Системы распознавания текста

Технология обработки текстовой информации

Необходимость в системах распознавания символов

- С помощью сканера достаточно просто получить изображение страницы текста в графическом файле. Однако работать с таким текстом невозможно: как любое сканированное изображение, страница с текстом представляет собой графический файл - обычную картинку. Текст можно будет читать и распечатывать, но нельзя будет его редактировать и форматировать. Для получения документа в формате текстового файла необходимо провести распознавание текста, то есть преобразовать элементы графического изображения в последовательности текстовых символов.

Программы распознавания текста

Преобразованием графического изображения в текст занимаются специальные программы распознавания текста (Optical Character Recognition - **OCR**).

Наиболее распространенные системы оптического распознавания символов:

- ABBYY FineReader
- CuneiForm от Cognitive

Получение электронного документа

1. Отсканировать изображение (с помощью ПО сканера);
2. Распознать структуру размещения текста на странице: выделить колонки, таблицы, изображения и т.д.
3. Выделенные текстовые фрагменты графического изображения страницы необходимо преобразовать в текст;
4. Проверка орфографии (если необходимо);
5. Сохранение в файл или передача текста в другое приложение, например в Word.

Методы распознавания символов

- Если исходный документ имеет типографское качество то задача распознавания решается **методом сравнения с растровым шаблоном**.
- При распознавании документов с низким качеством печати используется метод распознавания символов **по наличию** в них **определенных структурных элементов** (отрезков, колец, дуг и др.).

ABBYY FineReader

FineReader - омнифонтовая система оптического распознавания текстов. Это означает, что она позволяет распознавать тексты, набранные практически любыми шрифтами, без предварительного обучения. Особенностью программы FineReader является высокая точность распознавания и малая чувствительность к дефектам печати.

FineReader имеет массы дополнительных функций и удобный интерфес.

Оптимальное разрешение при сканировании

Оптимальным разрешением для обычных текстов является - 300 dpi и 400-600 dpi для текстов, набранных мелким шрифтом (9 и менее пунктов).

Сканирование в сером является оптимальным режимом для системы распознавания. В случае сканирования в сером режиме осуществляется автоматический подбор яркости. Если Вы хотите, чтобы содержащиеся в документе цветные элементы (картинки, цвет букв и фона) были переданы в электронный документ с сохранением цвета, необходимо выбрать цветной тип изображения. В других случаях используйте серый тип изображения.

Вопросы:

- Зачем нужны программы распознавания текста?
- Как происходит распознавание текста?
- Какие программы распознания текста вы знаете? Какими пользовались?
- Какое разрешение является оптимальным для сканирования текста, изображений?