**OPIS PRZEDMIOTU ZAMÓWIENIA**

Zamówienie jest realizowane przez Instytut Badań Edukacyjnych w ramach projektu Wspieranie dalszego rozwoju Zintegrowanego Systemu Kwalifikacji w Polsce (ZSK 6) w ramach programu Fundusze Europejskie dla Rozwoju Społecznego 2021 - 2027 (FERS), współfinansowanego ze środków Unii Europejskiej.

Przedmiotem zamówienia jest usługa polegająca na opracowaniu **narzędzia informatycznego do rozpoznawania w nieprzetworzonych danych tekstowych dot. ofert pracy ściśle określonych jednostek nazwanych**. Narzędzie zostanie przygotowane przy użyciu języka programowania Python oraz nowoczesnych (state-of-the-art) metod przetwarzania języka naturalnego, bazujących na dużych modelach językowych (large language models, LLM), lub na:

• metodach opartych na regułach (rule-based),

• klasycznych modelach uczenia maszynowego (np. CRF, SVM),

• sieciach neuronowych wyspecjalizowanych pod NER (np. BiLSTM-CRF, BERT-CRF),

• hybrydowych podejściach łączących różne metody.

**Przewidywany czas realizacji całości zamówienia - 6 miesięcy od dnia podpisania umowy.**

**ZADANIA WYKONAWCY:**

1. **Opracowanie koncepcji narzędzia**

Zamawiane narzędzie będzie się składać z modelu, jego oprogramowania oraz programów współpracujących z modelem. Wykonawca zaproponuje skuteczny model do rozpoznawania jednostek nazwanych oraz uzasadni jego wybór w formie pisemnej. Uzasadnienie wyboru powinno zawierać porównanie modelu na tle innych produktów/rozwiązań alternatywnych/konkurencyjnych przy użyciu benchmarków dla zadań typu „Named-entity recognition”. Ewaluacja i porównanie modeli powinny uwzględniać:

* miary precyzji, czułości, F1-score oraz innych dedykowanych wskaźników jakości dla zadań typu NER dla każdej kategorii jednostek nazwanych,
* porównanie z istniejącymi rozwiązaniami typu open-source (np. spaCy, NLTK, Stanza).,
* testy na różnorodnych danych, w tym edge cases i trudnych przykładach.

Powyższe informacje zostaną także umieszczone w ostatecznej wersji raportu metodologicznego, po akceptacji Zamawiającego.

Przygotowanie API do łączenia przygotowanego narzędzia z istniejącymi ogólnodostępnymi modelami LLM open source takimi jak Chat GPT czy Claude AI nie jest wystarczające do zrealizowania zadania. Model musi być wytrenowany na danych dostarczonych przez Zamawiającego i być częścią przygotowanego przez Wykonawcę oprogramowania. Model nie może mieć ograniczeń pojemności informacji możliwych do przetworzenia. Model nie może przyporządkowywać do ofert pracy nieistniejących kodów zawodów (w przeprowadzonej symulacji wybrany model przyporządkował 27 nieistniejących kodów zawodów na 48 ofert pracy).

1. **Przygotowanie danych**

Zleceniodawca dostarczy Wykonawcy plik tekstowy zawierający nieprzetworzone i nieoznakowane dane tekstowe nt. internetowych ofert pracy ogłoszonych na różnych komercyjnych portalach rekrutacyjnych w okresie 2021-2024. Dane tekstowe zawierać będą całość treści internetowych ofert pracy (nieprzetworzone dane jednostkowe dot. poszczególnych ofert pracy). Dane przedstawione będą w takim języku i układzie, w jakim zostały one ogłoszone (również w językach innych niż polski). Narzędzie powinno obsługiwać język polski oraz angielski. Wykonawca we własnym zakresie przeprowadzi procedurę przetwarzania danych tekstowych, w tym oznakowanie danych i tłumaczenie (o ile jest to potrzebne) oraz przedstawi szczegółowy opis podjętych w trakcie przygotowania danych etapów/czynności. Wykonawca zaproponuje:

* metody czyszczenia i normalizacji tekstu,
* techniki augmentacji danych,
* strategie radzenia sobie z niezrównoważonymi klasami,
* strategie transferu uczenia pomiędzy językami,
* metody radzenia sobie z różnicami w strukturze nazw własnych między językami,
* metody optymalizacji modelu pod kątem szybkości inferencji,
* strategie przetwarzania dużych wolumenów tekstu,
* możliwości przetwarzania strumieniowego (online) vs. wsadowego (batch).

