ПРОГРАММЫ ДЛЯ РАСПОЗНАВАНИЯ ТЕКСТА

Обухов Вадим Анатольевич
Тохирова Сарвиноз Гайратжон кизи
Исахонов Хушнидбек Муродилжон угли
Ферганский филиал Ташкентского университета
информационных технологий имени Мухаммада Ал-Хоразмий

Аннотация: Программы для распознавания текста - это программные решения, разработанные для автоматического преобразования текстовых данных, представленных в различных форматах (например, отсканированных изображений, рукописных заметок или фотографий), в машинночитаемый текст. Эти программы играют важную роль в автоматизации процессов обработки информации и являются ключевой частью в областях, таких как оптическое распознавание символов (OCR), обработка документов, анализ текста и многие другие.

Ключевые слова: программы, распознавание текста, точность, многоязычность, адаптивность, интеграция, OCR, обработка документов, автоматизация, машинное обучение.

1. ABBYY FineReader.

ABBYY FineReader - программа для оптического распознавания символов, разработанная российской компанией ABBYY.

Программа позволяет переводить изображения документов (фотографий, результатов сканирования, PDF-файлов) в электронные редактируемые форматы. В частности, в Microsoft Word, Microsoft Excel, Microsoft Powerpoint, Rich Text Format, HTML, PDF/A, searchable PDF, CSV и текстовые (plain text) файлы. Начиная с одиннадцатой версии файлы можно сохранять в формате djvu. Версия двенадцать поддерживает распознавание текста на сто девяноста языках и имеет встроенную проверку орфографии для сорока восьми из них.

FineReader безоговорочный лидер среди всех программ, распознающих текст на изображении.

Единственное обстоятельство, которое может разочаровать пользователей, состоит в том, что программа платная. Бесплатно распространяется только пробная версия на пятнадцать дней. За этот период разрешено сканирование пятидесяти страниц.

Достоинства: точное распознавание; огромное количество языков чтения; толерантность к качеству изображения-источника.

Недостаток: пробная версия на пятнадцать дней.

2. OCR CuneiForm.

CuneiForm (англ. cuneiform, - клинопись), Cognitive OpenOCR - свободно распространяемая открытая система оптического распознавания текстов российской компании Cognitive Technologies.

OCR CuneiForm была разработана компанией Cognitive Technologies как коммерческий продукт в 1993 году. Система поставлялась с наиболее популярными моделями сканеров, МФУ и ПО в России и мире: Corel Draw, Hewlet-Pachard, Epson, Xerox, Samsung, Brother, Mustek, OKI, Canon, Olivetti и др.

OCR – это использование технологии для идентификации и преобразования отсканированных рукописных или печатных текстовых символов в электронную форму, более легко распознаваемую компьютерами и другими программами. Базовый процесс распознавания включает изучение текста и перевод символов в код, который можно использовать для обработки данных. OCR иногда также называют распознаванием текста.

Технология состоит из сочетания аппаратного и программного обеспечения, которое используется с целью преобразования физических документов в машиночитаемый текст. Аппаратное обеспечение, такое как оптический сканер или специализированная монтажная плата, используется для копирования или чтения текста, в то время как программное обеспечение отвечает за расширенную обработку. Программное обеспечение может использовать искусственный интеллект для реализации более совершенных методов интеллектуального распознавания (ICR), таких как идентификация языков или стилей рукописного ввода.

OCR чаще всего используется для преобразования печатных юридических или исторических документов в PDF-файлы. После этого полученные электронные копии пользователи могут редактировать, форматировать при помощи обычных редакторов текста. Первым шагом процесса оптического распознавания является использование сканера с целью обработки физической формы документа. После копирования всех страниц программа OCR преобразует документ в двухцветную или черно-белую версию. Отсканированное растровое изображение анализируется на наличие светлых и темных областей. При этом темные области идентифицируются как символы, которые необходимо распознать, а светлые области – как фон. После этого темные области обрабатываются для поиска букв или цифр.

Существующие программы распознавания могут иметь разные методы работы, но, как правило, все они включают таргетинг на один символ, слово или блок текста. Для идентификации символов используются два основных алгоритма.

✓ Обработка распознаваемого материала происходит на примерах различных шрифтов и текстовых форматов.

✓ Распознавание основывается на использовании правил обнаружения признаков, касающихся особенностей конкретной буквы или цифры (ICR). С помощью функции обнаружения программное обеспечение оценивает данные документа в соответствии с правилами о том, как формируется буква или цифра. Например, заглавная буква «А» может храниться как две диагональные линии, пересекающиеся с горизонтальной линией посередине.

Особенности.

