xpdf の pdftotext
http://www.foolabs.com/xpdf/
を利用する

Windows

インストール

■ダウンロードと解凍

http://www.foolabs.com/xpdf/
から

```
xpdf-xxxxx-win32.zip
```

をダウンロードし、任意の場所に解凍する。

２．日本語対応

■ダウンロードと解凍

http://www.foolabs.com/xpdf/
から
xpdf-japanese.tar.gz
をダウンロードし、解凍。

以下のディレクトリ構成になるようにファイルを移動する。

```
xpdf root
    ....
    pdftotext.exe
    ....
    japanese
        add-to-xpdfrc
        Adobe-Japan1.cidToUnicode
        EUC-JP.unicodeMap
        ISO-2022-JP.unicodeMap
        README
        Shift-JIS.unicodeMap
        Shift-JIS.unicodeMap_bak

        CMap
```

■設定ファイルの準備

```
add-to-xpdfrc
```

を pdftotext.exe と同じ階層へ移動して、

```
xpdfrc
```

へファイル名を変更する。

・ディレクトリ構成

```
xpdf root
  ....
  pdftotext.exe
  xpdfrc
  ....
```

■設定ファイルの編集

xpdfrc を

```
#----- begin Japanese support package (2004-jul-27)
cidToUnicode  Adobe-Japan1  "D:\Program Files\xpdf-3.02pl4-win32\japanese\Adobe-Japan1.cidToUnicode"
unicodeMap  ISO-2022-JP  "D:\Program Files\xpdf-3.02pl4-win32\japanese\ISO-2022-JP.unicodeMap"
unicodeMap  EUC-JP    "D:\Program Files\xpdf-3.02pl4-win32\japanese\EUC-JP.unicodeMap"
unicodeMap  Shift-JIS  "D:\Program Files\xpdf-3.02pl4-win32\japanese\Shift-JIS.unicodeMap"
cMapDir    Adobe-Japan1  "D:\Program Files\xpdf-3.02pl4-win32\japanese\CMap"
toUnicodeDir      "D:\Program Files\xpdf-3.02pl4-win32\japanese\CMap"
#displayCIDFontTT  Adobe-Japan1  /usr/..../kochi-mincho.ttf
#----- end Japanese support package
```

のように各ファイルへのパスを絶対パスで指定する。
パスにスペースを含む場合は、必ずダブルクォーテーションで括る。

Linux

インストール

基本的には

```
yum install xpdf
```

でＯＫ.

バイナリやソースは
http://www.foolabs.com/xpdf/
からダウンロード出来る。

日本語対応

```
yum install xpdf
```

でインストールした場合は、設定済みなのでこの作業はしなくて良い。

■ダウンロードと解凍

自分でコンパイルした場合は、
http://www.foolabs.com/xpdf/
から
xpdf-japanese.tar.gz
をダウンロードし、任意の場所に解凍。

■設定ファイルの準備

```
add-to-xpdfrc
```

を ホームディレクトリへコピーして

```
.xpdfrc
```

へファイル名を変更する。

■設定ファイルの編集

.xpdfrc を add-to-xpdfrc を参考にパスを編集。

実行

```
pdftotext -enc Shift-JIS test.pdf
```

でテキストが抽出できる。
また、

```
-cfg
```

オプションで 設定ファイルを指定出来る。

半角が全角になってしまう

標準の動作だが気に入らないときは

```
Shift-JIS.unicodeMap
```

を書き換える。具体的には

```
0020 8140
0021 8149
0022 8168
0023 8194
0024 8190
0025 8193
0026 8195
0027 8166
0028 0029 8169
002a 8196
002b 817b
002c 8143
002d 815d
002e 8144
002f 815e
0030 0039 824f
003a 003b 8146
003c 8183
003d 8181
003e 8184
003f 8148
0040 8197
0041 005a 8260
005b 816d
```

```
005c 818f
005d 816e
005e 814f
005f 8151
0060 8165
0061 007a 8281
007b 816f
007c 8162
007d 8170
007e 8160
```

を

```
0020 007e 20
```

に変える。