УДК 004.912:81'322.2

ПОЛУЧЕНИЕ ИСХОДНОЙ ИНФОРМАЦИИ В РАМКАХ СИСТЕМЫ АВТОМАТИЗИРОВАННОГО АНАЛИЗА ТЕКСТОВЫХ ЗАИМСТВОВАНИЙ

Мищенко И.И.
к.т.н., доцент каф. КИСС Защелкин К.В.
Одесский Национальный Политехнический Университет, УКРАИНА

АННОТАЦИЯ. Рассмотрена задача получения текстовых фрагментов из стандартных форматов документов. Предложен способ получения текстовой информации при помощи промежуточного представления документов в формате PDF.

Введение. В учебном процессе высших учебных заведений весьма часто используется форма контроля знаний, при которой, студент получает внеаудиторное задание и в качестве результата его выполнения формирует некоторый набор текстовой информации [1]. Этот набор информации обычно представляется в виде совокупности файлов в одном из распространенных форматов текстовых документов, которая передается для анализа преподавателю или автоматизированной системе контроля знаний [2]. Для данной формы контроля знаний характерна проблема возможного наличия текстовых заимствований в блоках текстовой информации, переданных на проверку. Анализ текстовых заимствований представляет собой весьма трудоемкую задачу. В связи с этим проблема построения автоматизированных систем анализа текстовых заимствований является очень актуальной.

Цель работы. Цель данной работы состоит в формировании подходов к получению исходной текстовой информации из файлов, представленных в стандартных форматах текстовых документов в рамках системы анализа текстовых заимствований.

Основная часть работы. Первым этапом функционирования системы анализа текстовых заимствований системы является обработка документов, загруженных на сервер. На текущий момент, основными форматами текстовых документов являются DOC, DOCX, ODT и PDF. Все эти форматы имеют различную структуру и требуют нетривиальных действия по получению исходных текстов из структуры файла. Были проанализированы существующие программные средства (библиотеки), выполняющие извлечение данных из форматов документов DOC, DOCX, ODT. Однако все они характеризуются, во-первых низким качеством получения результата, а во-вторых значительным временем, необходимым для получения результата. В силу этого было принято решение использовать формат PDF в качестве промежуточной формы представления для получения исходных текстовых данных из документов. Было принято решение автоматически конвертировать загруженные на сервер файлы форматов DOC, DOCX, ODT в формат PDF, а затем считывать текст из этого формата. Для конвертации в формат PDF использовалась библиотека DocsToPdfConverter. Эта библиотека способна обрабатывать файлы формата DOC, DOCX, ODT, XLS и XLSX и переводить их в формат PDF за приемлемое временя. После конвертирования, обработанные файлы сохраняются в отдельный каталог, расположенный на сервере. Далее полученные файлы обрабатываются при помощи библиотеки IText, которая позволяет извлекать текстовые фрагменты из файлов формата PDF.

Выводы. В работе выполнен анализ возможных способов получения исходных текстов из текстовых документов стандартных форматов. Обоснован и программно реализован подход в рамках которого для получения текстовых фрагментов используется промежуточная форма представления документов в формате PDF.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Clark R. E-learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning / R. Clark, R. Mayere. – New Jersey: Wiley, 2016.

2. Защелкин, К.В. Автоматизированная система контроля знаний, основанная на использовании Web-сервисов хостинга IT-проектов / К.В. Защелкин, Е.Н. Иванова // Труды международной научно-практической конференции “СИЭТ-2016”. – Одесса, 2016. – С. 48-49.