УДК 004.91

Подорван Владислав Сергеевич, магистрант,
Амурский государственный университет

АВТОМАТИЗИРОВАННАЯ ПРОВЕРКА ДОКУМЕНТОВ WORD НА СООТВЕТСТВИЕ СТАНДАРТУ ОФОРМЛЕНИЯ

Аннотация. Работа посвящена разработке программного обеспечения автоматизированной проверки оформления документов Microsoft Word на соответствие стандарту.

Ключевые слова: Нормоконтроль документов, проверка соответствия формальному шаблону, pdf.

В процессе исторического развития общества формы фиксации и передачи информации претерпели значительные изменения: от письменности и бумаги до печатного станка, пишущих машинок и современных цифровых технологий обработки текста. Независимо от формы представления, к оформлению документа всегда предъявлялись требования, обеспечивающие его структурированность, единообразие и соответствие нормам форматирования.

В современных условиях данные требования сохраняют свою значимость, однако развитие информационных технологий и рост автоматизации многих процессов привели к значительному увеличению объёмов документооборота. При этом проверка соответствия документов установленным стандартам оформления по-прежнему часто выполняется вручную. Этот процесс является трудоёмким и требует значительных временных затрат, особенно при обработке больших объёмов документов. Кроме того, ручная проверка сопровождается риском субъективных ошибок и неоднородности оценивания, что снижает эффективность контроля и создаёт определённые трудности для организаций.

Собственно, цель данной работы заключается в разработке программного обеспечения автоматизированной проверки документов Word на соответствие стандартам оформления.

В качестве предметной области выступают методы автоматизированной проверки оформления документов на соответствие стандартам оформления, предметом – является оформление документа, а объектом – текстовые документы, созданные в Microsoft Word.

На сегодняшний день, в современных системах документооборота широкое распространение получили форматы DOCX и PDF.

Формат DOCX позволяет хранить полную информацию о структуре текста, стилях, абзацах, заголовках, таблицах и рисунках, что делает его удобным для автоматизированного анализа и проверки соответствия стандартам оформления. В большинстве систем анализа таких документов применяется многоуровневая проверка структуры и форматирования документа. Алгоритм обычно включает последовательный обход абзацев, проверку стилей текста на уровне абзаца и слова, а также анализ параметров форматирования – шрифта, размера, интервала, выравнивания и отступов [1] Такой подход позволяет выявлять несоответствия стандарту и формировать отчёт о нарушениях. Однако, как правило, такие системы не учитывают последовательность блоков в рамках документа и зависимости их оформления от расположения на странице. А также, многие решения ориентированы на конкретный тип документа, что ограничивает универсальность алгоритмов.

Формат PDF, напротив, используется преимущественно для финальной версии документа и обеспечивает неизменность визуального представления. Содержимое страницы в PDF представлено в виде неупорядоченного набора различных объектов: текстовых блоков, графических примитивов, шрифтов, изображений и инструкций по их отображению на странице. Что делает анализ сложнее из-за отсутствия явной логической структуры.

Следует отметить, что промышленные решения для интеллектуального анализа PDF-документов, как правило, имеют сложную архитектуру и нередко являются коммерческими разработками. Алгоритмы восстановления структуры документа, классификации блоков текста и определения параметров форматирования часто представляют собой коммерческую тайну.

В тоже время существуют различные библиотеки парсинга PDF-файлов. В рамках данной работы остановимся на pdfminer.six. Он позволяет извлекать текст, координаты каждого символа, шрифтовые атрибуты (размер, начертание, стиль), а также информацию о структуре страниц.

Собственное решение будет построено на основе pdfminer.six. Будут использованы его эвристики для поиска структурных элементов на странице, а затем с помощью конечного автомата будут формироваться высокоуровневые структурные элементы на подобие абзацев, причислений, рисунков и так далее. В общем виде разбиение страницы на структурные элементы приведено на рисунке 1.

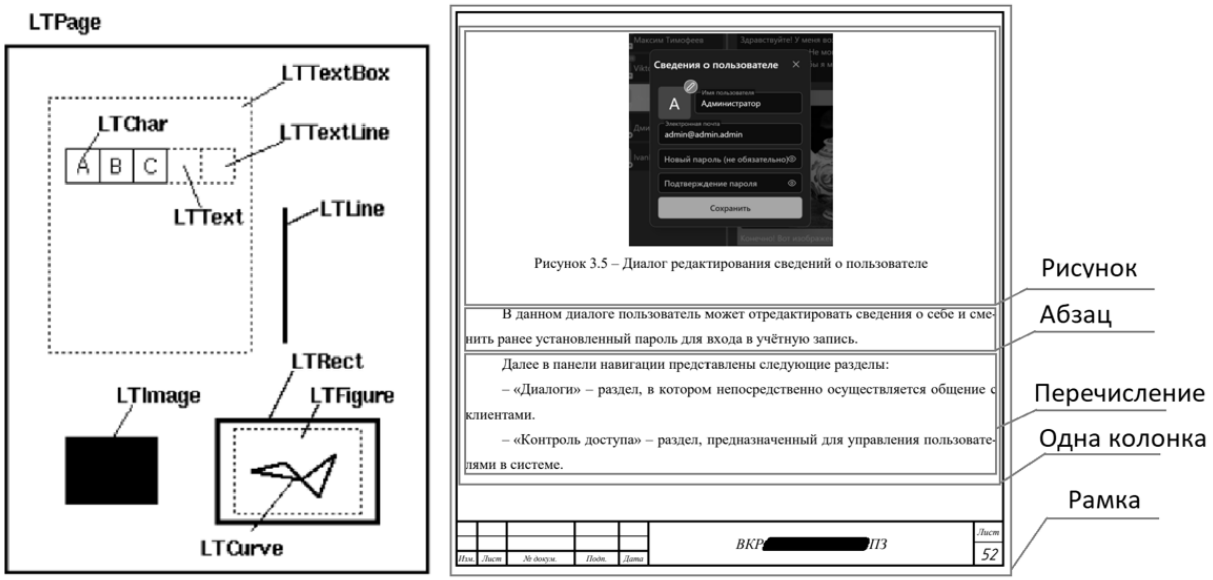


Рисунок 1. Структурная декомпозиция pdf файла (левый фрагмент изображения взят из [2])

Практическая значимость исследования определяется возможностью использования разработанного программного решения в различных организациях для повышения качества подготовки документов и оптимизации процессов нормоконтроля.

Список литературы:

1. Стариченко, Б. Е. Программа автоматизации контроля оформления текстовых документов / Стариченко Борис Евгеньевич, Устинов Максим Андреевич // Педагогическое образование в России. – 2018. – № 8. – С. 163-168.

2. Converting a PDF file to text [Электронный ресурс]. – Режим доступа: https://pdfminersix.readthedocs.io/en/latest/topic/converting_pdf_to_text.html#topic-pdf-to-text-layout. – 22.02.2026.