УДК 004.912

РАЗРАБОТКА АЛГОРИТМОВ ВЫДЕЛЕНИЯ И ИЗВЛЕЧЕНИЯ СТРОК, ПАРАГРАФОВ, ИХ СВОЙСТВ В PDF ДОКУМЕНТАХ

Марцинкевич В.И. (Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»),
Терещенко В. В. (Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»),
Бережков А. В. (Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»)
Научный руководитель - доцент, Горлушкина Н. Н. (Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»)

Введение. PDF - формат текстового документа, созданный компанией Adobe и который планировалось повсеместно использовать для печати файлов, содержимое которых не зависит от операционной системы или текстового редактора.
PDF формат представляет собой бинарный документ, состоящий из 4 элементов: заголовок, тело, xref таблица, trailer. Основой всего контента, содержащегося в документе, является тело, со множеством объектов, которые и характеризуют текстовое и графическое содержание файла [1].
В связи со сложностью формата возникает множество проблем при извлечении данных из него. Ключевой проблемой является отсутствие информации о структурных элементах в документе, то есть текстовые данные хранятся в виде независимых друг от друга символов или групп символов, а не параграфов или списков и д. как, например в тегированных форматах наподобие ODT или DOCX [2]. В работе описаны алгоритмы для формирования строк и параграфов, и извлечения их свойств и атрибутов в PDF документе.

Основная часть. Согласно проведенному исследованию, были выделены ключевые проблемы при формировании структурных элементов документа, которые включают в себя: отсутствие классической структуры текстового документа, отсутствие объектов типа таблиц, и соответственно маркировки табличного текста, и несоблюдение правил оформления отчетов студентами согласно нормативным актам [3,4]. Для решения поставленных проблем было принято решение в основу алгоритмов положить принципы, исходя из которых человек может определять различные структурные элементы в документе. В качестве ключевой переменной для выделения строк в тексте использовались данные о местоположении символов на странице по вертикали [5]. В основу алгоритма для выделения параграфов легла информация об отступах строк от левой границы страницы и межстрочный интервал. Для выделения используемых шрифтов и размера текста в строках и абзацах применялась информация о шрифте и размере каждого отдельного символа, входящего в эти объекты.

Выводы. Разработанные алгоритмы для формирования и выделения строк и параграфов из PDF документов позволили также извлекать информацию о свойствах и атрибутах текста. Следует отметить, что количество извлекаемых свойств уступает тегированным форматам. Созданные алгоритмы в дальнейшем планируется использовать при проведении автоматизированного нормоконтроля электронных документов формата PDF.

Список использованных источников:

1. PDF 2.0 "ISO 32000-2:2020(en), Document management — Portable document format — Part 2: PDF 2.0". www.iso.org.
2. Анализ возможностей парсинга электронных текстовых документов для автоматизации нормоконтроля / В. И. Марцинкевич, Г. С. Ларионова, В. В. Терещенко [и др.] // Экономика. Право. Инновации. – 2022. – № 3. – С. 39-49. – DOI 10.17586/2713-1874-2022-3-47-57.
3. Parinov, S. (2017). Extraction and visualisation of citation relationships and its attributes for papers in PDF. International Journal of Metadata, Semantics and Ontologies, 12(4), 195-203.
4. Schäfer, U., & Kiefer, B. (2011). Advances in deep parsing of scholarly paper content. In Advanced Language Technologies for Digital Libraries: International Workshops on NLP4DL 2009, Viareggio, Italy, June 15, 2009 and AT4DL 2009, Trento, Italy, September 8, 2009 (pp. 135-153). Springer Berlin Heidelberg.
5. Chao, H., & Fan, J. (2004). Layout and content extraction for pdf documents. In Document Analysis Systems VI: 6th International Workshop, DAS 2004, Florence, Italy, September 8-10, 2004. Proceedings 6 (pp. 213-224). Springer Berlin Heidelberg.