Instalacja OCRa dla faktur w systemie eDokumenty

Aktualna dokumentacja od wersji 6.53.0 znajduje się pod poniższym linkiem

OCR dla Faktur

Poniższa instrukcja przedstawia uruchomienie mechanizmu OCRowania faktur w systemie eDokumenty działających na systemie Linux. Mechanizm jest obsługiwany od wersji 5.2.77.

Poniższa instrukcja została przygotowana na bazie systemu Linux Debian9

```
apt-get update
```

```
apt-get install autoconf-archive automake g++ libtool libleptonica-dev pkg-config
apt-get install git
apt-get install poppler-utils
apt-get install libjpeg-dev libtiff-dev libpng-dev
apt-get install zbar-tools
apt-get install libpango1.0-dev
```

instalacja poppler-utils jeżeli wersja z dystrybucji jest mniejsza niż 0.86.0:

```
sudo apt-get purge poppler-utils
sudo apt-get install libopenjp2-7-dev libgdk-pixbuf2.0-dev cmake checkinstall
sudo apt-get install libfreetype6-dev libfontconfig-dev libcairo2-dev
sudo apt-get install debhelper dpkg libboost-dev libglib2.0-dev libfontconfig1-dev libjpeg-dev libpng-dev libtiff-dev libl
mkdir /usr/lib/poppler_utils
cd /usr/lib/poppler_utils
wget https://poppler.freedesktop.org/poppler-0.86.1.tar.xz
tar -xf poppler-0.86.1.tar.xz
cd poppler-0.86.1
mkdir build
cd build
cmake ..
sudo checkinstall make install
ldconfig
```

Komenda do przeprowadzenie testu popplera.

```
pdftotext -bbox-layout NAZWAPLIKUWEJSCIOWE.pdf NAZWAPLIKUWYJSCIOWEGO.html
```

Jeśli pakiety leptonica 1.74+ nie są dostępne w dystrybucji w takim przypadku, konieczna będzie komplikacja ze źródeł

```
mkdir /usr/lib/leptonica
cd /usr/lib/leptonica
wget http://www.leptonica.org/source/leptonica-1.82.0.tar.gz
gunzip leptonica-1.82.0.tar.gz
tar -xf leptonica-1.82.0.tar
cd leptonica-1.82.0
./configure
make
make install
```

```
mkdir /usr/lib/tesseract
cd /usr/lib/tesseract
git clone https://github.com/tesseract-ocr/tesseract.git tesseract-ocr
cd tesseract-ocr/
./autogen.sh
./configure
```

```
make
make install
ldconfig
```

```
cd /usr/local/share/tessdata/
wget https://github.com/tesseract-ocr/tessdata_fast/raw/master/script/Latin.traineddata
wget https://github.com/tesseract-ocr/tessdata_fast/raw/master/pol.traineddata
wget https://github.com/tesseract-ocr/tessdata_fast/raw/master/eng.traineddata
wget https://github.com/tesseract-ocr/tessdata_fast/raw/master/osd.traineddata
```

alternatywne źródło do pobrania:

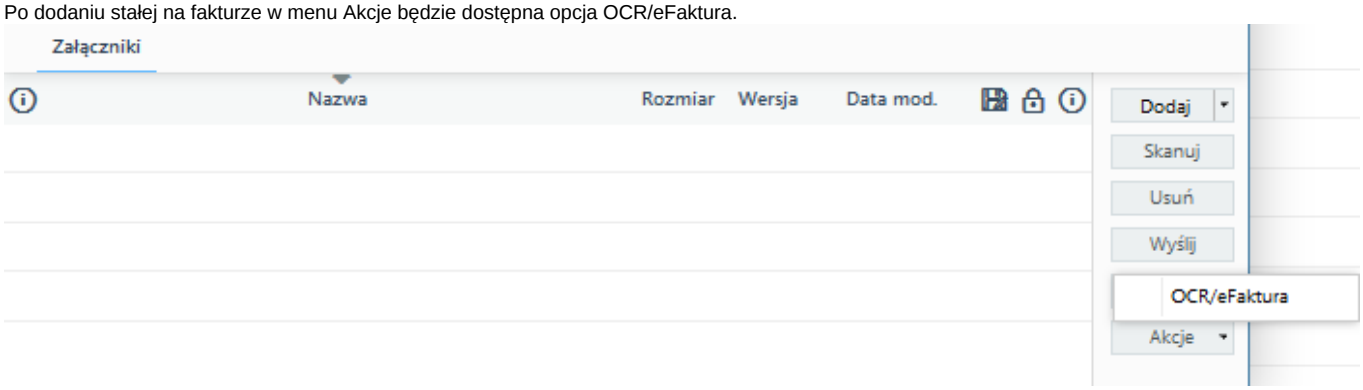
```
wget https://raw.githubusercontent.com/tesseract-ocr/tessdata/main/script/Latin.traineddata
wget https://raw.githubusercontent.com/tesseract-ocr/tessdata/main/pol.traineddata
wget https://raw.githubusercontent.com/tesseract-ocr/tessdata/main/eng.traineddata
wget https://raw.githubusercontent.com/tesseract-ocr/tessdata/main/osd.traineddata
```

