.

В рамках проведенного обзора автоматической обработки документов, было решено создать двухэтапную обработку с применением отдельных разработанных решений:

1. Разработанный метод автоматического извлечения содержимого PDF документов с использованием проверки текстового слоя, обеспечивающий достоверность извлечения и скорость обработки документов (блок 1, рисунок 1). Данный метод использует разработанные методы обработки изображений сканированных

документов (блок 2, рисунок 1) или внешнюю библиотеку для обработки PDF (блок 3, рисунок 1).

2. Разработанные методы обработки изображений сканированных документов для извлечения содержимого (блок 2, рисунок 1).
3. Разработанный метод восстановления иерархической структуры из извлеченного содержимого (блок 3, рисунок 1).

Во второй главе описаны новые методы, положенные в основу построения программного комплекса автоматической обработки текстовых документов. Представлено подробное описание разработанных методов обработки документов неструктурированных форматов, а именно:

1) Метод автоматического определения корректности текстового слоя PDF-документов. Метод включает разработанный бинарный классификатор определения корректности извлекаемого текста. На основании предсказания классификатора определяется каким способом эффективно обрабатывать PDF-документ. Время и качество обработки PDF методом измерялось на реальных данных с использованием разработанного программного комплекса.
2) Методы для автоматической обработки изображений сканированных документов. Методы применяются последовательно, а именно методы предварительной обработки изображений (исправление угла поворота, ориентации изображений), метод обнаружения и распознавания табличной информации и распознавания текста.
3) Метод восстановления иерархической структуры из содержимого документов. Метод основан на классификации текстовых строк с использованием “градиентного бустинга” деревьев решений. Разработанный метод показал лучшее качество на наборе данных международного соревнования FiNTOC 2022.

Метод автоматического определения корректности текстового слоя PDF-документов. Цель метода - эффективно извлечь текстовое содержимое из PDF-документов. Разработанный метод берет на себя функцию автоматического определения корректности входного PDF-файла. В методе решается каким из двух способов следует эффективно извлечь из него текстовое содержимое (Рисунок 2) в автоматической обработке. В случае, если текстовый слой PDF-файла поврежден (некорректен), то следует обработать PDF-файл как набор изображений сканированных документов через разработанные методы, описанный ниже. В случае, если текстовый слой PDF-файла не пустой и является корректным, то эффективнее извлечь текстовое содержимое путем анализа инструкций вывода на печать формата PDF с помощью внешней библиотеки. Таким образом, задача

сводится к бинарной классификации (корректен/некорректен) копируемый текст текстового слоя PDF. В данной работе используется внешняя PDF-библиотека для получения текста из PDF.

Учитывая особенность предметной области обрабатываемых документов. Первая страница в PDF-документе может быть представлена в виде изображения без текстового слоя (сканированная страница), в то время как остальные страницы PDF документа имеют текстовый слой и получены путем конвертации из хорошо структурированных форматов, таких как офисные форматы DOC/DOCX. По описанной причине, текст первой страницы анализируется и обрабатывается в методе отдельно от остальных страниц PDF документов.

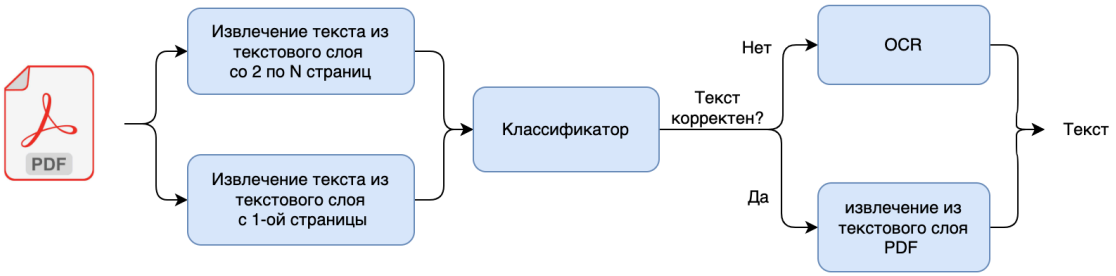

Рисунок 2 - Схема метода автоматической обработки текстовых PDF-документов.