Powyższe informacje zostaną także umieszczone w ostatecznej wersji raportu metodologicznego, po akceptacji Zamawiającego.

W zależności od wybranej metody/modelu Wykonawca we własnym zakresie powinien odpowiednio przygotować dane (czyszczenie, przekształcenie oraz inne manipulacje). Jako dane wejściowe narzędzie musi przyjąć przygotowane dane tekstowe. Jako wynik narzędzie musi zwrócić rozpoznane w tekście jednostki nazwane w formie słownika, w którym „kluczem” będzie nazwa jednostki, zaś „wartością” bezpośrednio zidentyfikowane jednostki nazwane. W przypadku, jeśli takich jednostek będzie więcej niż jedna, np. kilka wymaganych umiejętności, muszą one zostać podzielone na osobne wyrazy i zwrócone w postaci wektora. Przykład wymagań z ogłoszenia zatrudnienia, treść którego przedstawiono w sekcji Przykłady – [„wykształcenia średniego technicznego o specjalności elektronika / elektrotechnika, automatyka”, „doświadczenie w projektowaniu szablonów, programowaniu automatów, drukarki pasty, umiejętność doboru profilu pieca,”, „mile widziane doświadczenie w obsłudze linii SMT”, „znajomość procesów z zakresu montażu powierzchniowego”, „znajomość komponentów elektronicznych oraz technologii SMT”]. – [„wymaganie nr 1”, „wymaganie nr 2”]).

1. **Opracowanie narzędzia**

Zleceniodawca nie wymaga od Wykonawcy interfejsu do narzędzia. Narzędzie powinno być możliwe do zastosowania z ogólnodostępnym językiem programowania Python.

Narzędzie powinno skutecznie rozpoznawać w nieprzetworzonym tekście ściśle określone przez Zleceniodawcę typy jednostek nazwanych, wśród których są następujące jednostki nazwane (jednostki nazwane wysortowane malejąco od najbardziej istotnych dla Zleceniodawcy):

* wymagania na stanowisko,
* obowiązki zawodowe,
* typ umowy,
* benefity (motywatory pozapłacowe),
* jednostka geograficzna – lokalizacja/miejscowość,
* wymiar czasu pracy,
* tryb pracy,
* nazwa pracodawcy wraz z identyfikacją numeru REGON (umożliwienie identyfikacji jednostki gospodarczej i jednostki lokalnej).

Przykładowe rozpoznanie i oznakowanie wybranych jednostek nazwanych w tekstach internetowych ofert pracy z portalu rekrutacyjnego https://www.pracuj.pl/ oraz serwisu ogłoszeniowego https://www.olx.pl/praca/ zostało przedstawione w dwóch przykładach w sekcji Przykłady.

Wykonawca może korzystać z różnych narzędzi czy programów zewnętrznych pod warunkiem, że są one ogólnodostępne i nie wymagają one dodatkowych kosztów (są typu open source). Wykorzystywanie narzędzi odpłatnych jest możliwe wyłącznie po uzgodnieniu z Zleceniodawcą.

Narzędzie musi mieć możliwość funkcjonowania, trenowania i rozpoznawania jednostek nazwanych przy użyciu zarówno procesora (CPU), jak i karty graficznej (GPU, NVIDIA GeForce RTX 30 series oraz 40 series), a także kilku połączonych kart graficznych (Wykonawca nie jest ograniczony co do wyboru metody/sposobu łączenia kart graficznych, pod warunkiem, iż nie będzie ona wymagała nabycia kosztownego sprzętu/oprogramowania).

Narzędzie powinno być wystarczająco uniwersalne i elastyczne by Zleceniodawca (bez posiadania zaawansowanej wiedzy informatycznej) mógł również we własnym zakresie nauczyć narzędzie do rozpoznania innych niż wskazane przez Zleceniodawcę typów jednostek nazwanych.

Ponadto Wykonawca uwzględni następujące obszary aktualizacji i walidacji modelu:

* strategie aktualizacji modelu w miarę pojawiania się nowych danych,
* monitoring wydajności modelu w czasie,
* procedury re-treningu i fine-tuningu,
* testy jednostkowe dla poszczególnych komponentów,
* testy integracyjne całego systemu,
* testy wydajnościowe i obciążeniowe.

Powyższe informacje zostaną umieszczone w ostatecznej wersji raportu metodologicznego, po akceptacji Zamawiającego.

