Dane bez twarzy

Dokumenty mówią więcej, niż powinny. Twoim zadaniem jest sprawić, by zachowały sens, lecz utraciły wszystko, co zdradza człowieka - nie naruszając ich prawdy.

1. Wprowadzenie - opis organizacji, sytuacji i stanu aktualnego

Reprezentujemy konsorcjum pracujące nad modelem **PLLuM (Polish Large Language Model)** – modelem językowym dedykowanym dla języka polskiego. Naszą misją jest stworzenie suwerennej technologii AI, która będzie wspierać administrację, naukę i biznes w Polsce, uwzględniając specyfikę naszego języka i kultury.

Aby wytrenować model tej klasy, potrzebujemy ogromnych zbiorów danych tekstowych. Znaczną część tych zasobów stanowią zapisy konwersacji, fragmenty e-maili, posty z forów dyskusyjnych czy zgłoszenia urzędowe. Problemem, z którym się mierzymy, jest występowanie w tych tekstach **Danych Osobowych**. Zgodnie z RODO oraz standardami etycznymi, nie możemy trenować modelu na surowych danych zawierających imiona, nazwiska, numery PESEL czy adresy, ponieważ istnieje ryzyko, że model "zapamięta" te dane i ujawni je w przyszłości.

Obecnie stosowane rozwiązania oparte na zasobach słownikowych są niewystarczające, szczególnie w przypadku języka polskiego (fleksja, skomplikowana gramatyka) oraz konieczności rozróżnienia kontekstu (np. miasto będące miejscem akcji vs miasto zamieszkania). Potrzebujemy rozwiązania, które z chirurgiczną precyzją wyczyści dane, pozostawiając jednocześnie ich strukturę gramatyczną i sens, co jest kluczowe dla jakości treningu modelu.

2. Wyzwanie

Wyzwanie polega na stworzeniu algorytmu lub modelu uczenia maszynowego, który automatycznie wykryje w tekstach konwersacyjnych w języku polskim określone kategorie danych wrażliwych i zastąpi je odpowiednimi etykietami (tokenami zastępczymi, np. `{name}`, `{pesel}`).
Kluczową trudnością jest kontekstowość wypowiedzi. System musi radzić sobie z tekstami nieformalnymi, pojedynczymi zapytaniami oraz dłuższymi fragmentami dialogów.
Rozwiązanie musi poprawnie klasyfikować dane, odróżniając np. miasto wspomniane w kontekście opisu wycieczki `{city}` od miasta będącego częścią adresu zamieszkania `{address}`.
Dodatkowo punktowanym wyzwaniem jest moduł pozwalający na morfologicznie spójną generację danych syntetycznych do zanonimizowanych tekstów. Moduł ten zamienia etykiety na pasujące i odpowiednio odmienione tokeny z danej kategorii.

---

3. Oczekiwany rezultat

Oczekiwanym rezultatem jest **komponent programistyczny (biblioteka Python / skrypt)**, który przyjmuje na wejściu surowy tekst i zwraca jego zanonimizowaną wersję z podmienionymi encjami.

Użytkownikiem końcowym tego rozwiązania będą **Inżynierowie Danych i Badacze ML** z zespołu PLLuM. Dlatego rozwiązanie powinno być łatwe do integracji z potokami przetwarzania danych (data pipelines).
Ważne jest, aby narzędzie nie usuwało całych zdań, lecz podmieniało konkretne frazy na tokeny zastępcze:

- Przykład wejścia: *"Nazywam się Jan Kowalski, mój PESEL to 90010112345. Mieszkam w Warszawie przy ulicy Długiej 5."*
- Oczekiwane wyjście: *"Nazywam się {name} {surname}, mój PESEL to {pesel}. Mieszkam w {address}."*
- Przykładowe oczekiwane wyjście z modelu do generacji danych syntetycznych: "*Nazywam się Maria Nowak, mój PESEL to 12432486324. Mieszkam w Bielsku-Białej przy ulicy Szerokiej 5."*

---

4. Wymagania formalne

Projekt przesyłany do oceny powinien zawierać:

- **Repozytorium kodu** (np. GitHub/GitLab) z pełnym kodem źródłowym rozwiązania pozwalającym na uruchomienie go w środowisku organizacji.
- **Plik README** z instrukcją instalacji (np. `pip install -r requirements.txt`) oraz instrukcją uruchomienia na przykładowych danych.
- **Prezentację w formacie PDF** (maksymalnie 5 slajdów) opisującą zastosowane podejście (użyte modele, heurystyki).