CuneiForm позиционируется как система преобразования электронных копий бумажных документов и графических файлов в редактируемый вид с возможностью сохранения структуры и гарнитуры шрифтов оригинального документа в автоматическом или полуавтоматическом режиме. Система включает в себя две программы для одиночной и пакетной обработки электронных документов. Поддерживается смесь русского и английского языка. Бесплатная программа для считывания текстовой информации с изображений. Точность распознавания на порядок ниже, чем у предыдущей рассматриваемой программы. Но как для бесплатной утилиты, функционал все-таки на высоте.

Программа может прочитать и сохранять шрифт распознаваемого текста. В базе шрифтов содержится большинство используемых печатных шрифтов. Поддерживается даже распознавание текста, вышедшего из печатной машинки. Для обеспечения точности к процессу распознавания подключаются специальные словари, которые пополняют словарный запас из сканируемых документов.

Достоинства: бесплатное распространение; использование словарей для проверки правильности текста; сканирование текста с ксерокопий плохого качества.

Недостатки: относительно небольшая точность; небольшое количество поддерживаемых языков.

3. WinScan2PDF.

Это даже не полноценная программа, а утилита. Установка не требуется, а исполнительный файл весит всего в несколько килобайт. Процесс распознавания происходит предельно быстро, правда, полученные в его результате документы сохраняются исключительно в формате PDF.

Фактически весь процесс выполняется при нажатии трех кнопок: выбор источника, места назначения и запуска программы. Утилита предназначена для быстрой пакетной обработки множества файлов. Для удобства пользователей предусмотрен большой языковой пакет интерфейса.

Достоинства: портативность; быстрая работа; простота в использовании.

Недостатки: единственный формат файлов на выходе.

4. SimpleOCR.

Отличная небольшая программа для распознавания текстов с изображений. Поддерживает даже чтение рукописей. Беда в том, что русский не входит ни в языковой пакет интерфейса, ни в список поддерживаемых для распознавания языков.

Однако если необходимо отсканировать английский, датский или французский, то лучшего бесплатного варианта не найти.

В своей области программа обеспечивает точную расшифровку шрифтов, удаление шума и извлечение графических изображений. К тому же в интерфейс программы встроен текстовый редактор, практически идентичный WordPad, что значительно повышает удобство использования программы.

Достоинства: точное распознавание текста; удобный текстовый редактор; удаление шума с изображения.

Недостатки: полное отсутствие русского языка.

5. Freemore OCR.

Программа позволяет оперативно извлекать текст и графику с изображений. Софт поддерживает работу с несколькими сканерами без потери производительности. Извлеченный текст может быть сохранен в формате текстового документа или документа MS Office.

Кроме того, предусмотрена функция многостраничного распознавания.

Распространяется Freemore OCR бесплатно, однако, интерфейс только на английском. Но это обстоятельство никак не влияет на удобство пользования, потому как организованы элементы управления интуитивно понятным образом.

Достоинства: бесплатное распространение; возможность работы с несколькими сканерами; достойная точность распознавания.

Недостатки: отсутствие русского языка в интерфейсе; необходимость загрузки русского языкового пакета для распознавания.

6 Распознавание текста в FineReader.

Для эффективной работы со сканируемыми документами нужно знать, для чего нужна ABBYY FineReader, как пользоваться основными функциями программы и правильно запускать ее. Инструмент для сканирования предельно точно распознает текст в выбранном печатном документе, не перенося постранично информацию. Кроме того, программа старается сохранить шрифты, колонтитулы и разметку текста на странице максимально близко к оригиналу.

В меню выберите «Сервис», перейдите в «Опции» и укажите режим распознавания: тщательное или быстрое распознавание. Тщательный режим будет удобен для работы с некачественными текстовыми файлами, текстами на

цветном фоне или сложными таблицами. Быстрое распознавание рекомендовано для больших объемов файлов или когда ограничены временные рамки.

Чтобы не возникало сложностей при редактировании в ABBYY FineReader 12, разработчики создали интуитивно понятный интерфейс и удобную навигацию по пунктам. Отредактировать текст можно двумя способами: непосредственно в окне «Текст», либо выбрав на панели инструментов «Сервис» и далее «Проверка». Доступные средства для изменения текста находятся над окном «Текст» и включают в себя стандартный набор для редактирования шрифта, его размера, отступов и замены символов. Для редактирования непосредственно PDF-изображения, нужно зайти в меню в «Редактор изображений» и выбрать из списка нужную функцию.

