Содержание:

Что такое OCR

Представьте, вам надо оцифровать журнальную статью или распечатанный договор. Конечно, вы можете провести несколько часов, перепечатывая документ и исправляя опечатки. Либо вы можете перевести все требуемые материалы в редактируемый формат за несколько минут, используя сканер (или цифровую камеру) и программу для оптического распознавания символов (OCR).

Что такое OCR

ЧТО ПОДРАЗУМЕВАЮТ ПОД ТЕХНОЛОГИЕЙ

ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ СИМВОЛОВ

Оптическое распознавание символов (англ. Optical Character Recognition – OCR) – это технология, которая позволяет преобразовывать различные типы документов, такие как отсканированные документы, PDF-файлы или фото с цифровой камеры, в редактируемые форматы с возможностью поиска.

Предположим, у вас есть бумажный документ, например, статья в журнале, брошюра или договор в формате PDF, присланный вам партнером по электронной почте. Очевидно, для того чтобы получить возможность редактировать документ, его недостаточно просто отсканировать. Единственное, что может сделать сканер, – это создать изображение документа, представляющее собой всего лишь совокупность черно-белых или цветных точек, то есть растровое изображение.

Для того чтобы копировать, извлекать и редактировать данные, вам понадобится программа для распознавания символов, которая сможет выделить в изображении

буквы, составить их в слова, а затем объединить слова в предложения, что в дальнейшем позволит работать с содержимым исходного документа.

Какие принципы лежат в основе технологии FineReader OCR?

КАКИЕ ПРИНЦИПЫ ЛЕЖАТ В ОСНОВЕТЕХНОЛОГИИ FINEREADER OCR?

Наиболее совершенные системы распознавания символов, такие как ABBYY FineReader OCR, делают акцент на использовании механизмов, созданных природой. В основе этих механизмов лежат три фундаментальных принципа: целостность, целенаправленность и адаптивность (принципы IPA).

Изображение, согласно принципу целостности, будет интерпретировано как некий объект, только если на нем присутствуют все структурные части этого объекта и эти части находятся в соответствующих отношениях. Иначе говоря, ABBYY FineReader не пытается принимать решение, перебирая тысячи эталонов в поисках наиболее подходящего. Вместо этого выдвигается ряд гипотез относительно того, на что похоже обнаруженное изображение. Затем каждая гипотеза целенаправленно проверяется. И, допуская, что найденный объект может быть буквой А, FineReader будет искать именно те особенности, которые должны быть у изображения этой буквы. Как и следует поступать, исходя из принципа целенаправленности. Принцип адаптивности означает, что программа должна быть способна к самообучению, поэтому проверять, верна ли выдвинутая гипотеза, система будет, опираясь на накопленные ранее сведения о возможных начертаниях символа в данном конкретном документе.

Какая технология лежит в основе OCR?

КАКАЯ ТЕХНОЛОГИЯ ЛЕЖИТ В ОСНОВЕ OCR?

Компания ABBYY, опираясь на результаты многолетних исследований, реализовала принципы IPA в компьютерной программе. Система оптического распознавания символов ABBYY FineReader – единственная в мире система OCR, действующая в соответствии с вышеописанными принципами на всех этапах обработки документа.

Эти принципы делают программу максимально гибкой и интеллектуальной, предельно приближая ее работу к тому, как распознает символы человек. На первом этапе распознавания система постранично анализирует изображения, из которых состоит документ, определяет структуру страниц, выделяет текстовые блоки, таблицы. Кроме того, современные документы часто содержат всевозможные элементы дизайна: иллюстрации, колонтитулы, цветной фон или фоновые изображения. Поэтому недостаточно просто найти и распознать обнаруженный текст, важно с самого начала определить, как устроен рассматриваемый документ: есть ли в нем разделы и подразделы, ссылки и сноски, таблицы и графики, оглавление, проставлены ли номера страниц и т. д. Затем в текстовых блоках выделяются строки, отдельные строки делятся на слова, слова на символы.

Важно отметить, что выделение символов и их распознавание также реализовано в виде составных частей единой процедуры. Это позволяет в полной мере использовать преимущества принципов IPA. Выделенные изображения символов поступают на рассмотрение механизмов распознавания букв, называемых классификаторами.

