




МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО
ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное
образовательное учреждение высшего
образования
«Национальный исследовательский университет
ИТМО» (Университет ИТМО)

Кронверкский пр-т, д. 49, лит. А,
Санкт-Петербург, Россия, 197101
Тел.: (812) 480-00-00 | Факс: (812) 232-23-07
od@itmo.ru | itmo.ru

16.03.2026 № 40-09-21

УТВЕРЖДАЮ
Проректор по научной работе,
доктор технических наук, профессор

В. О. Никифоров
 2026 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

на диссертацию Беляевой Оксаны Владимировны на тему
“Автоматическое восстановление структуры текстовых документов”
представленную на соискание ученой степени кандидата технических наук по
специальности 2.3.5 “Математическое и программное обеспечение
вычислительных систем, комплексов и компьютерных сетей”

Актуальность темы диссертации

Диссертационная работа Беляевой О. В. посвящена исследованию и разработке методов автоматического извлечения содержимого и восстановления иерархической структуры текстовых электронных документов различных форматов и предметных областей. В условиях стремительного роста объемов неструктурированных данных все больше уделяется внимание исследованиям в области автоматической обработки, анализа многостраничных документов и их структуризации, что особенно важно в контексте увеличения числа прикладных систем, ориентированных на автоматический анализ большого массива данных.

Особую значимость приобретает развитие методов восстановления иерархической структуры документов. Иерархическое представление документов зачастую является неотъемлемой частью в решении ряда прикладных задач в области интеллектуального анализа документов, позволяет оптимизировано хранить и эффективно работать с документами.

Одним из наиболее распространенных форматов является PDF, автоматическая обработка которого вызывает трудности, поскольку формат ориентирован на корректную отрисовку содержимого при печати, а не на структурированное хранение информации.

Отсюда возникает потребность в исследовании методов, повышающих качество извлечения содержимого и его структуризации из плохо структурированных форматов документов, в том числе формата PDF.

В связи с этим, исследования, представленные в диссертационной работе Беляевой О. В., приобретают особую актуальность. В работе рассмотрены ключевые проблемы и предложены инновационные методы автоматической обработки PDF-документов, а также восстановления иерархической структуры из содержимого документов. Эти разработки вносят значительный вклад в развитие области автоматической обработки документов, предлагая решения, которые отвечают современным вызовам и способствуют повышению эффективности работы с большими массивами документов.

Структура и содержание диссертации

Диссертация состоит из введения и трех глав, заключения, списка литературных источников из 106 наименований и 3 приложений. Общий объем работы 152 страниц текста.

Во введении обоснована актуальность темы диссертации, сформулирована цель, задачи, показаны научная новизна, практическая ценность результатов работы, приведены основные положения, выносимую на защиту.

В первой главе приведен обзор состояния области автоматической обработки электронных документов, описан подход обработки, описаны основные методы извлечения содержимого документов, восстановления структуры, анализа PDF-документов, приведены результаты анализа причин некорректности содержимого PDF.

Вторая глава включает описание нового метода автоматической обработки PDF-документов с проверкой корректности текстового слоя для повышения качества и сокращения времени обработки PDF. Также показана общая схема автоматической обработки документов, включающая предложенный метод. В главе описан новый метод восстановления иерархической структуры текстовых документов, который использует расширенное признаковое пространство и учитывает оглавление в документах. В главе приведены экспериментальные обоснования предложенных методов, которые показывают лучшее качество обработки документов на разных наборах данных.

В третьей главе описывается расширяемая архитектура программного комплекса, включающего разработанные методы. В главе описана разработанная методика расширения программного комплекса для поддержки новых форматов и типов документов. Дополнительно приводится описание технической составляющей программного комплекса: внешний интерфейс, документация, особенности запуска.

В заключении подведены итоги и обобщены результаты проведенных исследований диссертации.

Научная новизна полученных результатов

К важнейшим результатам диссертационной работы, обладающих научной новизной, относятся следующие:

1. Предложен новый метод автоматического извлечения содержимого PDF документов с использованием проверки текстового слоя, обеспечивающий достоверность извлечения и скорость обработки документов. Данный метод впервые предлагает более эффективную обработку PDF по сравнению с базовыми решениями, что позволяет обходить проблему извлечения некорректного текста из PDF.

2. Предложен новый метод автоматического восстановления иерархической структуры из содержимого документов, который превосходит по качеству другие аналогичные решения на наборе данных международного соревнования FINTOC. Также метод демонстрирует высокие показатели качества на разных видах документов.

Значимость полученных результатов для науки и практики

Представленные в диссертации результаты могут быть использованы для разработки интеллектуально-аналитических систем анализа электронных документов, в которых могут решаться задачи поиска по актуальным данным, сбор статистической информации и другие задачи анализа данных, в том числе с использованием больших языковых моделей.

Практическая ценность диссертации подтверждается актами внедрения в бюджетные и коммерческие организации. Дополнительно, стоит отметить, что разработанный программный комплекс является открытым и расширяемым для обработки новых предметных областей и форматов документов за счет разработанной архитектуры и методики расширения.

Достоверность полученных результатов

Обоснованность положений диссертации подтверждается результатами проведенных экспериментов и анализом эффективности разработанных методов, а также многочисленной апробацией на конференциях и научных мероприятиях всероссийского и международного уровней и научными публикациями, среди которых три индексируются в базах Scopus и Web of Science. Кроме того, получено три свидетельства о государственной регистрации программ для ЭВМ.

Замечания

По диссертации имеются следующие замечания:

1. Метод определения корректности текстового слоя в PDF документах обучался и тестировался на русских и английских текстах, при этом не раскрыта его применимость и расширяемость для работы с документами на других языках.

2. В Таблице 2.1.6.3, где приведено сравнение аналогичных систем, также стоило бы описать параметры запуска для каждой системы.

3. Помимо описания выходного представления документа в разделе 3.5, было бы уместно привести примеры выходных представлений документа в HTML и JSON форматах.

Тем не менее указанные замечания не снижают положительную оценку работы и не уменьшают высокой значимости полученных результатов.

Заключение

Диссертационная работа Беляевой Оксаны Владимировны “Автоматическое восстановление структуры текстовых документов” является законченным научным исследованием по актуальной теме. В работе представлены результаты, имеющие важное научное и практическое значение и соответствуют специальности 2.3.5 - “Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей”, а ее автор Беляева О. В. заслуживает присуждения ученой степени кандидата технических наук.

Диссертационная работа обсуждалась на семинаре исследовательского центра в сфере искусственного интеллекта «Сильный искусственный интеллект в промышленности» федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет ИТМО». Настоящий отзыв рассмотрен и одобрен на семинаре исследовательского центра в сфере искусственного интеллекта «Сильный ИИ в промышленности», протокол № 4 от 23 марта 2026 г. На заседании присутствовали 8 человек.

Отзыв составил:

Научный руководитель исследовательского центра в сфере искусственного интеллекта «Сильный искусственный интеллект в промышленности»,
доктор техн. наук

Бухановский А.В.

Сведения о ведущей организации:


федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО» (Университет ИТМО).

Почтовый адрес: 197101, г. Санкт-Петербург, пр-т Кронверкский, д. 49, лит. А

Телефон: (812) 490 0000

Веб-сайт:

Адрес эл. почты: info@itmo.ru

Подпись 
УДОСТОВЕРЯЮЩАЯ
ЗАМЕСТИТЕЛЬ
ПРОЕКТНОГО
УНИВЕРСИТЕТА