Использованная литература

1. TOJIBOEV, I., RAYIMJONOVA, O., ISKANDAROV, U., MAKHAMMADJONOV, A., & TOKHIROVA, S. МИРОВАЯ НАУКА. МИРОВАЯ НАУКА Учредители: ООО" Институт управления и социально-экономического развития", (3), 26-29.
2. Tojiboev, I., Rayimjonova, O. S., Iskandarov, U. U., Makhammadjonov, A. G., & Tokhirova, S. G. (2022). ANALYSIS OF THE FLOW OF INFORMATION OF THE PHYSICAL LEVEL OF INTERNET SERVICES IN MULTISERVICE NETWORKS OF TELECOMMUNICATIONS. Мировая наука, (3 (60)), 26-29.
3. Muhammadjonov, A., & Toxirova, S. (2023). YARIMO ‘TKAZGICHLARNING TURLARI. ICHKI VA TASHQI YARIMO ‘TKAZGICHLAR. Research and implementation.
4. Isroilovich, H. A., & Abdimahamatovich, H. A. (2023). KIBERJINOYAT JAMIYAT UCHUN YANGI TAHDID SIFATIDA. World scientific research journal, 15(1), 249-252.
5. Abdimahamatovich, H. A., & Anatolyevich, O. V. (2022). SANOAT KORXONALARINING RIVOJLANISH TENDENSIYALARI. Journal of new century innovations, 11(1), 195-202.
6. Обухов, В. А., & Хакимов, А. А. (2022). ОСНОВЫ ИСПОЛЬЗОВАНИЯ РЕКУРСИВНЫХ ФУНКЦИЙ В СТРУКТУРАХ ДАННЫХ. Journal of new century innovations, 11(1), 92-99.
7. Обухов, В. А., & Хакимов, А. А. (2022). МОДЕЛИРОВАНИЕ РЕГИСТРОВ ПРОЦЕССОРА. Journal of new century innovations, 11(1), 169-178.
8. Abdullayeva, M., & Hakimov, A. (2023). ZAMONAVIY AXBOROTLASHGAN JAMIYATDA SANOAT KORXONALARIGA AXBOROT TEXNOLOGIYALARINING TADBIQI. Research and implementation.

9. Z. Qadamova TATU Farg'ona filiali magistri D.Sotivoldiyev Fiskal instituti dotsenti BIOLOGIK NEYRONLARNING MODELI, SUN'IY NEYRON TARMOQLARINING INSONIYAT HAYOTIDAGI AXAMIYATI// Международная научно-техническая конференция «Практическое применение технических и цифровых технологий и их инновационных решений», T
10. Набижонов , Р., & Обухов , В. (2023). ДАЛЬНЕЙШИЙ ВКЛАД БЛОКЧЕЙН-СЕТЕЙ В РАЗВИТИЕ ДИСТАНЦИОННОГО ОБРАЗОВАНИЯ. Research and Implementation. извлечено от https://fer-teach.uz/index.php/rai/article/view/772
11. Nabijonov, R. M. o'g'li, & Mamayeva, O. I. qizi. (2023). TA'LIM SIFATINI OSHIRISHDA ELEKTRON AMALIY DASTURIY PAKETLARNING AHAMIYATI. GOLDEN BRAIN, 1(25), 51–55. Retrieved from https://researchedu.org/index.php/goldenbrain/article/view/4782
12. Обухов, В., Ходжиматов Ж., & Набижонов , Р. (2023). РАЗВИТИЕ БЛОКЧЕЙН ТЕХНОЛОГИЙ В УЗБЕКИСТАНЕ: СОВРЕМЕННЫЕ ВЫЗОВЫ И ПЕРСПЕКТИВЫ. Research and Implementation. извлечено от https://fer-teach.uz/index.php/rai/article/view/768
13. Обухов , В., Хамидов Э., & Набижонов , Р. (2023). ПОЭТАПНОЕ ВНЕДРЕНИЕ БЛОКЧЕЙН ТЕХНОЛОГИЙ В РЕСПУБЛИКЕ УЗБЕКИСТАН. Research and Implementation. извлечено от https://fer-teach.uz/index.php/rai/article/view/770
14. Обухов, В. (2023). 5 СПОСОБОВ, КОТОРЫМИ БЛОКЧЕЙН ПОВЛИЯЕТ НА ИНДУСТРИЮ ОБРАЗОВАНИЯ. Engineering problems and innovations.
15. Акбаров, Д. Е., Кушматов, О. Э., Умаров, Ш. А., & Далиев, Б. С. (2021). Исследование Вопросов Необходимых Условий Идеально Стойких Алгоритмов Шифрования. Central asian journal of mathematical theory and computer sciences, 2(11), 65-70.
16. Ходжиматов, Ж. М., Хамидов, Э. Х., & Собиров, М. М. (2022). ОСНОВНЫЕ СОВРЕМЕННЫЕ ЯЗЫКИ ПРОГРАММИРОВАНИЯ. Journal of new century innovations, 11(1), 136-143.
17. Khamidovich, X. E., & Murodovichelnur, X. J. (2022). Computer-Vision Based Method for Human Action Recognition. International Journal of Innovative Analyses and Emerging Technology, 2(3), 44-47.
18. Хамидов, Э. Х., Собиров, М. М., & Ходжиматов, М. М. (2022). НЕЙРОННЫЕ СЕТИ RNN И LSTM. Journal of new century innovations, 11(1), 127-135.