УДК 004

Подходы к автоматизированному анализу оформления электронных документов формата docx

Комаров М.С. (магистрант, Университет ИТМО), **Тартынских П.С**.
(магистрант, Университет ИТМО)
Научный руководитель - **Насыров Н.Ф.** (аспирант, Университет ИТМО)

Стремительная цифровизация в сфере образовательных технологий требует создания информационных систем, направленных на обеспечение образовательного процесса. Так одним из востребованных направлений, является проверка подготовленных электронных документов на соответствие набору некоторых формальных правил. Это требуется в организациях различного профиля для поддержания единого стиля документооборота. Например, в учебных заведениях для оформления выпускных квалификационных работ в Университете ИТМО имеется положение, регламентирующее нормы оформления документа.

На данный момент нормоконтроль выполняется контролирующими лицами “вручную”, что, очевидно, требует больших временных затрат, а также отвлекает проверяющих от семантики, содержания документа. Поскольку требования форматирования являются формальными, анализ на соответствии документом этих требований может быть представлен в виде алгоритма, а это в свою очередь означает то, что этот процесс может быть автоматизирован.

Анализ инструментов и исследований, направленных на решение поставленной задачи, выявил недостаточное количество решений. Таким образом, авторами была сформулирована задача по обзору подходов для реализации такого инструмента - инструмента, способного проанализировать оформление электронных документов на их соответствие набору формальных правил и дать нормоконтролеру отчёт по результатам такого анализа.

На данный момент для редактирования и хранения электронных документов наиболее распространен формат docx, ключевой особенностью которого является возможность сохранения заданного пользователем стиля и форматирования документа, поэтому именно он рассматривается авторами в качестве формата входных файлов.

Объектная модель документа формата docx предоставляет разнообразные типы и интерфейсы, используемые для представления docx документа в объектном виде. Это облегчает проведение анализа свойств элементов документа. Поиск среди источников выявил две популярные реализации таких моделей.

Microsoft.Office.Interop.Word - официальная библиотека от Microsoft. Она предоставляет большое количество разнообразных объектов, организованных в виде иерархии, которая соответствует пользовательскому интерфейсу программы Word. Основное место среди них отводится объекту Document, предоставляющего все содержимое документа. Так, с помощью его свойства Paragraphs можно проводить анализ форматирования каждого абзаца документа.

Библиотека GemBox.Document - стороннее решение от компании GemBox. Это самостоятельный .NET компонент, с помощью которого можно создавать новые и взаимодействовать с существующими docx документами. Среди предоставляемых возможностей библиотеки можно выделить чтение и запись в отличных от docx форматах (xps, pdf и др.) и доступ к свойствам элементов документа (секций, абзацев).

Кроме того, в качестве подходов авторами рассмотрена методология поиска некоторых часто встречающихся ошибок путём перевода исходного документа формата docx в другие форматы представления электронных документов, такие как графические форматы, форматы xps, pdf, а также путём извлечения текстового содержимого электронного документа.

Таким образом, в работе представлен анализ способов проверки оформления электронного документа формата docx в контексте решения задачи автоматизации нормоконтроля. Приведено сравнение подходов, описаны их достоинства и недостатки, в качестве примера рассмотрены алгоритмы поиска некоторых типовых ошибок путём комбинации описанных методов, дана оценка сложности автоматизации для поиска тех или иных ошибок.

Авторы: Комаров М. С. _____

Тартынских П. С. _____

Научный руководитель: Насыров Н. Ф._____