Znak sprawy: XIV.263.**1**.2019 **Załącznik nr 1a do SIWZ**

**Załącznik nr 1 do OPZ**

**Zakres równoważności oprogramowania OCR**

1. Parametry funkcjonalne
   1. Praca z pełną wymaganą funkcjonalnością na serwerze Zamawiającego bez przesyłania jakiejkolwiek zawartości plików podlegających procesowi OCR poza sieć wewnętrzną Zamawiającego.
   2. Równoległe przetwarzanie w wielu wątkach na wielu fizycznych rdzeniach procesorów 64-bitowych.
   3. Możliwość ręcznego ustalania harmonogramu i priorytetów wykonywanych zadań.
   4. Automatyczne przeszukiwanie wskazanych folderów lokalnych, sieciowych oraz kierowanie do przetwarzania pojawiających się w nich plików.
   5. Możliwość dostosowywania działania oprogramowania i integracji z zewnętrznymi aplikacjami poprzez:
      1. VBScript
      2. JScript
      3. API sieciowy
   6. Definiowanie parametrów przetwarzania poprzez pliki XML
   7. Automatyczne wstępne przetwarzanie rozpoznawanych obrazów:
      1. Wykrywanie orientacji strony
      2. Wykrywanie i korekcja rozdzielczości
      3. Korekcja szumów
      4. Prostowanie obrazu (deskew)
      5. Usuwanie zanieczyszczeń obrazu (despeckle)
   8. Wykrywanie wielu osobnych bloków tekstowych na stronie i ustalenie ich kolejności
   9. Możliwość klasyfikacji i indeksowania dokumentów automatycznie według skryptu
   10. Obsługiwane formaty plików wejściowych:
       1. TIFF, w tym wielostronicowe pliki TIFF
       2. JPEG, JPEG2000
       3. BMP
       4. PNG
       5. PDF
   11. Obsługiwane formaty plików wyjściowych:
       1. TXT – tekst w formacie Unicode UTF-8 i UTF-16
       2. RTF
       3. DOC, DOCX
       4. XLS, XLSX
       5. PDF, PDF/A
       6. ALTO XML
   12. Rozpoznawanie co najmniej następujących języków:
       1. polski
       2. angielski
       3. niemiecki
       4. francuski
       5. rosyjski
       6. łacina
       7. ukraiński
       8. włoski
       9. czeski
       10. litewski
       11. hiszpański
       12. białoruski
       13. słowacki
       14. niderlandzki
       15. portugalski
       16. węgierski
       17. grecki
       18. szwedzki
   13. Wsparcie słownikowe co najmniej dla następujących języków:
       1. polski
       2. angielski
       3. niemiecki
       4. francuski
       5. rosyjski
       6. łacina
       7. ukraiński
       8. włoski
       9. czeski
       10. litewski
       11. hiszpański
       12. słowacki
       13. niderlandzki
       14. portugalski
       15. węgierski
       16. grecki
       17. szwedzki
   14. Udostępniona integracja poprzez komunikację z wykorzystaniem folderów wejściowych, z następującymi założeniami:
       1. w folderach wejściowych będą umieszczane pliki do rozpoznania OCR wraz z plikiem sterującym w postaci xml (każde zadanie musi być możliwe do opisania oddzielnym plikiem sterującym XML)
       2. będzie możliwość definiowania wielu zadań OCR poprzez umieszczanie materiału do rozpoznania w oddzielnych podfolderach danego folderu wejściowego (każdy podfolder to oddzielne zadanie).
       3. musi być możliwość definiowania wielu folderów wejściowych a dla każdego z folderów wejściowych musi być możliwość przypisania (z poziomu GUI administracji aplikacją OCR) oddzielnego profilu z domyślnymi parametrami przetwarzania.
       4. Plik sterujący xml to plik, który musi mieć możliwość określenia następujących parametrów dotyczących zadania przetwarzania OCR:
          1. określenie listy plików przeznaczonych do przetwarzania,
          2. przekazanie parametrów użytkownika (własnych, niezdefiniowanych w oprogramowaniu),
          3. określenie parametrów przetwarzania nadpisujących domyślne parametry pochodzące z profilu właściwego dla danego folderu wejściowego. Musi być możliwość definiowania tylko wybranych parametrów zaś dla sekcji nie zdefiniowanych w pliku sterującym mają być uwzględniane parametry domyślne ustawione w profilu właściwym dla danego folderu wejściowego. Musi być możliwość określenia dowolnego parametru jaki jest możliwy do ustawienia w GUI zarządzania aplikacją, w tym między innymi: metoda separacji plików wynikowych, formaty plików wyjściowych, sposób i miejsce docelowe zapisu, parametry tworzenia plików wyjściowych, wybór słowników (języków) dla rozpoznawanego materiału.
   15. Udostępniona integracja poprzez komunikację z wykorzystaniem, odpowiadających jeden do jeden folderom wejściowym, folderów wyjściowych, w których umieszczane są pliki zawierające wyniki procesu rozpoznawania OCR wraz z plikami informacyjnymi w postaci xml. Jeżeli w folderze wejściowym zadanie znajdowało się w podfolderze to w katalogu wyjściowym pliki wynikowe tego zadania muszą być umieszczone w podfolderze o identycznej nazwie jak w katalogu wejściowym.
       1. Plik informacyjny xml to plik, który zawiera informacje o przetworzonym zadaniu OCR i musi zawierać co najmniej następujące informacje:
          1. Listę plików wyjściowych wraz ze ścieżkami dostępu do tych plików i parametrami ich tworzenia (zgodnie z ustawieniami w profilu lub ustawieniami określonymi w pliku sterującym xml) a także wzorem nadawania nazwy.
          2. Listę plików wejściowych, które podlegały procesowi rozpoznawania OCR, wraz z informacjami przy każdym pliku o: dacie modyfikacji pliku, informacji o ewentualnych uwagach (warningach), bądź błędach przetwarzania danego pliku, statystyki danego pliku (łączna liczba znaków, liczba znaków niepewnych/nierozpoznanych).
          3. Informacje o relacji, z jakiego zadania cząstkowego (per plik wejściowy) wychodzą jakie pliki wyjściowe, wraz parametrami tworzenia pliku wyjściowego, jego lokalizacji, nazwie i regule tworzenia nazwy.
          4. Informacje o całym zadaniu, w tym: parametry preprocesingu, parametry separacji plików wyjściowych, parametry formatów wyjściowych (jakie formaty wyjściowe i parametry ich tworzenia), użyte słowniki, statystyki dla całego zadania (łączna liczba znaków, łączna liczba znaków niepewnych/nierozpoznanych).
2. Parametry silnika OCR
   1. W załączniku nr 2 do Opisu Przedmiotu Zamówienia - Obiekty testowe Zamawiający umieścił testowe skany obiektów dla każdego umieszczając pliki zawierające obrazy rastrowe.
   2. Oprogramowanie zostanie uznane za równoważne, jeśli uzyska dla trybu nie nadzorowanego przecz operatora, a więc z jednakowymi ustawieniami ręcznymi dla wszystkich obiektów testowych w każdej kategorii wyniki nie gorsze niż przedstawione w poniższej tabeli. Poza ustawieniami ręcznymi silnik OCR nie może być sprofilowany do dostarczanych przez zamawiającego skanów testowych (musi mieć bazę w wersji dostępnej dla testów dla niezdefiniowanego odbiorcy).
   3. Przez wynik gorszy rozumiane jest uzyskanie wartości większej.

