Подготовил: Козлов Даниил

# ТЕМА: СИСТЕМА ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ ТЕКСТОВ

Вступление

Что это такое OCR?

Как работает OCR?

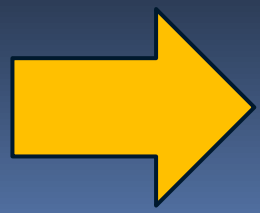
Варианты использования

Заключение

Литература

# Вступление

- Все чаще встречаются ситуации, когда человек сталкивается с задачей перевода рукописей или напечатанных на бумаге текстов на цифровые носители.



# Что это такое OCR?

- OCR или Optical Character Recognition – это система оптического распознавания символов, с помощью которой происходит преобразование изображений, к примеру фотографий печатного текста, файлов в PDF-формате, а также отсканированных документов, в текстовые форматы с возможностью их дальнейшего редактирования и наличием в них поиска.

# Как работает OCR?

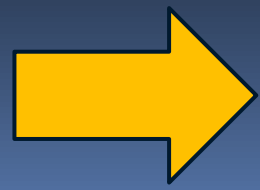
- Первым шагом процесса оптического распознавания является использование сканера с целью обработки физической формы документа. После копирования всех страниц программа OCR преобразует документ в двухцветную или черно-белую версию. Отсканированное растровое изображение анализируется на наличие светлых 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.

# Варианты использования

- Сканирование печатных документов в версии, которые можно редактировать с помощью обычных редакторов текста.
- Индексирование печатного материала для поисковых систем.
- Автоматизированная обработка и ввод данных.



- Расшифровка документов в текст, который может быть прочитан вслух для пользователей с нарушениями зрения.
- Архивирование исторической информации (газет, журналов), а также поиск по ним.
- Извлечение данных и передача в бухгалтерские программы (квитанции, счета).
- Размещение важных подписанных юридических документов в электронной базе данных.

- Распознавание номерных знаков с помощью камеры контроля скорости и программного обеспечения камеры с подсветкой.
- Сортировка писем для доставки почты.
- Перевод слов в изображении на заданный язык.
- Обеспечение поиска отсканированных книг.

# Заключение

- До того, как появилась технология OCR, единственным методом оцифровки бумажных носителей была ручная повторная печать текста. Этот процесс занимал много времени, а также часто приводил к ошибкам при печати. Использование OCR экономит время, помогает исключить ошибки, минимизировать усилия. Кроме этого, технология позволяет выполнять действия, которые недоступны для физических копий, например, может использовать сжатие в ZIP-файлы, выделять ключевые слова, размещать документы на веб-сайте, прикреплять их к электронной почте.

# Литература


- https://beorg.ru/press-centr/opticheskoe-raspoznavanie/
- https://artismedia.by/blog/chto-takoe-ocr-i-dlya-chego-ono-ispolzuetsya/

- Спасибо за внимание