Usługa DIGITAL_TAX – opis zadania

Problem: Krajowa Administracja Skarbowa, mimo rozwoju usług publicznych dostępnych za pomocą kanałów elektronicznych, musi przetwarzać informacje podatkowe składane przez Klientów KAS w formie papierowej. **W I półroczu 2020r. KAS przyjął blisko 5 mln deklaracji w formie papierowej (w tym deklaracji PIT-37 ponad 2 mln.)**. Przetworzenie danych otrzymywanych w postaci papierowej do postaci cyfrowej jest czasochłonne i generuje koszty po stronie KAS. Często z uwagi na czasochłonność rezygnuje się z wprowadzania wszystkich danych do systemów KAS, co może negatywnie wpływać na zdolności analityczne służb skarbowych.

Pomysł: Dostarczenie rozwiązania wykorzystującego technologię transformacji tekstu pisanego i drukowanego na edytowalną postać cyfrową. Mechanizm na podstawie spodziewanej wartości w danym polu deklaracji, wnioskuje z dokumentu drukowanego i zapisuje tekst, dane w formie zdigitalizowanej, zapewnia podniesienie skuteczności automatycznej konwersji od skanu do pliku XML poprzez wykorzystanie w procesie rozpoznawania tekstu danych z Systemów Krajowej Administracji Skarbowej (w tym słowników). Produkt końcowy umożliwia dalsze przetwarzanie danych cyfrowych w innych aplikacjach i systemach KAS.

Założenia podstawowe

1. Budowa oprogramowania wykorzystującego technologię OCR (optical character recognition) czyli zestaw technik służących do rozpoznawania znaków w pliku graficznym o postaci rastrowej oraz technologię ICR (Intelligent Character Recognition) pozwalającą odczytywać blokowe pismo ręczne do rozczytywania deklaracji podatkowych składanych w formie papierowej.

2. Mechanizm transformacji ma być na tyle czuły, że do sporządzenia obrazu dokumentu pracownicy jednostek KAS wykorzystywać będą obecnie posiadany (średniej jakości) sprzęt do skanowania wspierający proces przetwarzania danych z formularzy składanych w wersjach papierowych.

3. W kwestii aspektów technicznych istotne będzie oparcie propozycji o rozwiązania open source (nieoferowane aktualnie na rynku jako kompletne rozwiązania).

Celem zadania jest przede wszystkim uzyskanie rozwiązania zapewniającego powiązanie między plikiem skanu (wejście do procesu), wynikiem procesu OCR (produkt pośredni), plikiem XML (plik wynikowy z procesu zawierający deklarację podatkową w formie cyfrowej) oraz podniesienie skuteczności automatycznej konwersji od skanu do pliku XML poprzez wykorzystanie w procesie rozpoznawania tekstu danych z Systemów Krajowej Administracji Skarbowej (w tym słowników).

Celem jest więc dostarczenie jednostkom KAS rozwiązania pozwalającego na przekształcenie wersji drukowanej deklaracji w jej edytowalną postać cyfrową (z poprawnością rozpoznawania prawidłowości tekstu na jak najwyższym poziomie). Deklaracje podlegające przekształceniu składane są do KAS w formie papierowej, zwykle wypełnione są ręcznie, ewentualnie są to wydruki z edytorów tekstu bądź programów do wypełniania deklaracji oferowanych przez podmioty niezależne. Zaletą jest fakt, że każdy z tych dokumentów ma formę wzoru deklaracji publikowanego przez Ministerstwo Finansów, więc dokument taki zawiera pozycje deklaracji wg schematu wymaganego przez MF. Rozwiązanie jest możliwe do

szybkiego wdrożenia w KAS z uwagi na ustalony wzór deklaracji podatkowych (układ określony w przepisach prawa).

Rozwiązanie wypracowane w formie usługi DIGITAL_TAX, po przekonwertowaniu obrazu na tekst (zeskanowanego dokumentu w pliku pdf, jpg otrzymanego drogą elektroniczną) przenosi dane do systemu KAS w odpowiednie pola. Algorytmy będące częścią systemu, przeszukują rozpoznaną treść w celu znalezienia określonych parametrów np. NIP, adres, przychód, koszty uzyskania przychodu, co umożliwia systemowi zaklasyfikowanie wyrażeń znajdujących się w okolicy słów kluczowych do właściwych pól. Dodatkowo system weryfikuje format danych (np. czy dana wartość jest datą), cyfrę kontrolną (np. NIP), a także analizuje zgodność rozpoznanych danych ze słownikami czy bazami danych KAS.

System wspomaga się danymi z systemów Krajowej Administracji Skarbowej m.in.: danymi z rejestru podatników i płatników, systemu poboru podatków ale również słownikiem adresów TERYT, słownikiem organizacji pożytku publicznego, słownikiem walut NBP i innymi.