|  |  |  |  |
| --- | --- | --- | --- |
| **Obiekt testowy** | **format** | **Błędy CER**  (Character Error Rate) | **Błędy WER**  (Word Error Rate) |
| Chleb rzucony umarłym | TIFF | 0,38 | 1,13 |
| JPG | 0,39 | 1,16 |
| Chłopi 1 | JPG | 0,36 | 1,01 |
| Jaszczur | TIFF | 0,77 | 1,71 |
| JPG | 0,82 | 1,81 |
| Dzieje Tristana i Izoldy | JPG | 0,69 | 3,09 |
| Kwiaty grzechu | TIFF | 1,22 | 2,86 |
| JPG | 1,20 | 2,79 |
| Czerwony kapturek | TIFF | 23,15 | 33,18 |
| JPG | 26,42 | 39,81 |

* 1. Wyniki opisane w punkcie 2.2 mają być uzyskane przez porównanie wyników OCR z prawidłowymi tekstami ground truth za pomocą oprogramowania do weryfikacji Ocrevaluation Tool (lub równoważnego tj. dającego dla danego zestawu plików identyczne wyniki) opisanego pod adresem:

<https://sites.google.com/site/textdigitisation/ocrevaluation/>

i dostępnego publicznie pod adresem:

<https://github.com/impactcentre/ocrevalUAtion>