Формальная постановка задачи выглядит следующим образом. Извлеченный с помощью PDF библиотеки g текст x_N из первых N страниц документа d можно задать как $x_N = g(d)$. Результат бинарного классификатора f тогда задается выражением $f(g(d)) \in [0;1]$, где 0 - корректный, 1 - некорректный текст. Таким образом, описанный алгоритм можно представить следующей формулой:

$$document = \begin{cases} correct & \text{if } f(x_N) = 0 \& x_N \notin \emptyset \\ incorrect & \text{otherwise} \end{cases}$$

где, $x_N \notin \emptyset$ это не пустой текст из первых N текстовых страниц документа d.

В рамках исследования модели классификации представлено сравнение разных моделей машинного обучения на собранных реальных и синтетическом наборах данных. Сравнение различных моделей классификации проводилось на двух типах входных признаках извлекаемых из текста:

1. Извлечение n-грамм из текста с последующим применением меры TF-IDF;
2. На эвристических правилах, извлекаемых из текста.

Синтетический набор данных для обучения классификатора f генерировался двумя способами, на основе текстов на русском и английском языках, полученных из Wikipedia:

1) Декодирование текста заведомо неправильной кодировкой;
2) Сохранения текста на изображении и распознавание оптическим распознаванием текста (Optical Character Recognition - OCR) методом с заведомо неправильным языком распознавания;

Таким образом, был сгенерирован сбалансированный синтетический набор данных мощностью в 15 тысяч текстов.

Для сравнения моделей классификации был собран несбалансированный тестовый набор реальных данных, состоящий из 257 PDF документов объемом в среднем в 51 страницу, где 233 документа корректных, остальные 24 некорректные PDF. Некорректных PDF-документов меньше, чем корректных, что соответствует дисбалансу данных в реальных условиях.

Результаты сравнения моделей классификации на двух типах признаков на синтетическом наборе данных представлено в таблице 1. Для вероятностных моделей MLE, Logistic Regression порог классификации подбирался на валидационной выборке.

Обучение моделей проводилось на тренировочном наборе синтетических данных. Время, представленное в Таблице 1 состоит из вычисления признаков классификации и времени вычисления модели. По результатам (Таблица 1) модели показывают схожую высокую точность, но по времени быстрее всего обработка происходит с Logistic Regression и MLE. Согласно результатам (Таблица 2) на реальных собранных данных, наиболее эффективными (качество, время) являются методы XGBoost на эвристических признаках, MLE на биграммах. В качестве классификатора была выбрана модель деревьев решений XGBoost на эвристических признаках, поскольку она не требует подбора вероятностного порога на валидационной выборке.

Таблица 1 - Результат моделей классификации на тестовом синтетическом наборе текстов (на тестовой выборке).

Модель	**Признаки**	**F1-score**	**Time (ms)**
Logistic Regression	TF-IDF	0.990	**1.221**
	Custom features	0.810	1.452
Random Forest	TF-IDF	0.998	9.593
	Custom features	0.998	10.322
XGBoost	TF-IDF	0.998	7.104
	Custom features	0.998	6.276
RuBert	BERT Features	**1.000**	82.213

MLE	Unigram	0.986	2.013
	Bigram	0.996	2.535
	Trigram	0.996	2.921
MLE with Laplace	Unigram	0.985	1.996
	Bigram	0.991	2.665
	Trigram	0.980	2.891

Таблица 2 - Результат моделей классификации на тестовом наборе реальных текстов.

Models	**Features**	**Precision**	**Recall**	**F1**
Logistic Regression	TF-IDF	0.922	0.868	0.886
	Custom features	0.897	0.739	0.791
Random Forest	TF-IDF	0.923	0.872	0.889
	Custom features	0.957	0.953	0.955
XGBoost	TF-IDF	0.916	0.837	0.863
	Custom features	0.961	0.961	0.961
RuBert	BERT Features	**0.970**	**0.965**	**0.966**
MLE	Unigram	0.940	0.942	0.932
	Bigram	0.940	0.938	0.939
	Trigram	0.942	0.930	0.934
MLE with Laplace	Unigram	0.949	0.946	0.935
	Bigram	0.952	0.953	0.952
	Trigram	0.955	0.957	0.955

Метод был успешно интегрирован в итоговый программный комплекс, описанный в главе 3. Существует два основных способа извлечения текстовой информации из файла PDF: использование технологии OCR и непосредственное чтение его текстового слоя с помощью внешних библиотек PDF. Предложенный метод автоматически определяет наличие и корректность текстового слоя и решает, как обрабатывать PDF-файл.