Na potrzeby hackathonu dane z powyższych systemów i słowników zasymulowane zostały danymi dołączonymi w pliku XLS, jednakże rozwiązanie musi być gotowe także do pobierania tych danych poprzez API (budowa API programistycznego będzie wyżej punktowana).

Treści rozpoznane przez system OCR-ujący, które znajdują się wokół kluczowych słów zyskują najwyższy wskaźnik prawdopodobieństwa poprawności, co przy jednoczesnym wykorzystaniu technologii ICR czyli narzędzi odwołujących się do dodatkowych źródeł danych zwiększa efektywność rozpoznanego tekstu. Oczekiwany format wyjściowy danych to format XML zgodny ze strukturą danej deklaracji – analogicznie do deklaracji elektroniczne składanych przez podatnika z wykorzystaniem istniejących rozwiązań (system e-Deklaracje).

Deklaracje przekonwertowane z ich postaci papierowej do edytowalnej postaci cyfrowej (po OCR i ICR) zostają skierowane do procesu korekty przez użytkownika. Rozwiązanie ‘podpowiada’ użytkownikowi tekst co do którego nie ma pewności, że został rozpoznany właściwie - oznaczając te obszary tekstu – tak by użytkownik skupił się na weryfikacji obszarów oznaczonych jako niepewne. **Jednocześnie rozwiązanie ma zapewniać możliwość modyfikacji przez użytkownika całości danych (nie tylko obszarów wskazanych jako obszary do korekty).**

Dążyć należy do tego by użytkownik weryfikował poprawność przekształcenia danych, korygując ‘propozycję’, którą uzyska dzięki rozwiązaniu, o ile te zawierają błędy zgłoszone przez system. Mechanizm podpowiada, sygnalizuje użytkownikowi, które z pozycji mogą zawierać błędne wpisy – do weryfikacji przez użytkownika i naniesienia przez niego ewentualnej korekty.

Wejściem do zadania są przygotowane przez MF/AK:

1. zeskanowane deklaracje PIT-37
2. Schemy xsd w wersji 25 i 26 formularza PIT-37
3. Przykładowe wynikowe Pliki xml wybranych deklaracji
4. Dane testowe z Systemów KAS I słowników (format XLS).

← Wyjściem:

1. GUI dla użytkownika (pracownika KAS), w którym ma on możliwość poprawienia danych, z wyróżnionymi obszarami słabo rozpoznanymi – tak by użytkownik mógł rozstrzygnąć o ostatecznych wartościach, które zostaną utrwalone w systemach KAS.
2. Pliki XML deklaracji, zgodne z dostarczonymi XSD, z danymi zaakceptowanymi ostatecznie przez użytkownika.

W GUI o którym mowa w punkcie 1 powyżej, użytkownik ma wgląd, w kontekście danej deklaracji, do:

- skanu danej deklaracji
- udostępnionego wygenerowanego pliku wynikowego w formacie XML po OCR zgodnego ze schemą XSD
- treści wynikowej w postaci zwizualizowanej deklaracji, gdzie treści OCR zawierają jak najmniej obszarów, których nie udało się rozpoznać prawidłowo, a w razie trudności w rozstrzygnięciu poszczególnych zapisów, jakość rozpoznania treści deklaracji została poprawiona w oparciu o dane z systemów KAS (w tym dane podatnika, słownik TERYT, OPP itp.), jak i zgodne ze strukturą danej wersji schemy xsd formularza. Poprawność rozpoznawania tekstu powinna być udoskonalona także w oparciu o warunki opisane na samych drukach deklaracji, takie jak zależności między pozycjami (np. w danej pozycji oczekiwana jest suma dwóch poprzedzających).

Ostatecznie obszary, które nie udało się rozpoznać właściwie (pomimo sprawdzenia jak wyżej) wyróżnione są spośród pozostałego tekstu – tak by użytkownik mógł je poprawić i zdecydować o ostatecznych wartościach, jakie zostaną wprowadzone do systemów KAS (czyli zapisane w formie wynikowego pliku XML). Poza pozycjami wątpliwymi użytkownik musi także mieć możliwość poprawienia pozostałych pozycji deklaracji.

Ekran na którym użytkownik pracuje by obejrzeć wyniki prac ICR to ekran na którym użytkownik podejmuje decyzje co do treści danych, które ostateczne zostaną utrwalone w systemach KAS – stąd kluczowa dla oceny tego rozwiązania jest ergonomia tego ekranu i przepustowość procesu decyzyjnego. Zasadniczym wyznacznikiem sukcesu tego procesu będzie więc wykonanie transformacji papieru do wersji cyfrowej przy jak najmniejszym obciążeniu operatora systemu pracą ręczną.

