УДК 004.021:004.912

Проектирование приложения для структурного анализа документа с шаблоном

И.Н. Магомедов

Балтийский государственный технический университет «ВОЕНМЕХ» им. Д.Ф. Устинова

Структурный анализ документа с шаблоном, является рутинной задачей для человека и может приводить к пропуску ошибок. Задача усугубляется в случаях, когда документы начинают разрастаться или они в первой итерации имели очень большой объём. Для решения данной проблемы была поставлена задача проектирования приложения для структурного анализа документа с шаблоном.

В данной работе представлено спроектированное программное решение данной задачи.

При проектировании были рассмотрены структуры текстовых документов docx и odt. Оба документа представляют из себя zip архивы и разделены на файлы формата Office Open XML и OpenDocument. Разметка документов схожа, имеются подобные теги разметки. Главная особенность структуры внутреннего представления документов в данных форматах в том, что текст делится на параграфы, представляющие из себя абзацы в тексте. Таблицы, нумерованные списки и изображения используют отдельную аналогичную разметку.

В ходе проведённой работы были выделены особенности в виде деления элементов текстового документа на отдельные блоки как в формате docx и odt. Такое структурирование данных можно использовать для разделения текстового документа на отдельные структурные блоки (элементы), что упростит задачу анализа документа.

Были рассмотрены возможные алгоритмы для модуля сравнения документов, и сам процесс работы программы. Далее представлен наиболее подходящий разработанный алгоритм:

1. Программе подаётся шаблон на вход;
2. Шаблон анализируется, то есть выделяются основные части;
3. Программе подаётся файл для сравнения;
4. Программа выделяет отдельные блоки документа;
5. Программа поочерёдно заносит в буфер один из блоков исходного документа и сравнивает его с элементами шаблона;
6. Сравнение продолжается пока не будет найден элемент похожий на блок из исходного документа или не наступит конец шаблона;
7. Оценка схожести:

a. если блок найден, то он выделяется зелёным цветом;

b. если блок частично совпадает с элементом шаблона, то жёлтым выделяются схожие части блока;

c. если блок не найден, выделяется красным.

8. В конце оба документа отображаются пользователю в виде: шаблон отображается слева без каких-либо пометок, а документ с пометками открывается справа;
9. Возможности после сравнения документов:

a. возможность редактирования документа, при этом входной документ остаётся неизменным;

b. возможность сохранения документа с пометками;

c. возможность сохранения документа после внесения правок, как с пометками, так и без них;

d. возможность быстрого поиска по средствам клика на элемент (блок) шаблона (исходного документа) или на участок элемента (блока) шаблона (исходного документа).

При сравнение структурных элементов будет рассматриваться не только содержание, а также стили и расположение объектов. Приложение будет показывать специальным знаком (стрелка вверх или вниз) позицию откуда был перемещён блок.

Данное приложение имеет широкий круг применения, от учебных заведений до корпоративного сектора, например, может применяться при оформлении договоров, технических заданий, рабочих и учебных планов.

Приложение позволит ускорить процесс создания документов за счёт быстрой проверки документа на ошибки и наличие нужных структурных элементов, а также даёт возможности редактирования и внесения исправлений без использования нескольких программ. Также позволяет создавать новые шаблоны на основании имеющегося шаблона и документа.

1. MSDN – сеть разработчиков Microsoft [Электронный ресурс] URL: https://msdn.microsoft.com/ru-ru (дата обращения: 19.03.2018).
2. Магомедов И.Н. Сопоставление структур форматов документов odt и docx // Старт-2017: Тезисы докладов III Общероссийской молодёжной науч.-техн. конф. – СПб: Балт. гос. техн. ун-т., 2017. – С. 48.