Чтобы оценить влияние классификатора на общее качество извлечения текста, был собран двухстраничный набор данных, включающий 30 корректных PDF-документов и 30 некорректных, каждый из которых в среднем составлял 2 страницы. Исходный текст был подготовлен вручную для каждого документа, и каждый документ был обработан с помощью трех режимов обработки:

1. Обработка только методами извлечения содержимого из изображений сканированных документов (с OCR) (блок 2 на Рисунке 1);
2. Обработка только внешней PDF-библиотекой извлечения текста из текстового слоя PDF (блок 3 на Рисунке 1);

3. Обработка с разработанным методом автоматического определения корректности текстового слоя (блок 1 на Рисунке 1).

Таблица 3 - Среднее время извлечения содержимого одной страницы PDF-документа в разработанном программном комплексе Dedoc.

Набор данных	Время извлечения содержимого с использованием OCR (обработка режимом 1)	Время извлечения содержимого с использованием разработанного метода (обработка режимом 3)	Время извлечения содержимого с использованием PDF-библиотеки (обработка режимом 2)
Двухстраничный набор данных	3.336 сек/стр.	2.665 сек/стр.	1.260 сек/стр.
Набор данных реальных документов	3.374 сек/стр.	0.616 сек/стр.	0.167 сек/стр.

Таблица 4 - Точность извлеченного текста в различных режимах обработки PDF-документа в разработанном программном комплексе Dedoc на двухстраничном наборе данных.

Character Accuracy (Mean Levenshtein ratio) с использованием OCR (обработка режимом 1)	Character Accuracy (Mean Levenshtein ratio) с использованием разработанного метода (обработка режимом 3)	Character Accuracy (Mean Levenshtein ratio) с использованием PDF-библиотеки извлечения текстового слоя (обработка режимом 2)	Точность бинарной классификации (метрика Accuracy)
0.906	**0.939**	0.589	0.9

Результаты тестов представлены в Таблицах 3 и 4. В Таблице 3 представлено время выполнения каждого режима обработки на реальном наборе данных, состоящем в среднем из 51 страницы. Режим обработки 3 с разработанным методом в 5 раз быстрее режима 1 с OCR, при этом по точности превосходит на 3.3% Character Accuracy. Таким образом, разработанный метод показывает наилучшее время и качество обработки PDF-документов.

В Таблице 5 приведено сравнение результатов обработки некорректных PDF документов из размеченного набора. Как видно из Таблицы 5, существующие открытые системы, которые умеют обрабатывать структурированные и неструктурированные форматы документов плохо обрабатывают некорректные PDF документы.

Таблица 5 - Точность извлеченного текста разными системами для некорректных PDF документов с размеченным текстом из набора двухстраничных документов.

Открытые системы для автоматической обработки PDF документов	Среднее значение Character Accuracy (Mean Levenshtein ratio)
Разработанный программный комплекс	**0.91**
docling	0.097
unstructured	0.08

Методы автоматического извлечения текстового содержимого из изображений сканированных документов. Работа содержит обзор решений обработки сканированных документов. После обзорной части, приведено описание каждого этапа обработки изображений в программном комплексе с замером качества распознавания на собранных или сгенерированных наборах. Набор методов обеспечивает извлечение текстовой и табличной информации для распространенного домена сканированных финансовых, юридических и технических документов. Страницы таких документов характеризуются белым фоном, темным текстом, таблицами с явными границами. Документы рассматриваются на русском и английском языках. Ориентация страниц документов может быть разная 0, 90, 180, 270 градусов. Текст может быть многоколоночный. Макет страниц типа манхеттен, когда текстовые блоки расположены параллельно друг к другу. Итоговая точность распознавания текста на страницах разработанными методами приведена в таблице 6 на собственно размеченном наборе данных.

Точность извлекаемого текста (Character Accuracy) в программном комплексе замеряется с использованием расстояния Левенштейна для каждой страницы:

$$CharacterAccuracy = \frac{n - E}{n}$$

, где n - число символов на странице, E - число ошибок распознавания текста на странице (количество операций редактирования удаления/вставки по Левенштейну).

В разработанном программном комплексе можно изменить обработчик изображений документов на другой с применением других методов распознавания.

Таблица 6 - Точность распознанного текста с применением методов обработки сканированных документов.

