Умный конвертер из PDF в FB2

Пояснительная записка
к итоговому проекту
по треку Искусственный Интеллект
Samsung Innovation Campus

Выполнил:
Торотенков Дмитрий Борисович

Москва, 2024

Актуальность

Электронные книги завоевывают всю большую популярность среди читателей. Поэтому задача создание контента для чтения с помощью электронных ридеров достаточно актуальна.

Практически вся литература на сегодняшний день так или иначе оцифрована и лежит в открытом доступе на просторах интернета. Наибольшее распространение на сегодняшний день получил формат PDF. Данный формат может объединять разноплановую информацию (картинки, текст, таблицы) в один документ.

Проблема заключается в том, что этот формат воспринимается электронными ридерами как графический. То есть, на экране своей электронной книги пользователь видит одну большую картинку, которая масштабируется по размеру дисплея устройства. Получается, что какого бы не были размера страницы в исходном PDF-документе, все содержимое в таком же виде поступает пользователю. Это было бы приемлемо, если бы дисплей ридера достигал 10 дюймов. Но такие модели стоят неоправданно дорого (начиная от 40 тысяч рублей). Поэтому наибольшее распространение получили 5.5 и 6-ти дюймовый ридеры. И, естественно, чтение непосредственно PDF файлов затруднительно.

Лучше всего адаптированы (так как и создавались для этих целей) к чтению на электронных книгах форматы EPUB и FB2. Первый из которых является международным, а второй получил наибольшее признание на территории СНГ. FB2 = наиболее простой формат, представляющей по своей сути простой XML файл с текстом и картинками записанных с помощью base64. С такой простотой связан и главный

минус данного формата – данные располагаются строго друг под другом и выравниваются по левому краю. EPUB – это целый архив, по структуре напоминающий web-страницу, где каждый лист будущей электронной книги размечается отдельно. Это сильно усложняет создание книг в этом формате, поэтому данной работой занимаются издательства.

Анализ данных обстоятельств наводит на мысль: конвертировать данные из неудобного для чтения с электронного ридера PDF в разработанный для этого FB2 (или EPUB). Анализ существующих решений, позволяющих сделать это онлайн, показывает, что полноценных решений данной проблемы нет. Все протестированные онлайн конвертеры либо пропускали часть информации (такое может быть, когда изображение в PDF файле не помечено как рисунок), создавали абсолютно нечитабельное форматирование (разбивая текст на абзацы по строкам в PDF файле). А в случае, когда PDF представляет из себя отсканированный документ, то так в FB2 эти изображения и записывали (т.е. не представляют никакой практической пользы).

Проблемы не возникает, если уже на просторах интернета можно найти интересующую книгу в формате FB2 или EPUB. Но как показывает практика, готовый FB2 файл можно найти лишь для художественной литературы, изданной более 10 лет назад. За все время исследования не было найдено ни одного готового файла с технической литературой (Под словом техническая литература понимаются именно академические учебники и иная литература, содержащая большое количество формул, таблиц и пояснительных рисунков). Это легко объяснимо тем, что данная литература изобилует графически сложным

контентом, который крайне проблематично записать в FB2 формат с помощью традиционных методов конвертации.

Именно на решение этой проблемы направлен данный проект. А учитываю отсутствие готовых решений в данном сегменте, данный проект имеет хорошие перспективы дальнейшей коммерциализации, как посредством добавления рекламы, так и продажи авторских прав.

Целевая аудитория данного проекта – это активные пользователи электронных книг (по данным опроса более 13% россиян посещают сайт Литрес). Особенно актуален этот проект людям, часто читающим именно техническую литературу.

Идея реализации

Главная идея данного проекта – сохранять графическую и трудно передаваемую текстовую информацию (например, математические формулы) в виде изображений. А простой текст так и записывать в будущую FB2 книгу.

Основан проект использование нейросетевой модели для распознавания содержимого на странице документа PDF, задача которой определить, где на странице текст, где сопроводительная информация, а где будущие изображения (таблицы, формулы, картинки). Дальше процесс будет различаться, в зависимости от того, с каким типом PDF необходимо работать. Если предстоит обработать отсканированные документы, то данная модель сегментации – это единственный источник получения информации. Так то, что модель будет считать картинками будет вырезано и сохранено, как отдельные изображения. Места, которые

модель воспримет, как текстовые поля, в дальнейшем будут обработаны специальной OCR-системой.

При обработки текстовой информации с помощью OCR-систем возможно появление нелепых опечаток и странных ошибок. Поэтому необходимо дополнить данную систему корректировщиком, который опираясь на правила русского языка, пытался устранить эти проблемы.

Иначе обстоит дело, в случае с программно-генерируемыми PDF документами. В этих файлах, непосредственно, хранится текст, изображения, таблицы. Поэтому исчезает надобность в использовании OCR-системы для распознавания текста. При этом роль модели сегментации в процессе обработки такого документа не уменьшается. Без неё было бы невозможно отследить математические формулы на странице, а также многие рисунки могут оказаться не рисунками. Тогда они упускаются из виду и итоговый файл теряет часть информации. Проверку корректности текса в данном случае проводить не нужно.

Взаимодействие с пользователем необходимо реализовать в удобном и понятном виде. Наилучшим образом подойдет создание телеграмм-бота, который будет принимать PDF файл и возвращать документ в формате FB2. Так как вычислительные мощности в данном проекте строго ограничены, то необходимо реализовать живую очередь, во избежание ошибок при конвертации.

Дальнейшая реализация

После формирования и создания основного скелета проекта, для повышения лояльности пользователей и

привлечения новых клиентов необходимо добавить новые функции.

Интересным примером такой функции может стать создание автоматической аннотации к каждой книги. При этом практический интерес может представлять сохранения текстовых данных при конвертации для последующего обучения языковых моделей (естественно, данные действия будут осуществляться с согласия пользователя).

Заключение

В силу отсутствия прямых аналогов и многочисленности целевой аудитории, данный проект обладает значительным коммерческим потенциалом. При этом, грамотная реализация проекта, может привести к увеличению целевой аудитории (например, повышению популярности электронных книг среди студентов).

Очень заманчиво иметь в кармане сотню учебников, при чтении которых не так сильно устают глаза (в сравнении со смартфоном или планшетом). А именно такие перспективы открывает реализация данного проекта.