Repozytorium Cyfrowe BN

OCR

Repozytorium współpracuje z wybranymi rozwiązaniami OCR w celu uzyskania warstwy tekstowej z obiektów, które zawierają postać graficzną tekstu (skany, zdjęcia). Na chwilę obecną wdrożona jest współpraca z dwoma rozwiązaniami:

- ABBYY Recognition Server (wersja 3.5)
- DocWorks

1. ABBYY Recognition Server

Opis oprogramowania (istotny z punktu widzenia RCBN)

ABBYY Recognition Server (RS) dostarcza kilku metod współpracy które są opisane szczegółowo w dokumentacji. W przypadku RCBN współpraca z RS została oparta na wymianie plików przy pomocy współdzielonych katalogów. Rozróżniamy trzy typy katalogów w tym podejściu:

- Katalog wejściowy (IN) – jest to katalog w którym umieszczane są zadania do przetworzenia przez RS. Wraz z plikami można umieścić plik XML (tak zwany XML ticket) który poza wskazaniem kompletnej listy plików należących do danego zadania będzie również określał dodatkowe niestandardowe parametry (nadpisujące parametry określone w danym profilu)
- Katalog wyjściowy (OUT) – RS umieszcza w nim gotowe pliki wynikowe wraz z plikiem XML opisującym szczegóły zadania i wyników
- Katalog błędów (EXC) – RS umieszcza w nim pliki źródłowe zadania w którym napotkał na problemy wraz plikiem XML opisującym napotkany problem.

W ABBYY RS można określić wiele profilów w których każdy ma odrębne katalogi IN, OUT oraz EXC. Profil to zestaw predefiniowanych ustawień z jakimi zadanie zostanie przetworzone. W ramach tych ustawień możemy wyróżnić między innymi (wybrane ustawienia i skrócony opis):

- Określenie metody współpracy (tutaj katalogi współdzielone)
- określenie czasu po którym zadanie zostanie podjęte (licząc od daty ostatniej modyfikacji katalogu/pliku), czas ten zapobiega przedwczesnemu podjęciu zadania kiedy system współpracujący jeszcze jest w trakcie wysyłania materiału
- określenie katalogów IN, OUT i EXC
- określenie czy pliki wynikowe będą w formie jeden plik każdego wskazanego formatu na całe zadanie (forma scalona) jeden plik dla każdego wskazanego formatu na każdy plik w zadaniu
- wskazanie słowników z jakich RS ma korzystać podczas procesu rozpoznawania. Tutaj należy się uwaga iż nie należy wybierać zbyt wielu słowników jednocześnie w ramach pojedynczego profilu gdyż skutkuje to pogorszeniem jakości wyników zamiast ich polepszeniem oraz wpływa znacząco na czas przetwarzania

- określenie czy ma być uwzględniona w procesie weryfikacja oraz po spełnieniu jakich warunków weryfikacja ma się uruchomić (zawsze, nigdy, po wystąpieniu określonego % nierozpoznanych znaków)
- jakość OCRa – wpływa na czas przetwarzania
- określenie plików wynikowych

W powyższych ustawieniach występuje taka definicja jak „weryfikacja" . Polega to na tym, iż w przypadku zakwalifikowania danego zadania do etapu weryfikacji, wyniki procesu OCR przed opublikowaniem (np. w katalogu OUT) zostaną poddane weryfikacji przez użytkownika. Weryfikacja następuje poprzez dedykowane oprogramowanie zainstalowane na komputerze klienta (użytkownika) w którym użytkownik może podjąć wybrane zadanie, podejrzeć przetworzone wyniki i wprowadzić odpowiednie poprawki w razie potrzeby. Dopiero po zaakceptowaniu wyniku OCR przez kontrolera wyniki danego zadania zostaną opublikowane w katalogu OUT skąd system współpracujący będzie mógł je pobrać i zaaplikować do swojej bazy. Krok ten jest o tyle istotny iż znacząco wpływa na wydłużenie procesu OCR – zadanie będzie tak długo w trakcie procesu dopóki ktoś go nie podejmie, opracuje i zaakceptuje (bądź odrzuci – co skutkuje wrzuceniem zadania do katalogu EXC).