**PRODUKTY ZAMÓWIENIA:**

**Narzędzie informatyczne posiadające następujące cechy:**

* wieloplatformowość (musi funkcjonować na różnych systemach operacyjnych, w tym na różnych wersjach czy generacjach systemów operacyjnych);
* odporność na zmiany wersji stosowanych w produkcie zewnętrznych bibliotek (frameworków);
* wielojęzyczność (produkt musi umieć operować treściami w języku polskim oraz angielskim);
* odpornośćna błędy gramatyczne, odmiany oraz inne specyficzne cechy językowe (itd. żargon, homonimy, skróty, takie jak: Sp. z o.o., S.A., UoP, um. zlec itd.).

**Wykonawca przekaże Zamawiającemu:**

1. Wstępny raport metodologiczny (w terminie 1 miesiąca od zawarcia umowy), zawierający:
   1. opis wybranej metody, modelu lub architektury z uzasadnieniem wyboru, w tym: przegląd rankingów dla zadań typu NER, wybór rankingu, na podstawie rankingu wybór najlepszego modelu, porównywanie modeli z tego rankingu, opis pracy tego modelu oraz uzasadnienie dlaczego ten model został wybrany.
2. Ostateczny raport metodologiczny (w terminie 6 miesięcy od zawarcia umowy), zawierający:
   1. opis wybranej metody, modelu lub architektury z uzasadnieniem wyboru, w tym: przegląd rankingów dla zadań typu NER, wybór rankingu, na podstawie rankingu wybór najlepszego modelu, porównywanie modeli z tego rankingu, opis pracy tego modelu oraz uzasadnienie dlaczego ten model został wybrany
   2. opis procedury przetwarzania danych przekazanych od Zleceniodawcy do poziomu niezbędnego do wykorzystania przez opracowane narzędzie (w tym opis procedury oznakowania danych, o ile to będzie potrzebne) w tym:
      1. metody czyszczenia i normalizacji tekstu,
      2. techniki augmentacji danych,
      3. strategie radzenia sobie z niezrównoważonymi klasami,
      4. strategie transferu uczenia pomiędzy językami,
      5. metody radzenia sobie z różnicami w strukturze nazw własnych między językami,
      6. metody optymalizacji modelu pod kątem szybkości inferencji,
      7. strategie przetwarzania dużych wolumenów tekstu,
      8. możliwości przetwarzania strumieniowego (online) vs. wsadowego (batch);
   3. opis obsługi narzędzia (instrukcja) uwzględniający:
      1. proces trenowania modelu do wyosobnienia nowych (innych niż wskazanych przez Zleceniodawcę) jednostek nazwanych oraz proces rozpoznania jednostek wskazanych przy użyciu przygotowanych danych;
      2. strategie aktualizacji modelu w miarę pojawiania się nowych danych,
      3. monitoring wydajności modelu w czasie,
      4. procedury re-treningu i fine-tuningu,
   4. opis wyników oraz skuteczności opracowanego narzędzia a także wyników testów zawartych w nim skryptów, w tym:
      1. testy jednostkowe dla poszczególnych komponentów,
      2. testy integracyjne całego systemu,
      3. testy wydajnościowe i obciążeniowe;
   5. pozostałe treści, powstałe w trakcie realizacji zamówienia.
3. Skrypty I-IV skompilowane w postaci biblioteki (frameworku) w języku programowania Python (w terminie 6 miesięcy od zawarcia umowy):
   1. skrypty informatyczne zawierające algorytm do przetwarzania danych surowych (analogicznych do tych, które zostaną przekazane przez Zleceniodawcę) do poziomu niezbędnego do wykorzystania przez opracowane narzędzie wraz z objaśnieniem poszczególnych jego części;
   2. skrypty informatyczne zawierające algorytm do trenowania wybranego przez Wykonawcę modelu na odpowiednio przygotowanych danych wraz z objaśnieniem poszczególnych jego części;
   3. skrypty informatyczne zawierające algorytm do rozpoznawania w nieprzetworzonych danych tekstowych ściśle określonych jednostek nazwanych przy użyciu odpowiednio przygotowanych danych i wytrenowanego modelu wraz z objaśnieniem poszczególnych jego części;
   4. skrypty informatyczne zawierające algorytm do eksportowania wyników modelu (punkt III) w formie tabelarycznej wraz z objaśnieniem poszczególnych jego części;
4. Materiały szkoleniowe: podręcznik użytkowania, przykłady użycia dla programistów i tutorial (lub tutoriale) wideo.