В системе ABBYY FineReader применяются классификаторы следующих типов: растровый, признаковый, контурный, структурный, признаково-дифференциальный и структурно-дифференциальный. Растровый и признаковый классификаторы анализируют изображение и выдвигают несколько гипотез о том, какой символ на нем представлен. В ходе анализа каждой гипотезе присваивается определенная оценка (так называемый вес). По итогам проверки мы получаем список гипотез, проранжированный по весу (то есть по степени уверенности в том, что перед нами именно такой символ). Можно сказать, что в данный момент система уже «догадывается», на что похож рассматриваемый символ.

После этого в соответствии с принципами IPA ABBYY FineReader проводит проверку выдвинутых гипотез. Это делается с помощью дифференциального признакового классификатора.

Кроме того, следует отметить, что ABBYY FineReader поддерживает 192 языка распознавания. Интеграция системы распознавания со словарями помогает программе при анализе документов: распознавание происходит более точно и упрощает дальнейшую проверку результата с учетом данных об основном языке документа и словарной проверки отдельных предположений. После подробной обработки огромного числа гипотез программа принимает решение и

предоставляет пользователю распознанный текст.

Распознавание цифровых фотографий

РАСПОЗНАВАНИЕ ЦИФРОВЫХ ФОТОГРАФИЙ

Изображения, полученные при помощи цифровой камеры, отличаются от отсканированных документов или PDF, представляющих собой изображение.

У них зачастую могут быть определенные дефекты, например искажения перспективы, засветки от фотовспышки, изгибы строк. При работе с большинством приложений такие дефекты могут существенно усложнить процесс распознавания. В связи с этим последние версии ABBYY FineReader содержат технологии предварительной обработки изображения, которые успешно выполняют задачи по подготовке изображений к распознаванию.

Как пользоваться OCR-программами

КАК ПОЛЬЗОВАТЬСЯ OCR-ПРОГРАММАМИ

Технология ABBYY FineReader OCR проста в использовании – процесс распознавания в целом состоит из трех этапов: открытие (или сканирование) документа, распознавание и сохранение в наиболее подходящем формате (DOC, RTF, XLS, PDF, HTML, TXT и т. д.) либо перенос данных напрямую в офисные программы, такие как Microsoft® Word®, Excel® или приложения для просмотра PDF.

Кроме того, последняя версия ABBYY FineReader позволяет автоматизировать задачи по распознаванию и конвертации документов с помощью приложения ABBYY Hot Folder. С помощью него можно настраивать однотипные или повторяющиеся задачи по обработке документов и увеличить производительность работы.

Какие преимущества вы получаете от работы с OCR-программами

КАКИЕ ПРЕИМУЩЕСТВА ВЫ ПОЛУЧАЕТЕ ОТ РАБОТЫ С OCR-ПРОГРАММАМИ

Высокое качество технологий распознавания текста ABBYY OCR обеспечивает точную конвертацию бумажных документов (сканов, фотографий) и PDF-документов любого типа в редактируемые форматы. Применение современных OCR-технологий позволяет сэкономить много сил и времени при работе с любыми документами. С ABBYY FineReader OCR вы можете сканировать бумажные документы и редактировать их. Вы можете извлекать цитаты из книг и журналов и использовать их без перепечатывания. С помощью цифровой фотокамеры и ABBYY FineReader OCR вы можете моментально сделать снимок увиденного постера, баннера, а также документа или книги, когда под рукой нет сканера, и распознать полученное изображение. Кроме того, ABBYY FineReader OCR можно использовать для создания архива PDF-документов с возможностью поиска.

Весь процесс преобразования из бумажного документа, снимка или PDF занимает меньше минуты, а сам распознанный документ выглядит в точности как оригинал!

Выводы

Технология OCR является бесспорно одной из самых полезных. Автор данной работы сам несколько раз прибегал к помощи данной технологии и остался доволен.

Источники

http://old.ci.ru/inform16_02/p_22text.htm

https://www.kp.ru/guide/raspoznavanie-teksta.html

https://www.abbyy.com/ru-ru/finereader/what-is-ocr/