PDF2OCR 簡易マニュアル

イメージ形式で作られた PDF データの文字は、そのままでは選択やコピーなどの処理が行えません。
PDF データを OCR(文字認識)処理してテキストデータ(TEXT,RTF,HTML,XML 形式)に変換することで、その結果に対し検索やコピーなどの処理ができます。
本プログラムは PDF データの文字を OCR 処理して、テキスト文書へ出力するソフトです。

Kernel Computer System
カーネルコンピュータシステム株式会社

本社：パッケージ販売部
〒221-0056
横浜市神奈川区金港町 6-3 横浜金港町ビル
TEL：045-442-0500　FAX：045-442-0501
URL：https://www.kernelcomputer.co.jp

特 徴

- PDF の入力をサポート(PDF2.0 までサポート)
- Adobe ライセンスは必要ありません。
- 文字の認識範囲指定が可能です。(全景も可能)
- 入力されたファイルがマルチページの場合、出力するページを指定する事が可能です。
- 認識した文字列毎の開始位置、文字の高さを出力する事が可能です。
- 認識した文字数を出力する事ができます。
- ページ番号、作成日付などを出力する事ができます。
- OCR 化したテキストデータは以下のフォーマット形式で出力できます。
  - TEXT
  - RTF
  - HTML
  - XML
  - XLS
  - 透明テキスト付き PDF
- 指定したキーワードで入力データを仕分けすることが可能です。

動作環境

Windows 7 (32/64bit)
Windows 8 (32/64bit)
Windows 8.1 (32/64bit)
Windows 10 (32/64bit)
Windows 11 (64bit)
Windows Server 2008 (32bit)
Windows Server 2008 R2
Windows Server 2012
Windows Server 2012 R2
Windows Server 2016
Windows Server 2019
Windows Server 2022

価格

PDF2OCR　：　30 万円(税抜き)

制限事項

デフォルトの学習文字ファイル(MdtOcr.upt)が認識辞書のあるフォルダに
存在するのでしたら、削除して下さい。

入力できる PDF ファイルのバージョンは 2.0 までです。

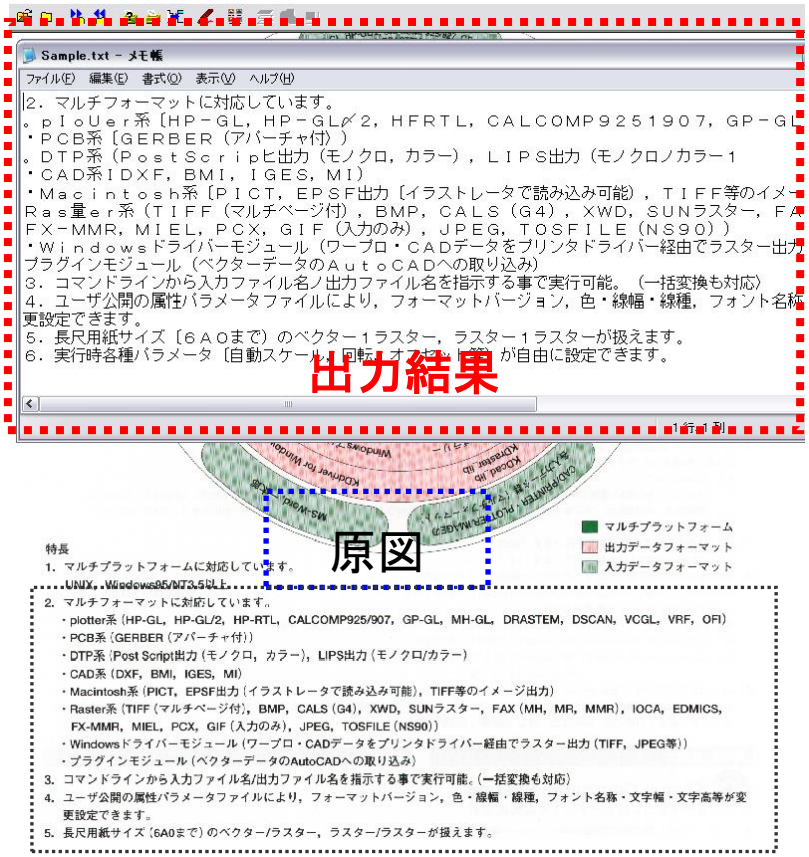
入力データのサイズが A3 を越える場合には、OCR 処理の範囲を指定し OCR 処理を行うか、
-c**を利用し、サイズが A3 に収まるようクリッピング処理を行った上で OCR 処理を行って下さい。
処理範囲の設定では、矩形の辺が 20 ピクセルより小さい場合は無効となります(無視されます)。

活字文書 OCR ライブラリ

本プログラムの OCR 処理は株式会社 NTT データ NJK の「活字文書 OCR ライブラリ」を使用しています。
プログラムを実行する時は OCR 辞書が必要ですので、OCR 辞書の置く場所は必ず指定して下さい。
指定方法は OCR 属性ファイルによって行います。

出力サンプル

・テキストサンプル(原図の下の枠を OCR 処理し、テキストに出力したサンプルです。)



・RTF(リッチテキスト)出力サンプル