Dodatkowo - ważne by zaprojektowane rozwiązanie oznaczało (zapisywało informacje o faktycznym operatorze: automat bądź konkretny użytkownik), które rekordy wprowadził automat, a w które ingerował użytkownik i je ostatecznie zmodyfikował.

Ostatecznym wynikiem działania usługi jest plik xml zawierający dane ze skanu deklaracji podatkowej (istotne jest żeby jak najlepiej odwzorować dane z deklaracji). Zaprojektowane rozwiązanie ma przetworzoną deklarację papierową w jej wynikowej postaci xml (zgodnej ze strukturą danej wersji schemy xsd) przekazywać na zewnątrz aplikacji.

Rozwój rozwiązania

Planuje się, że rozwiązanie docelowo obejmie swoim zasięgiem pozostałe deklaracje podatkowe, stąd usługa musi być zarówno skalowalna (wg aktualnych danych w formie papierowej wpływa do KAS ok. 5 mln deklaracji rocznie) jak i konfigurowalna – tak żeby mogła przy zerowej albo minimalnej ingerencji w kod obsłużyć także inne deklaracje (opisywane w formie schem XSD). W kwestii aspektów technicznych istotne będzie oparcie propozycji o rozwiązania open source, a w szczególności o rozwiązania innowacyjne (nieoferowane aktualnie na rynku jako kompletne rozwiązania). Architektura rozwiązania musi więc przygotować rozwiązanie do:

1. Integracji z zewnętrznym systemem obiegu dokumentów/źródła skanów.
2. Integracji z zewnętrznym systemem wykonującym OCR/ICR (przekształcenie skanu na tekst).
3. Rozproszenia architektury wydajnościowego i geograficznego (skalowalność wszerz).
4. Integracji z systemami KAS.
5. Zagadnieniami bezpieczeństwa (ochrona danych wrażliwych, ograniczony dostęp do danych, zabezpieczenie danych na każdym etapie przetwarzania).
6. Wdrożenia pełnej rozliczalności działań użytkownika/Systemu w procesie.

<u>Złóż swój projekt wraz z następującymi plikami:</u>

1. PROJECT NAME_ Project description.ZIP [obowiązkowy]:
   - Opis projektu w jęz. angielskim lub jęz. polskim
   - Max 500 słów

2. PROJECT NAME_PRESENTATION. ZIP [obowiązkowy]:

   - Folder zawierający informacje, materiały, opisy, screenshots, prezentacje, linki do filmów i innych wizualizacji Twojego projektu, interfejs prototypu itp.,
   - Prezentacja (w jęz. angielskim lub jęz. polskim, format PDF, max 10 slajdów).

3. PROJECT NAME_ MANUAL.ZIP : (obowiązkowy)

Video URL (obowiązkowo)
- Przygotuj video dot. projektu lub video gdzie opowiecie o Waszym projekcie
- Max 60 sec.
- w jęz. angielskim lub jęz. polskim

Link do Demo projektu (obowiązkowy)
- Pamiętaj o przekazaniu danych do logowania dla JURY

Repozytorium URL (obowiązkowy)
- Całość w jednym repozytorium (jeśli Twój projekt składa się z kilku części, modułów upewnij się, że zamieściłeś je w oddzielnych folderach)
- Zapewnij możliwość sprawdzenia on-line
- Repozytorium zawierające projekt rozwiązania lub link do niego wraz z opisem krok po kroku jak uruchomić rozwiązanie w Windows lub Linux,
- Zalecany skrypt automatyzujący proces,
- Lista frameworków, kompilatory, maszyny wirtualne, kontenery, środowisko uruchomieniowe, wymagania sprzętowe,
- Lista dodatkowych narzędzi, bibliotek i dodatkowych programów (np. bazy danych), z których korzysta aplikacja,
- Instrukcja uruchomienia – co należy zrobić, żeby uruchomić projekt,
- Instrukcja obsługi – dla administratora – gdzie/jak powinny być dostarczane skany, jak się zalogować do aplikacji, gdzie znaleźć xml po ocr a gdzie ostatecznie poprawione,
- Podstawowa, skrócona instrukcja obsługi dla użytkownika – jak posługiwać się aplikacją,
- Opis architektury.

4. OTHER Files. ZIP (obowiązkowy):

   - wygenerowany plik wynikowy w formacie XML po OCR zgodny ze schemą XSD
   - plik XML deklaracji, zgodne z dostarczonymi XSD, z danymi zaakceptowanymi ostatecznie przez użytkownika. Produkt końcowy usługi to plik XML zawierający dane

ze skanu deklaracji (ważne, zachowaj poprawność rozpoznawania prawidłowości tekstu na jak najwyższym poziomie)