Opis sposobu współpracy

Obiekt można wysłać do procesowania w oprogramowaniu ABBYY RS z dwóch miejsc:

- etap workflow „wybór formy prezentacji"
- zakładka OCR dostępna w edycji obiektu

Jak wcześniej wspomniano RCBN współpracuje z ABBYY RS przy pomocy współdzielonych katalogów. Zasoby każdego obiektu wytypowanego do procesu OCR przy pomocy RS jest przez RCBN wgrywane są odpowiedniego katalogu wejściowego, który to katalog obserwuje RS i po wgraniu materiału rozpoczyna proces rozpoznawania z określonymi parametrami. Każdy obiekt jest wgrywany do oddzielnego katalogu którego nazwa jest tworzona według wzoru: „UID_TASKID" – w przypadku obiektów wysyłanych w trakcie workflow; „UID" - w przypadku obiektów wysyłanych z formatki OCR dostępnej w edycji obiektu. Do procesu OCR wysyłane są wszystkie pliki źródłowe contentów podstawowych znajdujących się w sekcji „paginacyjnej" obiektu, niezależnie od formatu. Pomijane są natomiast te pliki które dotyczą contentów ukrytych. Nazewnictwo wgrywanych plików jest odpowiednio określone według następującego wzoru: … . Wraz z plikami do katalogu wgrywany jest specjalny plik XML (tzw. XMLticket) w którym zawarta jest lista wszystkich plików wchodzących w ramach danego zadania. Ten plik jest wgrywany jako pierwszy i dzięki zdefiniowaniu w nim całej listy plików do przetworzenia zapobiega się sytuacji (zabezpieczenie dodatkowe) w której ABBYY RS rozpoczęłby procesowanie tego obiektu przed wgraniem wszystkich plików. W pliku XML można również przekazać dodatkowe parametry nadpisujące ustawienia profilu zdefiniowane w oprogramowaniu ABBYY RS

jednakże na chwilę obecną RCBN nie korzysta z tej możliwości i użytkownik nie ma możliwości ustawienia tych parametrów od strony GUI RCBN.

Po przetworzeniu wyniki są umieszczane w oddzielnym katalogu który z kolei jest monitorowany przez RCBN i po wgraniu gotowych wyników importuje je do systemu. Są zdefiniowane określone zasady umieszczania wyników OCR w obiekcie:

- W przypadku wyników per content pliki ALTO oraz TXT są umieszczane w contencie bazowym jako alternatywny stream (pliki ALTO pod kluczem „download_alto", zaś pliki txt pod kluczem „text") pozostałe formaty są umieszczane przy contencie bazowym jako content alternatywny. Podczas dopasowywania pomijane są te contenty które są oznaczone jako ukryte.
- W przypadku wyników per content, jeżeli mamy do czynienia z plikami typu PDF, EPUB lub MOBI i dla danego profilu skonfigurowanego po stronie RCBN jest ustawienie wskazujące aby wyniki poddawać procesowi scalania, wówczas system nie umieszcza ich jako contenty alternatywne tylko wszystkie pojedyncze pliki scala do jednego pliku i umieszcza jako content zbiorczy. Pliki PDF są scalane za pomocą skryptu …. . Pliki EPUB i MOBI są scalane za pomocą skryptu Epubmerge.py Każdy z tych skryptów kopiuje sobie pliki na udział lokalny po czym łączy wszystkie pliki i zapisuje pod nazwą składającą się z odpowiednio spreparowanej metadanej „title".
- W przypadku wyników per zadanie (obiekt), plik jest wgrywany jako content zbiorczy (kolejny w kolejności). Dodatkowo jeżeli jest to plik PDF to tekst z każdej kolejnej strony jest wyciągany i zapisywany do streamu alternatywnego „text" przy contencie bazowym odpowiadającym kolejności stronie w pliku PDF (zachowując zasadę iż contenty ukryte się nie liczą).

W przypadku kiedy zadanie w OCR zostanie zakwalifikowane jako błędne ABBYY umieści pliki w odpowiednim katalogu przeznaczonym na zadania błędne. Ten katalog jest również monitorowany przez RCBN i w razie wykrycia wadliwego zadania odnotuje ten fakt w swojej bazie. Jeżeli obiekt był wysłany o procesu OCR w ramach workflow wówczas system oznaczy odpowiednie zadanie workflow dotyczące OCR jako zakończone i w zależności od ścieżki workflow przeniesie dane zadanie do odpowiedniego kroku następnego. Dodatkowo nieprawidłowość zostanie odnotowana poprzez dodanie do listy obiektów które nie przeszły procesu OCR.

Na potrzeby weryfikacji prowadzony jest odpowiedni dziennik zapisywany w bazie danych w tabeli „ocr_status". Każde zadanie wysyłane do OCR jest wpisywane do tej tabeli z informacją kiedy zostało wysłane, jaką ścieżką oraz ustawiony status na „IN_PROCESS". W momencie wysyłania każdego obiektu do OCR sprawdzana jest najpierw ta tabela pod kątem czy dany obiekt nie jest przypadkiem wysłany do OCR – jeżeli tak to ponowne wysłanie do OCR zostanie zaniechane, co zapobiega wysłaniu wiele razy tego samego dokumentu do procesu OCR. W chwili odebrania informacji o zakończeniu procesowania OCR w bazie odszukiwana jest informacja o ostatnim statusie wysłania obiektu do OCR i

wpis zostaje zaktualizowany o datę odebrania wyników oraz zaktualizowany jest status: „SUCCESS" kiedy wynik był w katalogu „OUT" lub „FAIL" jeżeli wynik był w katalogu „EXC".