**Ponadto Wykonawca zapewni:**

1. Wsparcie po zakończeniu prac i przekazaniu produktu w zakresie naprawiania błędów które mogą powstać w trakcie eksploatacji narzędzia (w okresie 6 miesięcy po zakończeniu umowy).
2. Szkolenie dla Zleceniodawcy dotyczące zastosowania wszystkich funkcjonalności opracowanego narzędzia informatycznego. Szkolenie będzie przeprowadzone online w formule synchronicznej i będzie trwało minimum dwie godziny. Wykonawca przewidzi czas na dyskusję, podczas której odpowie na wszystkie pytania Zamawiającego.
3. Redakcję językową raportu merytorycznego oraz podręcznika użytkownika oraz zadba o przystępną stronę wizualną obu produktów.

**ZESPÓŁ WYZNACZONY DO REALIZACJI ZAMÓWIENIA**

Zespół wyznaczony do realizacji zamówienia powinien składać się co najmniej z:

* specjalisty ds. programowania, który posiada doświadczenie badawcze związane z wykorzystaniem technik kodowania oraz metod rozpoznawania języka naturalnego (natural language processing, NLP) oraz wykazuje się praktyczną znajomością języków programowania,
* specjalisty ds. programowania, który posiada doświadczenie badawcze związane z wykorzystaniem technik kodowania oraz metod uczenia maszynowego (machine learning, ML) oraz wykazuje się praktyczną znajomością języków programowania, jak również doświadczeniem w pracy z narzędziami i bibliotekami do uczenia maszynowego,
* specjalisty ds. baz danych, który posiada doświadczenie w przygotowaniu, architekturze i zarządzaniu bazami danych wykorzystywanymi do pracy badawczej/analitycznej.

**PRZYKŁADY**

Treść ogłoszenia o pracy nr 1:

<https://www.olx.pl/oferta/praca/monter-pomocnik-swiatlowody-CID4-IdkUwlE.html>

„Monter/Pomocnik [1] – Światłowody .Warszawa, Wawer [2]. Pełny etat [3], Umowa o pracę [4]. Odpowiednie doświadczenie zawodoweDyspozycyjność: Elastyczny czas pracy Pracownicy z Ukrainy: 🇺🇦 Запрошуємо людей з України (Zapraszamy pracowników z Ukrainy).Szukamy osoby, która:wie jak spawać światłowody i robić pomiary ,wie jak budować relacje światłowodowe ,wie jak budować instalacje na obiektach,będzie wyjeżdżać na krótkie delegacje (każdy weekend w domu),ma prawo jazdy kat. B, a najlepiej B+E [5].Jeśli czegoś nie umiesz, ale chcesz umieć, to pomożemy. Inwestujemy w ludzi z motywacją. Finansujemy szkolenia i uprawnienia dopasowane „do zawodnika” [7].Najważniejsze informacje o nas:Zawsze wypłacamy pensje na czas.Oferujemy prywatną opiekę zdrowotną [7].Szukamy ludzi na stałe.Kupujemy 100% sprzętu potrzebnego do pracy.Szybko podejmujemy decyzje – nie jesteśmy częścią żadnej dużej korporacji.Jako Net Com System [6] od 20 lat projektujemy, budujemy, zaciągamy, dmuchamy, spawamy i naprawiamy światłowody. Robiliśmy projekty dla każdej dużej firmy telekomunikacyjnej w tym kraju. Zatrudniamy ponad 70 osób, pracujemy w Polsce [2] i Niemczech [2].Prosimy o dodanie w CV klauzuli: Wyrażam zgodę na przetwarzanie danych osobowych zawartych w niniejszym dokumencie do realizacji procesu rekrutacji zgodnie z ustawą z dnia 10 maja 2018 roku o ochronie danych osobowych (Dz. Ustaw z 2018, poz. 1000) oraz zgodnie z Rozporządzeniem Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (RODO).”

[1] – Tytuł oferty pracy

[2] – Lokacja / Miejscowość

[3] – Wymiar pracy

[4] – Rodzaj umowy

[5] – Wymagania

[6] – Nazwa pracodawcy / Nazwa firmy

[7] – Benefity / Bonusy pozapłacowe

Treść ogłoszenia o pracy nr 2:

<https://www.pracuj.pl/praca/operator-linii-smt-zielona-gora-slubicka-4,oferta,1003398841>

