

# Optical Character Recognition

---

Optical Character Recognition is a technology to convert the text in images or drawings into a machine readable format. It's usually used for scanned and fax documents. i-net PDFC uses this technology for several pre-installed filters. For a more detailed description of the different filter use-cases please visit the [filter documentation](#).

## Tesseract OCR

Tesseract is an open source OCR software which is used by the OCR plugin of i-net PDFC. The current version 4 of Tesseract uses well trained neuronal networks for the recognition process and thus provides a very high accuracy on printed text. Inaccuracies may still occur due to text styling, underlines or background noise.

Please note that Tesseract does neither support nor recognize hand written text.

### Requirements

Although Tesseract is rather robust, there are several requirements for an optimum recognition result:

- The language of the document has to be correctly detected or configured. E.g. if the language of the documents is set to 'English', the software will not recognize non-English characters like umlauts or characters with acutes.
- i-net PDFC is shipped with only English recognition data. Further languages have to be installed manually.
- If images are used for the recognition, the optimum resolution is 300 DPI. This is the default resolution for scanned documents.
- The text background should have a high contrast and little to no noise.
- The text has to be aligned on horizontal baselines.
- The font should be similar to common and pre-installed fonts. Fonts with large serifs may confuse the recognition. Please have a look at the [supported fonts](#) for further details.

### Add additional languages

If another language than English is required, the training data for this language has to be installed manually. To do so please visit <https://github.com/tesseract-ocr/tessdata> and download the corresponding \*.traineddata:file. Once downloaded, please copy the file into <installation folder>/lang/tessdata. Please restart i-net PDFC to activate the language file.

### Performance note

Optical Character Recognition requires a lot of system resources in terms of memory and CPU time. So the filters that rely on OCR should be activated only if required.

### Linux/Mac

For Linux/Mac it required an installed tesseract version 4. See [tesseract install](#). For adding language, should copied the language files in the installation folder to lang\\tessdata.

## Possible Errors

```
read_params_file: parameter not found: enable_new_segsearch
```

Please check the language files of Tesseract. This error is most likely caused by corrupt or outdated language files.

---

```
!strcmp(locale, "C"):Error:Assert failed:in file ../../src/api/baseapi.cpp,  
line 191  
!strcmp(locale, "C"):Error:Assert failed:in file ../../src/api/baseapi.cpp,  
line 201
```

For this is needed to change the environment variables. Give the following line in the shell

```
export LC_ALL=C
```

and start pdfc in this shell.

---

## Windows

If Tesseract doesn't work, it may require additional components to run. Please note that it's important to add both variants: 64-bit / x64 and 32-bit / x86.

<https://www.microsoft.com/en-ca/download/details.aspx?id=48145>