Jak wspomniano wcześniej, po stronie oprogramowania ABBYY RS można zdefiniować szereg profilów. Każdy z tych profilów które chcemy dać użytkownikowi do użytku w RCBN musimy najpierw skonfigurować po stronie Repozytorium. Konfigurację tą zapisujemy w pliku academica.properties która wygląda następująco:

```
ocr1.name=byPage_ALTO_PDF-70/300
ocr1.inDir=smb://172.26.106.50/ocr-workspace-aca/acaOcrP1_in/
ocr1.outDir=smb://172.26.106.50/ocr-workspace-aca/acaOcrP1_out/
ocr1.excDir=smb://172.26.106.50/ocr-workspace-aca/null/
ocr1.merge=true
ocr2.name=merged_imagePDF
ocr2.inDir=smb://172.26.106.50/ocr-workspace-aca/acaOcrM1_in/
ocr2.outDir=smb://172.26.106.50/ocr-workspace-aca/acaOcrM1_out/
ocr2.excDir=smb://172.26.106.50/ocr-workspace-aca/null/
ocr2.merge=false
ocr.user=userName
ocr.password=passwordValue
ocr.retrieve.enabled=true
epub.merge.command=/opt/academica/resources/epubmerge.py
ocr.sendingThreads=2
ocr.enabled=true
ocr.cron=0 */5 * * * ?
```

gdzie:

- ocr1, ocr2 – określa kolejny zdefiniowany profil
- ocrX.name – nazwa profilu
- ocrX.inDir – katalog IN profilu
- ocrX.outDir – katalog OUT profilu
- ocrX.excDir – katalog EXC profilu
- ocrX.merge – określa czy wybrane pliki należy scalać
- ocr.user – nazwa użytkownika do łączenia się do udziałów SMB
- osr.password – hasło użytkownika do łączenia się do udziałów SMB
- ocr.retrieve.enabled – określa czy dana instancja RCBN powinna brać udział w procesie pozyskiwania wyników OCR
- epub.merge.command – wskazuje ścieżkę do pliku używanego do łączenia EPUB
- ocr.sendingThreads – liczba wątków wysyłających zgłoszone obiekty do OCR
- ocr.enabled – określa czy dana instancja RCBN bierze udział w wysyłaniu zgłoszonych obiektów do OCR
- ocr.cron – określa z jaką częstotliwością system ma sprawdzać katalogi OUT i EXC

Na chwilę obecną na serwerach produkcyjnych są dostępne następujące ścieżki OCR:

- **byPage_ALTO_PDF** – pliki wynikowe ALTO dla każdej strony oraz PDF zbiorczy dla całej publikacji, włączony słownik polski i angielski, bez weryfikacji
- **byPage_ALTO** – pliki wynikowe ALTO dla każdej strony, włączony słownik polski i angielski, bez weryfikacji
- **merged_imagePDF** – wynikiem jest zbiorczy PDF jeden na całą publikację bez warstwy tekstowej (moim zdaniem jedna z bardziej bezsensownych ścieżek), bez weryfikacji
- **byPage_ALTO_PDF_verify** – pliki wynikowe ALTO dla każdej strony oraz PDF zbiorczy dla całej publikacji, włączony słownik polski i angielski, w tej ścieżce każda pozycja musi przejść etap weryfikacji przez redaktora przed wytworzeniem wyników (weryfikacja odbywa się w oddzielnym oprogramowaniu instalowanym dla wybranych osób)
- **byPage_pref** – jest to to samo co byPage_ALTO_PDF tylko ma wyższy priorytet jak już zadanie zostanie wgrane na serwer OCR (w repozytorium nie ma wyższego priorytetu), wynikiem są pliki ALTO dla każdej strony + jeden zbiorczy PDF na całą publikację, włączone słowniki polski i angielski, bez weryfikacji
- **byPage_ALTO_PDF_latin** – pliki wynikowe ALTO dla każdej strony oraz PDF zbiorczy dla całej publikacji, włączony słownik polski, angielski i łaciński, bez weryfikacji
- **byPage_ALTO_PDF_latin_verify** – pliki wynikowe ALTO dla każdej strony oraz PDF zbiorczy dla całej publikacji, włączony słownik polski, angielski i łaciński, w tej ścieżce każda pozycja musi przejść etap weryfikacji przez redaktora przed wytworzeniem wyników (weryfikacja odbywa się w oddzielnym oprogramowaniu instalowanym dla wybranych osób)
- **byPage_ALTO_latin** – pliki wynikowe ALTO dla każdej strony, włączony słownik polski, angielski i łaciński, bez weryfikacji