Набор данных	**Количество изображений документов**	**Character accuracy[3]**
Черно-белые сканированные изображения документов	83	97.541

Метод восстановления иерархической структуры. Для второго этапа двухэтапной обработки текстовых документов был разработан метод автоматического восстановления иерархической структуры из содержимого с форматированием, полученного на предыдущем этапе (блок 4 на рисунке 1).

Иерархическая структура документа - это иерархическое представление структурных частей документа так, чтобы части документа имели свой уровень в иерархии и имелась возможность навигации по документу.

В разработанном методе выделяются два типа структурных элементов:

- На уровне заголовков. Здесь восстанавливается иерархия заголовков в документе. Примером иерархии заголовков документа является “Оглавление”.
- На уровне текстовых блоков. Здесь восстанавливается иерархия текстового содержимого внутри секций документа. Сюда относятся разделение на параграфы и списки. Приоритеты таких структурных элементов ниже заголовочных.

Пример иерархической структуры документа представлен на рисунке 3, где ближе к корню иерархического дерева расположены структурные элементы с уровнем небольшим. В листьях дерева расположены такие, как элементы списка или логически неделимый обычный текст (параграфы).

Метод восстановления структуры состоит из следующих шагов:

1. Обнаружение строк ‘Оглавления’. Данный шаг применяется только для документов с Оглавлением, представленным в начале документа. К таким документам относятся ВКР (Диплом / магистерская диссертация), технические задания, набор данных соревнования FINTOC2022.

[3] S. V. Rice. *Measuring the Accuracy of Page-Reading Systems*. Ph.D. dissertation,41 University of Nevada, Las Vegas, 1996

2. Получение машиночитаемых численных свойств текста (матрицы признаков) на основе текстовых регулярных выражений и его визуальных свойств в виде матрицы признаков.
3. Обучение модели ИИ (классификатора строк) на полученной матрице признаков, используемой в качестве входных данных модели. Модель должна классифицировать строки документа согласно конкретному типу структуры (согласно предметной области). Структура документа определяет список классов строк. Каждый класс определяется своим приоритетом в предметной области. В качестве классификатора использовался метод градиентного бустинга над деревьями решений XGBoost.
4. Постобработка результатов предсказаний модели. На данном этапе уточняются тип структурных элементов согласно регулярным выражениям. Например, классификатор определяет класс заголовка как “named_item”, а далее можно уточнить вид данного заголовка (глава или статья). Таким образом, обеспечивается возможность изменить приоритет структурных элементов без переразметки тренировочного набора данных. Код постобработки отличается для каждой предметной области (типа) документа.
5. Построение иерархии, определения уровня структурных элементов на основе полученных приоритетов строк.

Рисунок 3 - Пример представления иерархической структуры документа “Жилищный кодекс”.

В методе выделяется три группы признаков:

- *Не-относительные признаки форматирования*. В эту группу относятся признаки форматирования, не вычисляемые относительно других строк с форматированием: полужирный текст, курсив, зачеркнутый текст, подчеркнутый текст, цвет текста;
- *Относительные признаки форматирования*. Значения признаков вычисляются для всех строк документа, далее вычисляется медиана значений, и значением признака принимается величина строки относительно медианы. К таким относительным признакам относятся: отступ строки слева (расстояние), межстрочное расстояние, размер шрифта;
- *Текстовые признаки строки*. К данной группе относятся признаки на основе регулярных выражений, а именно булевые признаки срабатывания регулярных выражений. Здесь определяется уровень вложенности для списковых строк.
- *Дополнительные признаки.* Индикатор того, что строка фигурирует в оглавлении документа.

Для оценки метода были размечены наборы данных документов с предметными областями (типами) “Нормативно-правовые акты” (НПА), “Технические задания” (ТЗ) и “Выпускные квалификационные работы” (ВКР). Каждый набор характеризуется своим набором классов строк, описание которых представлено в таблице 7 вместе с количественной информацией. В технических заданиях и выпускных работах также присутствуют списки на уровне текстовых блоков. Собранные наборы содержат размеченные документы различных форматов DOCX, PDF, TXT. Разметка данных заключалась в задании класса для текстовых строк документов с помощью внешней системы разметки.

Дополнительно был разработан бинарный классификатор текстовых строк (строка 4 в таблице 9) для определения является ли текстовая строка началом нового параграфа в документе или нет. Таким образом, делается объединение строк в текстовые блоки, иными словами, происходит выделение параграфов. Стоит отметить, что классификатор параграфов применяется только для неструктурированных или слабоструктурированных документов PDF и форматов изображений, в силу того, что у хорошо структурированных форматов, таких как DOCX, HTML информация о начале нового параграфа задана в формате документа. Например, для DOCX начало параграфа задает тэг <pPr>, а в HTML параграф оформлен в теги <p> или .