„Operator linii SMT [1] Lumel S.A. [2] O firmieSłubicka 4, Zielona GóraZielona Góra, lubuskie [3] ważna jeszcze ponad miesiącdo 11 lipca 2024 umowa o pracę [4] pełny etatp [5] pracownik fizycznypraca stacjonarna [6] Praca od zarazLumel S.A. [1] Słubicka 4Zielona Góra [3]Sprawdź jak dojechaćTwój zakres obowiązkówwspieranie procesu produkcyjnego,monitorowanie ciągłości produkcji,programowanie i optymalizacja linii SMT,monitorowanie ilości braków z linii SMT w celu zapewnienia wysokiego poziomu produktywności,wdrażanie działań korygujących procesy produkcyjne [7]. Nasze wymaganiawykształcenia średniego technicznego o specjalności elektronika / elektrotechnika, automatyka,doświadczenie w projektowaniu szablonów, programowaniu automatów, drukarki pasty, umiejętność doboru profilu pieca,mile widziane doświadczenie w obsłudze linii SMT,znajomość procesów z zakresu montażu powierzchniowego,znajomość komponentów elektronicznych oraz technologii SMT [8].To oferujemystabilne zatrudnienie w oparciu o umowę o pracę [4],atrakcyjne wynagrodzenie wraz z transparentnym systemem premiowym,profesjonalne wdrożenie do pracy na obejmowanym stanowisku,szkolenia wspierające rozwój zawodowy,pakiet świadczeń socjalnych, m.in.: dopłaty do wakacji, świąt, karty multisportowe,atrakcyjną ofertę ubezpieczenia grupowego,prywatną opiekę medyczną,możliwość korzystania z kompleksu sportowo-rekreacyjnego LUMEL ARENA (bilard, tenis stołowy, siłownia oraz multisportowa hala),możliwość zdobycia tytułu \"Pracownika miesiąca\",nagrody za usprawnienia procesu pracy oraz za skuteczne polecenie kandydata do pracy,pożyczki na preferencyjnych warunkach,możliwość udziału w akcjach charytatywnych,duży, bezpłatny parking,dobrą atmosferę i pełne wsparcie przełożonego i Prezesa przez cały okres pracy w Lumel [9]. Benefitydofinansowanie zajęć sportowychprywatna opieka medycznadofinansowanie szkoleń i kursówubezpieczenie na życiebrak dress code’uparking dla pracownikówdofinansowanie wypoczynkupaczki świąteczne [9] Lumel S.A. [2] Przewiń do profilu firmyNależymy do czołowych europejskich producentów urządzeń automatyki przemysłowej oraz precyzyjnych odlewów ciśnieniowych. Działamy na rynku od 1953 r.LUMEL S.A. [2] specjalizuje się w produkcji urządzeń automatyki przemysłowej przeznaczonej do pomiarów, przetwarzania, regulacji, rejestracji oraz transmisji i wizualizacji procesów przemysłowych, usługach w zakresie projektowania i wykonawstwa systemów automatyki, usługach w zakresie EMS oraz doradztwie technicznym.Lumel Alucast Sp. z o.o. [2] jest znanym na świecie producentem wysoko- ciśnieniowych odlewów z aluminium. Z powodzeniem łączymy blisko 70-letnie doświadczenie z innowacyjnymi technologiami. Dostarczamy kompleksowe rozwiązania począwszy od projektowania i wykonania narzędzi i form, przez odlewanie, obróbkę, malowanie i montaż. Nasze odlewy dostarczamy do wielu gałęzi przemysłu, takich jak: motoryzacja, automatyka, ogrzewnictwo i chłodnictwo, przemysł oświetleniowy, elementy pomp i sprężarek.Aplikuj szybkoAplikuj szybkoZapiszDrukujUdostępnijOferta z szybkim aplikowaniemco to?Administratorem Danych Osobowych są: Lumel S.A. [2] z siedzibą w Zielonej Górze przy ul. Słubickiej 4 [3] oraz Lumel Alucast Sp. z o.o. [2] z siedzibą w Zielonej Górze\nprzy ul. Słubickiej 1 [3]. Dane będą przetwarzane w celu przeprowadzenia procesu rekrutacji. Osobie, której dane dotyczą, przysługuje prawo dostępu do treści swoich danych oraz do ich poprawiania.PracaZielona Góra [3] Praca fizycznaMonterzy / Serwisanci / Elektrycy.”

[1] – Tytuł oferty pracy

[2] – Nazwa pracodawcy / Nazwa firmy

[3] – Lokacja / Miejscowość

[4] – Rodzaj umowy

[5] – Wymiar pracy

[6] – Typ pracy

[7] – Zakres obowiązków

[8] – Wymagania

[9] – Benefity / Bonusy pozapłacowe.