УДК 004.91

Подорван Владислав Сергеевич, магистрант,
Амурский государственный университет

АВТОМАТИЗИРОВАННАЯ ПРОВЕРКА PDF-ДОКУМЕНТОВ НА СООТВЕТСТВИЕ ТРЕБОВАНИЯМ ОФОРМЛЕНИЯ

Аннотация. Работа посвящена разработке программного подхода к автоматизированной проверке электронных текстовых документов в формате PDF на соответствие установленным требованиям оформления. Рассматривается метод анализа документов, основанный на последовательной сегментации содержимого, восстановлении абстрактного синтаксического дерева и применении высокоуровневых правил проверки.

Ключевые слова: Нормоконтроль документов, автоматизированная проверка документов, PDF, анализ структуры документа, абстрактное синтаксическое дерево, семантический анализ.

В современных условиях цифровизации образовательной и профессиональной деятельности значительное количество текстовых документов создаётся и распространяется в электронном виде. Несмотря на развитие средств автоматизации документооборота, к оформлению таких документов по-прежнему предъявляются требования, определяющие их структуру, последовательность расположения элементов и параметры форматирования. Соблюдение данных требований особенно важно при подготовке выпускных квалификационных работ, пояснительных записок, технической документации и различных отчётов.

Проверка соответствия документов установленным требованиям зачастую выполняется вручную. Данный процесс предполагает анализ большого количества параметров: наличия обязательных разделов, правильности оформления заголовков, абзацев, таблиц и рисунков, а также взаимного расположения структурных элементов. При увеличении объёма проверяемых документов ручной нормоконтроль становится трудоёмким и может сопровождаться субъективными ошибками.

Целью данной работы является разработка программного подхода к автоматизированной проверке PDF-документов, позволяющего анализировать их структуру, оформление и отдельные семантические характеристики.

В современных системах электронного документооборота широкое распространение получили форматы DOCX и PDF. Первый из них позволяет хранить сведения о структуре документа, стилях, абзацах, таблицах и других элементах. Благодаря этому автоматизированный анализ DOCX может выполняться с использованием содержащейся в файле структурной информации. Однако при преобразовании документа в PDF часть логических сведений может быть утрачена.

Формат PDF ориентирован преимущественно на сохранение визуального представления документа. Его содержимое может включать текстовые и графические объекты, информацию о шрифтах и инструкции по размещению элементов на странице. При этом логические связи между объектами не всегда представлены явно, что усложняет определение принадлежности текстовых фрагментов к заголовкам, абзацам, таблицам и рисункам [1].

Для извлечения информации из PDF существуют специализированные программные библиотеки. В рамках данной работы используется библиотека pdfminer.six, позволяющая получать текстовое содержимое документа, координаты символов и текстовых блоков, а также характеристики используемых шрифтов [2]. Полученные данные служат исходной информацией для последующего восстановления структуры документа.

Предлагаемый подход к автоматизированному нормоконтролю основан на трёх взаимосвязанных этапах обработки. На первом этапе выполняется сегментация документа, в ходе которой определяются крупные структурные части, такие как введение, содержание, заключение и список использованных источников. Для обнаружения таких элементов используется текстовое представление PDF и регулярные выражения, задаваемые во внешней JSON-конфигурации.

На втором этапе выполняется восстановление абстрактного синтаксического дерева (AST), отражающего структуру содержимого каждой страницы. В отличие от предыдущего этапа, здесь анализируются не только крупные разделы, но и вложенные элементы: абзацы, таблицы, рисунки, подписи и другие блоки. Построение дерева осуществляется на основе конфигурационного описания узлов. Каждый узел определяет собственные параметры обнаружения и возможную структуру дочерних элементов. Такой подход позволяет рекурсивно восстанавливать иерархию содержимого документа с сохранением информации о расположении элементов на страницах.

На третьем этапе к восстановленному структурному представлению применяются высокоуровневые правила проверки. Они позволяют анализировать как отдельные характеристики элементов, так и взаимосвязи между ними. Например, может проверяться наличие подписи к рисунку, её расположение относительно соответствующего изображения, последовательность структурных блоков и соответствие оформления установленным требованиям. Таким образом, проверка не ограничивается анализом отдельных параметров форматирования, а учитывает контекст и структуру документа.

Результатом работы алгоритма является структурированное представление документа и набор обнаруженных несоответствий. При этом проверка выполняется не только на заключительном этапе: ошибки могут возникать уже при поиске обязательных разделов или восстановлении структуры отдельных страниц. Например, отсутствие требуемого раздела выявляется при сегментации, невозможность обнаружения ожидаемого элемента фиксируется при построении AST, а нарушение правил расположения подписи относительно рисунка определяется на этапе высокоуровневой проверки. Все выявленные ошибки объединяются в общий массив, каждый элемент которого содержит идентификатор, описание нарушения, сведения о затронутых элементах и пояснение.

Описанный подход позволяет разделить процесс проверки на независимые по назначению, но взаимосвязанные этапы. Сегментация определяет общую структуру документа, построение AST формирует представление его внутреннего содержимого, а высокоуровневые правила позволяют оценивать соответствие полученной структуры требованиям оформления. При этом использование внешней JSON-конфигурации обеспечивает возможность задавать проверяемые элементы и условия их обнаружения отдельно от программной реализации алгоритма. Это упрощает адаптацию подхода к различным требованиям без необходимости изменять основные механизмы обработки PDF.

Таким образом, разработанный подход направлен на решение задачи автоматизированного нормоконтроля PDF-документов в условиях отсутствия явно заданной логической структуры. Сочетание текстового анализа, восстановления иерархии элементов и проверки связей между ними позволяет выявлять как структурные, так и оформительские несоответствия. Предложенная организация алгоритма может служить основой для создания программных средств, сокращающих объём ручной проверки и обеспечивающих более последовательное применение требований к оформлению документов.

Список литературы:

1. ISO 32000-1:2008. Document management – Portable document format – Part 1: PDF 1.7. – Geneva: International Organization for Standardization, 2008. – 756 с.

2. pdfminer.six Documentation. Converting a PDF file to text [Электронный ресурс]. – Режим доступа: https://pdfminersix.readthedocs.io/en/latest/topic/ converting_pdf_to_text.html (дата обращения: 15.11.2026).