Точность обученных классификаторов строк с постобработкой результатов метода представлена в таблице 8. Точность измерялась с использованием перекрестной проверки на 10 итерациях.

Для форматов изображений был разработан бинарный классификатор текстовых строк на начало параграфа/продолжение параграфа. Классификатор позволяет выделять параграфы в тексте. Классификатор параграфов работает по таким же признаках, что и другие классификаторы структур.

Для обработки документов из неизвестных предметных областей, в которых может не быть оглавления, создан отдельный обработчик. Здесь, применяется бинарная классификация для обнаружения параграфов и на этапе постобработки выделяется иерархическая структура с помощью списковых регулярных выражений. Таким образом, если неизвестна предметная область документа, то путем выделения параграфов и базовой списковой иерархии извлекается минимальная иерархическая структура из содержимого документа.

Таблица 7 - Описание используемых наборов данных документов.

Тип документа	**Количество документов**	**Количество строк**	**Распределение строк по типам (классам)**
Закон РФ	349	20471	титульный лист: 2058 заголовок: 4104 приложение: 696 сноска: 266 автор: 261 текст: 13086
Техническое задание	51	8556	титульный лист: 208 заголовок: 892 пункт: 2891 содержание: 677 текст: 3888
Диплом / магистерская диссертация	50	25545	титульный лист: 1159 заголовок: 1074 текст: 23312
Начало параграфа/продолжение параграфа	50	7502	начало параграфа: 2710 продолжение параграфа: 4472

Таблица 8 - Точность классификации строк по типу (предметным областям) документов.

Тип документа	**Точность классификации (accuracy)**	**F1-мера (macro)**
Закон РФ	0.91057	0.81411

Техническое задание	0.88010	0.80836
Диплом / магистерская диссертация	0.95152	0.94894
Начало параграфа/продолжение параграфа	0.91326	0.91491

Точность разработанного метода была также оценена (Таблица 9) на общедоступном наборе данных соревнования FINTOC2022, содержащего финансовые документы на английском языке. Документы набора имеют формат PDF с качественным текстовым слоем. В соревновании решается две задачи:

- Задача 1: детекция заголовков в документе;
- Задача 2: определение уровня вложенности заголовков;

Для решения каждой задачи используется разработанный метод, а именно извлечение содержимого с форматированием, извлечение признаков каждой текстовой строки, обнаружение оглавления документа, получение матрицы признаков, предсказание классификации, постобработка. Различие в том, что для первой задачи на выходе классификатора вычисляется 2 класса (заголовок/не заголовок), а для второй задачи классификатором предсказывается уровень вложенности заголовка от 0 до 9. Поскольку разметка набора FINTOC содержит только иерархическую структуру из заголовков (нет структуры внутри секций), то для решения второй задачи участвовали не все текстовые строки, а только заголовочные.

Итоговое сравнение для двух задач (задача “1” и “2”) разработанного метода с другими решениями на наборе данных FINTOC2022 на английском языке представлено в Таблице 9.

Таблица 9 - Сравнение с другими методами на наборе данных FINTOC2022 English.

	F1-мера обнаружения заголовка (Задача 1)	**Inex-F1-мера определения уровня заголовка (Задача 2)**	**Level Accuracy точность определения уровня заголовка (Задача 2)**	**Гармоническое среднее F1-меры и точности Level Accuracy определения уровня заголовка (Задача 2)**
Christopher Bourez (2021)	0.830	53.6	30.6	38.95

CILAB	0.738	56.5	27.5	36.99
swapUNIBA	0.793	63.6	42.9	51.23
Предложенный метод	**0.900**	**68.8**	**58.4**	**63.17**

В задачах 1 и 2 правильно обнаруженными заголовками является те, которые совпали по расстоянию Левенштейна более 0.85. Правильно обнаруженным заголовком является:

- для задачи 1: текста заголовков (обнаруженного и в разметке) совпали по расстоянию Левенштейна более 0.85 и по номеру страницы;
- для задачи 2: текста заголовков (обнаруженного и в разметке) совпали по расстоянию Левенштейна более 0.85 , по номеру страницы и уровню заголовка в разметке (от 0 до 9).

