INFO: TesseractEngine - Overview

Deprecation NOTICE:

IN the DotImage world, TesseractEngine refers specifically to Tesseract version 2. (2.0.1 to be precise). This engine was deprecated in DotImage 11.0, and removed from 11.1 - please use Tesseract3Engine for our most recent Tesseract based option

Q104XX - INFO: Tesseract3Engine - Overview (Coming soon)

TesseractEngine

The Tesseract Engine, class name TesseractEngine, is an open source engine that Atalasoft provides without charge for those who purchase the OCR Package. It is a commercial quality OCR engine originally developed at HP between 1985 and 1995. HP and UNLV open-sourced this engine in 2005.

Features

The Tesseract engine is fast and runtime royalty free although it is not quite as powerful as the other engines supported by DotImage. In particular, it lacks segmentation and it is not very good at recognizing low quality documents.

Supported Languages

The TesseractEngine supports the following languages:

- Dutch
- English
- French
- German
- Italian
- Portuguese
- Spanish

Supported Output Formatters

INFO: TesseractEngine - Overview

The TesseractEngine supports the following output formatters and provides a structure that allows you to build your own.

- Text
- PDF

Deployment

The assemblies listed below are required for deployment.

- Atalasoft.dotImage.OCR.Tesseract
- Atalasoft.dotImage
- Atalasoft.dotImage.OCR
- Atalasoft.dotImage.Lib
- System
- System.Data
- System.Drawing

Additionally, the Tesseract language files must be accessible. These are automatically placed in the DotImage directory during toolkit installation. When deploying, you must either copy the OcrResources to your application directory or tell the engine their location explicitly by passing it into the TerractEngine constructor. Please see the TesseractEngine class documentation for additional information.

Example

The Tesseract Engine is used in exactly the same way as the other OCR engines, all of which inherit from the same base class, Atalasoft.dotImage.OCR.

Special Considerations

Once the Tesseract Engine is used, recognize is called with a language, you cannot change to an alternate language. The initialization happens the first time an document in recognized. Attempting to change the language an any time beyond that point results in an exception being thrown.

See Also

OCR Engine

GlyphReaderEngine
RecoStarEngine

Original Article:

Q10363 - INFO: TesseractEngine - Overview

Atalasoft Knowledge Base
https://www.atalasoft.com/kb2/KB/50124/INFO-TesseractEngine-Overview