Table of Contents

Informatica OCR plugin

Image Processing: Need

Vasts amount of critical data remains trapped in image files as they are difficult to process and thus cannot be integrated with other systems. There are no market tools currently available which provide seamless integration of scanned text files with ETL/relational systems.

About the tool

Informatica OCR plugin is a powercenter based tool which leverages the image processing capabilities of ABBYY FineReader and the parsing capabilities of Informatica DT Studio to convert and process image files.

The plugin comprises of a simple powercenter workflow. The workflow consists of a mapping which triggers a DT service. The DT code uses java to invoke the ABBYY engine on the server which does the initial conversion of source files from image to text. The text is stored in memory which then can be parsed by the DT service as per the business requirements and the relevant information returned in a relational format to PowerCenter.

Installation Instructions for ABBYY FineReader

1. Download ABBYY FineReader Engine CLI for Linux(version 9.0 or later) from http://www.ocr4linux.com/en:download (The trial license is valid for processing only 100 pages)
2. Log on to your Linux server using root/admin privileges .Place the file anywhere on your linux server which is accessible to you. To extract the gz file use the following command:
   `$ gunzip file.gz`

3. To untar the .tar file , use the following command:
   `$ tar –xf file.tar`

4. A file called abbyyocr.run will be created in the directory where you executed the tar command. Run the file:
   `$ sh abbyyocr.run`

   By default the program will be installed at /opt/ABBYYOCR

Activating the license:

1. Go to the installation directory(/opt/ABBYYOCR) and look for the file activatefre.sh. Execute the following command to activate your license:
   `$ sh activatefre.sh`

2. Choose to activate using email and enter the serial number you received when you first downloaded the software.
3. Copy the text containing the information required for activation and paste it into the message body of the email addressed to fre-activation-robot@abbyy.com
4. You should receive an email containing the activation file. Paste this file into your installation directory.

*Note: Make sure that folder **/var/lib/frengine9/.elf** and all its contents have the necessary permissions for the non-root user which powercenter uses to execute the workflows.

How to use

To use the OCR_ABBYY, we need to follow the following general steps:

1. Import the workflow **wf_OCR_ABBYY.XML** using the repository manger on the Powercenter client to your destination folder
2. Place the DT service **OCR_V1** at the following path on the server:
   /../9.1.0/DataTransformation/ServiceDB
3. Place the **OCR.jar** file at the following path on the server:
   /../9.1.0/DataTransformation/externLibs/user
4. Place the source file **ocr_source_file_list.txt** at the following server path:
   /../9.1.0/server/infa_shared/SrcFiles
5. Execute the workflow. The output file in the default target file directory **/../9.1.0/server/infa_shared/TgtFiles** should contain a text file **tgt_ocr_output1.out**

Note: The above source file list contains the full path of the image files which need to be processed. These image files should be accessible to the Informatica integration service. A sample source file as well as a typical image file is included in the zip file