Precision это отношение правильно предсказанных заголовков к общему числу найденных заголовков в документе. Recall - это отношение корректно предсказанных заголовков к числу заголовков в разметке. F1 - гармоническое среднее между Precision и Recall. Точность второй задачи определяется по формуле:

$$LevelAccuracy = \sum \frac{E'_{ok}}{E_{ok}}$$

, где E_{ok} - это количество предсказанных заголовков с совпадающей страницей из разметки, E'_{ok} - это количество предсказанных заголовков с совпадающей страницей и уровнем вложенности совпадающей с разметкой.

Гармоническое среднее (harmonic mean) определение уровня заголовка вычисляется между F1-мерой и точностью Level Accuracy определения уровня заголовка.

В третьей главе описана архитектура программного комплекса, обладающая свойствами расширяемости. В главе описана методика расширения программного комплекса для добавления поддержки новых форматов и типов документов. Кроме того в главе представлены особенности сборки программного комплекса в виде python-библиотеки, так и в виде api-сервиса.

Программный комплекс. Особенностью программного комплекса является его расширяемость за счет возможности добавления новых форматов документов и новых

предметных областей документов для восстановления иерархической структуры. Описание архитектура системы представлена в разделах 3.1 и 3.2.

Общая схема архитектуры представлена на Рисунке 4 и состоит 4-х основных расширяемых компонент:

- *Конвертирование* (при необходимости) полученного документа в один из форматов, поддерживаемых программным комплексом. Таким образом, избегается проблема создания большого количества обработчиков при поддержке большого перечня форматов документов;
- *Обработка для извлечения содержимого и метаданных (форматирования)* из документа в промежуточное представление. На данном этапе выбирается необходимый обработчик, который извлекает содержимое (текст, таблицы) с форматированием из конкретного формата документа.
- *Восстановление* из промежуточного представления документа иерархической *структуры* в зависимости от предметной области документа: определение типа (класса) каждого элемента (строки) и его значимости (уровня вложенности) внутри документа;
- *Преобразование* промежуточного представления структурированного документа *в выходной формат* согласно древовидному представлению.

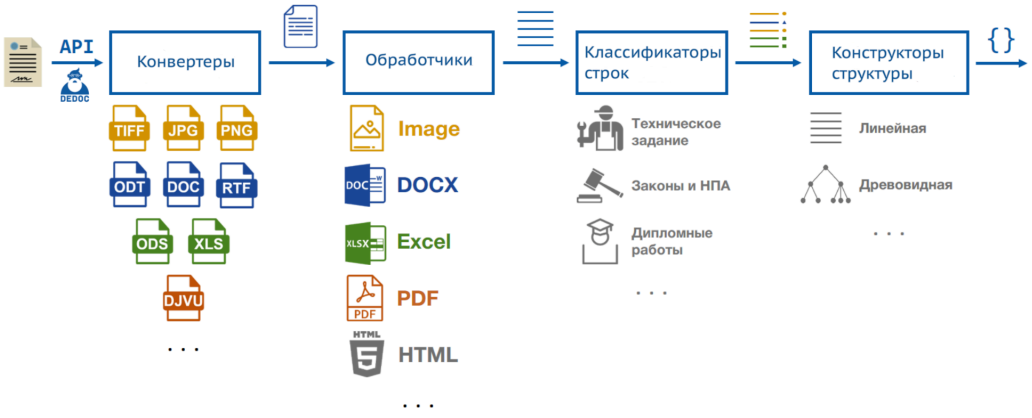

Рисунок 4 - Общая схема архитектуры программного комплекса.

В разделе 3.3 описана методика для добавления поддержки нового формата документа и типа структуры документа.

В разделе 3.4 представлено описание внешнего API-интерфейса системы, используемый механизм контейнеризации и возможность использования системы как

внешнюю Python-библиотеку. Примеры обработанных документов представлены на Рисунке 5.

В разделе 3.5 представлено описание внутреннего представления документа в системе. Выходной формат обрабатываемых документов унифицирован и может быть представлен в форматах HTML, json. Выход также представляется в виде иерархического дерева, где глубина узлов дерева задается глубиной восстановленной иерархической структуры документа.

