никами, которые используют вращательное движение большого мотора, преобразованное в поступательное. После чего детали определенного размера перемещаются по конвейеру в контейнер, закрепленный за ними.

Заключение. Таким образом, нами создан прототип робота на базе конструктора Lego Mindstorms EV3, осуществляющего сортировку мусора.

Результаты исследования внедрены в учебный процесс образовательного центра факультета математики и информационных технологий Витебского государственного университета имени П.М. Машерова «IT-академия МИР будущего».

1. Закон Республики Беларусь «Об обращении с отходами» - 2019 от 10 мая 2019 г. № 186-З [Электронный ресурс]. - Режим доступа: https://ohranatruda.of.by/zakon-respubliki-belarus-ob-obrashchenii-s-otkhodami-2019-s-izmeneniyami-ot-10-maya-2019-g-186-z.html. - Дата доступа: 19.03.2022.

2. ПОСТАНОВЛЕНИЕ СОВЕТА МИНИСТРОВ РЕСПУБЛИКИ БЕЛАРУСЬ от 28 июля 2017 г. №567 [Электронный ресурс]. Режим доступа: https://pravo.by/document/?guid=12551&p0=C21700567&p1=1&p5=0. – Дата доступа: 19.03.2022.

РАЗРАБОТКА МОДУЛЯ ПРЕОБРАЗОВАНИЯ ДОКУМЕНТОВ ИЗ ФОРМАТА PDF

Корниенко А.А.,
магистрант 1 курса ВГУ имени П.М. Машерова, г. Витебск, Республика Беларусь
Научный руководитель – Семенов М.Г., канд. физ.-мат. наук, доцент

В современном мире различные сферы деятельности всё чаще требуют разработки документации в электронном виде. Эта документация позже помогает её пользователям быстрее разобраться в новой технологии или изучить некоторую проблему. Однако многие коммерческие проекты и научные работы используют уже существующие сложные системы со своими тщательно проработанными документациями. Как правило, такая информация публикуется с помощью PDF файлов, потому что они позволяют добиться наибольшего сходства с результатами печати. Однако находящиеся в активной разработке документации, как правило, хранятся с помощью других форматов, например RTF или Word, которые содержат большое количество дополнительных данных о последовательности текста, отсутствующих в PDF. В связи с этим, возникает необходимость разработки и применения инструментов автоматического преобразования документов этих форматов. Несмотря на большое количество существующих бесплатных инструментов для подобного преобразования форматов, они, как правило, не предоставляют гибкость настройки, достаточной для того, чтобы избежать дополнительной обработки сгенерированных документов, что может быть компенсировано разработчиком, самостоятельно реализовавшим модуль для преобразования таких документов.

Цель данной работы – разработка модуля, предоставляющего базовые возможности преобразования PDF в другие форматы.

Материал и методы. Материалом для исследования послужила официальная документация форматов PDF и DOCX. При разработке модуля использовались библиотеки iText7 [1] и OpenXML SDK [2]. Обе имеют открытый исходный код. Несмотря на то, что первая имеет платные функции, они необязательны для анализа документа.

Результаты и их обсуждение. При реализации данного модуля были поставлены следующие требования:

1. Модуль должен считывать PDF файлы любых версий ниже 2.0.
2. Входные данные не должны быть преобразованы в итоговый формат напрямую. Должна быть определена промежуточная структура.
3. Модуль должен преобразовывать полученное представление в заданный формат с сохранением пропорций, заданных в исходном файле.

Проанализированный с помощью разработанного модуля файл может быть представлен с помощью структуры, содержащей следующие элементы:

1. Группы символов, соответствующие последовательностям текста с одинаковым стилем.
2. Слова, включающие в себя одну или несколько последовательностей символов, не содержащих пробелов, или расстояния между которыми не превышают заданного интервала между символами. Слова также должны содержать информацию о координатах левого нижнего угла, ширине и высоте слова.
3. Строки, содержащие последовательность слов. Строки также должны содержать информацию о положении левой и верхней границы, ширине и высоте строки.
4. Группы строк, соответствующие последовательностям строк с одинаковыми отступами между ними, от одного или обоих краёв страницы.
5. Абзацы, соответствующие группам строк, отделенных друг от друга отличием в расстоянии между двумя строками или отступом первой строки в абзаце.
6. Столбцы, соответствующие последовательностям содержимого, расположенного в виде двух вертикальных потоков, не зависящих друг от друга. Текст, расположенный в виде столбцов отличается от текста с интервалом, как, например, в случае нумерации или списка перечисления, последовательностью идентификаторов элементов помеченного содержимого (MCID).
7. Строки макета, соответствующие горизонтальным последовательностям содержимого, размещенного с помощью одинакового количества вертикальных столбцов. Строка макета меняется, когда меняется количество столбцов.

Таким образом, модуль способен распознавать текст, расположенный в нескольких столбцах, содержащий абзацы, отделяющиеся увеличенным расстоянием между строками или красной строкой (рисунок 1).

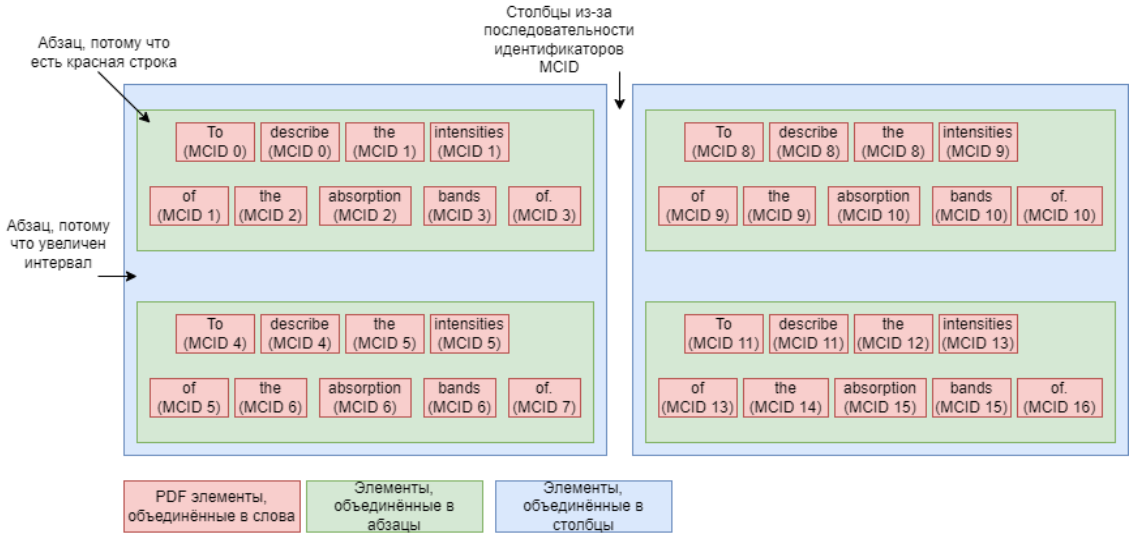

Рисунок 1 – Общая структура анализа документа

Заключение. В результате выполнения работы был разработан модуль, предоставляющий функционал анализа и преобразования PDF документов.

На данный момент продолжается работа над добавлением недостающего функционала и улучшением производительности приложения.

1. Официальная документация iText7 [Электронный ресурс] – Режим доступа: https://itextpdf.com/ru/products/itext-7/itext-7-core. – Дата доступа: 07.03.2022.

2. Официальная документация библиотеки OpenXML SDK [Электронный ресурс]. – Режим доступа: https://docs.microsoft.com/en-us/office/open-xml/open-xml-sdk. – Дата доступа: 08.03.2022.