УДК 004.624, 004.912

Архитектура конвертера документов для системы семантического анализа текстов на естественном языке

Ткешелашвили А.М., Университет ИТМО, Санкт-Петербург
научный руководитель: Клименков С.В., Университет ИТМО, Санкт-Петербург

В рамках разработки системы семантического анализа текстов на естественном языке ведутся работы по созданию модулей, решающих различные прикладные задачи по обработке текста и извлечению из него информации. Исходный текст в общем случае может быть представлен в различных текстовых форматах, таких как DOC, ODT, TXT, PDF, XLS и другие. Для унификации входных данных разрабатываемых модулей, а также для возможности применения их алгоритмов к документам различных форматов, была разработана отдельная подсистема, занимающаяся конвертацией исходных документов во внутренний формат. Структура внутреннего формата отражает специфику решаемой задачи и ориентирована на сохранение семантически значимой информации, а именно смысловых связей между элементами документа, а не на их визуальное отображение.

Целью работы является разработка гибкой архитектуры модулей конвертации документа, позволяющей обрабатывать большинство современных форматов, содержащих текстовую информацию.

Конвертер документов имеет модульную структуру, в рамках работы разработано три модуля, соответствующие трём типам документов: текстовые документы, электронные таблицы и иллюстрационные документы. В работе проведен анализ существующих API для преобразования форматов, в рамках которого выбран API LibreOffice, обеспечивающий унифицированное преобразование сотен форматов документов. Конвертер запускает LibreOffice в серверном режиме, обеспечивая многопоточную одновременную обработку множества операций конвертации.

В рамках работы был проведен архитектурный анализ и синтез модульной структуры конвертера, отдельно выделены обработчики форматов, определение и фильтрация стилей, формирование связей и атрибутов, необходимых для последующего семантического анализа, вывод результатов в разработанный внутренний формат в виде JSON объектов или XML-документа с выбором необходимых для реализации преобразования API.

Автор _____ / Ткешелашвили А.М.
Научный руководитель _____ / Клименков С.В.
Руководитель образовательной программы _____ / Алиев Т.И.

Контакты автора: annamt@yandex.ru, 8 (812) 325-25-53