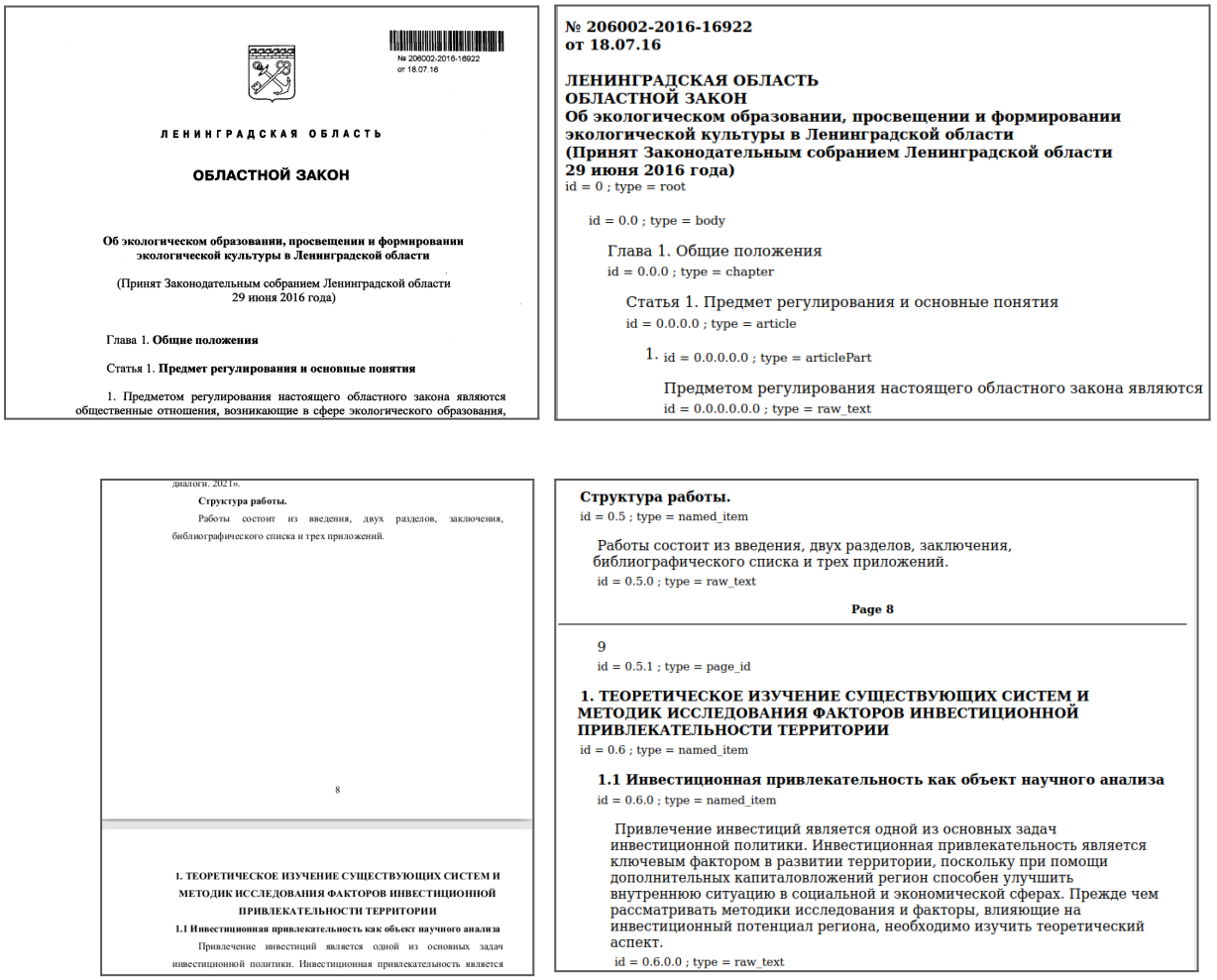

Рисунок 5 - Примеры извлечения содержимого и восстановления структуры разработанными методами и программным комплексом. В примере выше обработано изображение сканированного документа типа НПА. В примере ниже обработан ВКР формата PDF.

В **заключении** приведены основные результаты работы:

1. Разработан обладающий научной новизной метод автоматического извлечения содержимого PDF-документов с использованием проверки текстового слоя, обеспечивающий достоверность извлечения и скорость обработки документов.