Po pobraniu, zainstalowaniu oraz skompilowaniu pakietów ostatnim elementem jest dodanie stałej w config.inc domyślnie

```
vim /home/edokumenty/public_html/apps/edokumenty/config.inc
```

```
define('USE_NEW_OCR_FOR_EINVOICE', TRUE);
```

Po dodaniu stałej na fakturze w menu Akcje będzie dostępna opcja OCR/eFaktura.



Pakiety niezbędne do działania Bufora OCR - Python 2

```
apt-get install rabbitmq-server
apt-get install supervisor
apt-get install python-opencv
apt-get install python-pip
pip install pika
apt install python-pil

# pdftk
apt-get install pdftk
```

Pakiety niezbędne do działania Bufora OCR - Python 3 (dla systemu Debian 10)

```
apt-get install rabbitmq-server
apt-get install python3-opencv
apt-get install python3-pip
apt-get install python3-pika
apt-get install python3-toml
apt install python3-pil
apt-get install supervisor
```

```
# pdftk
apt-get install pdftk
```

Poniższy krok można pominąć. Poszczególny skryptu supervisora możemy wywołać poprzez polecenie python3. Zmiana domyślnej wersji.

```
sudo update-alternatives --install /usr/bin/python python /usr/bin/python3.6 2
sudo update-alternatives --config python
```

Użycie wersji jako domyślnej

```
sudo update-alternatives  --set python /usr/bin/python3.6
```

Znane problemy:

1. Brak pakietu libpng12.so.0. W logach OCR pojawia się komunikat:

tesseract: error while loading shared libraries: libpng12.so.0: cannot open shared object file: No such file or directory

Sprawdzamy czy pakiet istnieje:

```
ls -ld $(locate -r libpng.*\.so.*)
```

Komenda powinna zwrócić nam:

```
lrwxrwxrwx 1 root root     19 kwi 18 22:12 /usr/lib/x86_64-linux-gnu/libpng16.so -> libpng16.so.16.28.0
lrwxrwxrwx 1 root root     19 kwi 18 22:12 /usr/lib/x86_64-linux-gnu/libpng16.so.16 -> libpng16.so.16.28.0
-rw-r--r-- 1 root root 206768 kwi 18 22:12 /usr/lib/x86_64-linux-gnu/libpng16.so.16.28.0
lrwxrwxrwx 1 root root     11 kwi 18 22:12 /usr/lib/x86_64-linux-gnu/libpng.so -> libpng16.so
```

Jeśli otrzymamy taką informację konieczne będzie ponowne kompilowanie leptonici oraz tesseract

Kompilowanie tesseract dla 1 wątku

```
./configure --disable-openmp
```

1. Problem z convertowanie jpg do PDF

W logach php mamy komunikat

```
[23-Sep-2020 12:28:46 Europe/Warsaw] ReadyCls\OCR\OcrEngine - pdftoppm fails with message: [1]
```

lub

```
convert-im6.q16: attempt to perform an operation not allowed by the security policy `PDF' @ error/constitute.c/IsCoderAuth
```

W pliki vim /etc/ImageMagick-6/policy.xml należy zakomentować linię

```
<policy domain="coder" rights="none" pattern="PDF" />
```

Przetwarzanie w tle (Bufor OCR)

Dotyczy Ready_ w wersji 6.52.1+

Wykorzystujemy supervisor do uruchomienia dwóch workerów (skrypty w języku Python), które znajdują się w katalogu domowym systemu (najczęściej: /home/edokumenty/bin).
Skrypty to: **worker_ocr.py** oraz **ocr_result.py**

Domyślne konfiguracje umieszczone są w katalogu home/edokumenty/etc/. Przed uruchomieniem należy usunąć z nazwy _default.

BUFFOR OCR osobna maszyna

W celu rozłożenie obciążenia, które w dużym stopniu generuje OCR możemy wydzielić go na osobną maszynę.

W tym celu na środowisku gdzie działa RabbitMQ tworzymy nowego użytkownika i nadajemy mu odpowiednie uprawniania:

```
rabbitmqctl add_user UZYTKOWNIK HASLO
rabbitmqctl set_user_tags UZYTKOWNIK administrator
rabbitmqctl set_permissions -p / UZYTKOWNIK ".*" ".*" ".*"
```

Następnie dane do nowo utworzonego konta uzupełniamy w konfiguracji na maszynie eDokumentyOCR

```
vim /home/edokumenty/etc/rabbimq.toml
```

Po uzupełnieniu danych konieczny jest restart workerów

```
supervisorctl reload
```