Благодаря методу точность извлечения текста повышается на 3.3% (Character Accuracy), а время обработки сокращается более чем в 5 раз по сравнению с обработкой изображений сканированных документов;

2. Разработан обладающий научной новизной метод восстановления иерархической структуры из содержимого документов. Метод демонстрирует высокое качество на размеченных документов трех типов и показывает лучшие результаты на наборе данных международного соревнования FINTOC;
3. Разработана расширяемая архитектура программного комплекса, которая позволяет добавлять поддержку обработки новых форматов и типов структур документов. Архитектура позволяет обрабатывать документы в автоматическом режиме и приводить обрабатываемые документы к единому унифицированному виду;
4. На основе разработанной архитектуры и методов реализован программный комплекс в виде открытой библиотеки/системы, позволяющий автоматически обрабатывать документы разных форматов и типов структур с целью извлечения их содержимого и восстановления структуры в едином унифицированном виде. Внедрения программного комплекса подтвердили актуальность и практическую значимость диссертации.

Публикации автора по теме диссертации

1. Belyaeva O. Dedoc: A Universal System for Extracting Content and Logical Structure From Textual Documents / Belyaeva O., Bogatenkova A., Turdakov D. // 2023 Ivannikov Ispras Open Conference (ISPRAS). — IEEE, 2023. — P. 20-25.
2. Anastasiia Bogatenkova. ISPRAS@FinTOC-2022 Shared Task: Two-stage TOC Generation Model / Anastasiia Bogatenkova, Oksana Vladimirovna Belyaeva, Andrew Igorevich Perminov, Ilya Sergeevich Kozlov. // In Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022, Marseille, France. European Language Resources Association. — 2022. — P. 89–94.
3. Kozlov I. Ispras@ fintoc-2021 shared task: Two-stage toc generation model / Kozlov I. Belyaeva. O., [et al.] // Proceedings of the 3rd Financial Narrative Processing Workshop. — 2021. — P. 81-85.
4. Belyaeva O. V. Automatic verification of the text layer correctness in PDF documents / Belyaeva O. V., Golodkov A., Bukhatov B. // 2024 Ivannikov Memorial Workshop (IVMEM). — IEEE, — 2024. — P. 1-7.

5. Golodkov A.O. Real Application of CNN Interpretation Methods: Document Image Classification Model Errors' Detection and Validation / Golodkov A.O., Belyaeva O.V., Perminov A.I. // Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS). — 2023. — Vol 35. — P. 7-18. — (ВАК).
6. Bogatenkova A. O. A.I. Logical structure extraction from scanned documents / Bogatenkova A. O. Kozlov I. S., Belyaeva O. V., Perminov // Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS). — 2020. — Vol 32. — P. 175-188. — (ВАК).
7. Belyaeva O.V. Synthetic data usage for document segmentation models fine-tuning / Belyaeva O.V., Perminov A.I., Kozlov I.S. // Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS). — 2020. — Vol 32. — P. 189-202. — (ВАК).
8. Perminov A.I. Loss functions for train document image segmentation models / Perminov A.I., Turdakov D.Yu., Belyaeva O.V. // Programming and Computer Software. – 2023. — Vol 49. — P. 574-589. — (ВАК, WoS).
9. M. S. Akopyan. Text Recognition on Images from Social Media / M. S. Akopyan, O. V. Belyaeva, T. P. Plechov and D. Y. Turdakov // *2019 Ivannikov Memorial Workshop (IVMEM),* Velikiy Novgorod, Russia. — 2019. — P. 3-6. — (WoS).
10. A. O. Bogatenkova. Generation of Images with Handwritten Text in Russian / A. O. Bogatenkova, O. V. Belyaeva, A. I. Perminov // Programming and Computer Software. — 2024. — Vol 50. — P. 483-492. — (ВАК, Scopus).

Свидетельства о государственной регистрации программы для ЭВМ

1. Puredoc: сервис обработки изображений документов / Беляева О.В., Богатенкова А.О., Перминов А.И., Голодков А.О., Шевцов Н.С., Рахматуллаев Т.А., Михайлов А.А., Зыкин Я.И. ; ФГБУН Институт системного программирования РАН. — No 2023688256; заявл. 15.12.2023 (Рос. Федерация).
2. Docreader / Козлов И.С., Беляева О.В., Богатенкова А.О., Перминов А.И.; ФГБУН Институт системного программирования РАН. — No 2020666950; заявл. 21.12.2020 (Рос. Федерация).
3. Dedoc / Козлов И.С., Беляева О.В., Богатенкова А.О., Перминов А.И.; ФГБУН Институт системного программирования РАН. — No2020667079; заявл. 21.12.2020 (Рос. Федерация).