---

5. Wymagania techniczne

Rozwiązanie powinno być przygotowane w języku **Python**.

Wymagane klasy anonimizacji: System powinien rozpoznawać i anonimizować następujące kategorie:

1. Dane identyfikacyjne osobowe
    - `{name}` – imiona.
    - `{surname}` – nazwiska.
    - `{age}` – wiek.
    - `{date-of-birth}` – data urodzenia.
    - `{date}` – inne daty wydarzen pozwalające identyfikować osobę (np. w rozmowie medycznej „przyjęto 23.09.2023 r.”)
    - `{sex}` – płeć (jeśli wyrażona explicite w formie danej wrażliwej, np. w formularzu/deklaracji).
    - `{religion}` – wyznanie.
    - `{political-view}` – poglądy polityczne.
    - `{ethnicity}` – pochodzenie etniczne/narodowe.
    - `{sexual-orientation}` – orientacja seksualna.
    - `{health}` – dane o stanie zdrowia
    - `{relative}` – relacje rodzinne, które ujawniają tożsamość danej osoby (np. „mój brat Jan”, „syn Kowalskiego”, „córka pana Nowaka”)

2. Dane kontaktowe i lokalizacyjne
    - `{city}` – miasto (kontekst: opis miejsca zdarzenia, lokalizacja ogólna, nieadresowa).
    - `{address}` – pełne dane adresowe (ulica, numer domu/lokalu, kod pocztowy oraz miasto w kontekście miejsca zamieszkania).
    - `{email}` – adresy e-mail.
    - `{phone}` – numery telefonów.
3. Identyfikatory dokumentów i tożsamości:
    - `{pesel}` – numery PESEL.
    - `{document-number}` – numery dokumentów (np. dowodów osobistych, paszportów, legitymacji, prawa jazdy, itp).
4. Dane zawodowe i edukacyjne:
    - `{company}` – nazwa pracodawcy.
    - `{school-name}` – nazwa szkoły powiązana z osobą (jeśli unikalna).
    - `{job-title}` – stanowisko lub pełniona funkcja
5. Informacje finansowe
    - `{bank-account}` – numer rachunku bankowego, dane konta bankowego
    - `{credit-card-number}` – numery kart płatniczych.
6. Identyfikatory cyfrowe i loginy
    - `{username}` – nazwy użytkowników, loginy oraz identyfikatory w mediach społecznościowych
    - `{secret}` – sekrety takie jak hasła użytkowników czy klucze API.

Moduł wspierający generację danych syntetycznych powinien pozwalać na podmianę etykiety na pasujące kategorią wartości, po odpowiednim zapewnieniu dopasowania morfologicznego do reszty tekstu.

Ograniczenia i preferencje:

- **Rozwiązanie offline:** Ze względów bezpieczeństwa narzędzie NIE MOŻE korzystać z zewnętrznych API (np. OpenAI, Google Cloud NLP). Całe przetwarzanie musi odbywać się lokalnie.
- **Dozwolone technologie:** Można korzystać z modeli językowych dostępnych open-source (np. HerBERT, PolBERT, PLLuM) oraz bibliotek NLP (HuggingFace Transformers, Flair, SpaCy), pod warunkiem, że ich licencja pozwala na komercyjne użycie.
- **Wydajność:** Rozwiązanie powinno być skalowalne – docelowo będzie przetwarzać terabajty danych, więc wydajność inferencji jest istotna.

---

6. Sposób testowania i/lub walidacji

Zespoły otrzymają zbiór treningowy (przykładowe zdania z oznaczonymi encjami). Ocena końcowa odbędzie się na **ukrytym zbiorze testowym** przygotowanym przez organizatorów. Rozwiązanie zostanie uruchomione na ukrytym zbiorze, a wyniki zostaną porównane ze wzorcem ("złotym standardem"). Główną metryką oceny będzie **średni F1-score** dla

wszystkich klas. Dodatkowo nacisk zostanie położony na minimalizację **False Negatives** (FN, czyli sytuacji, gdzie dane wrażliwe nie zostały wykryte – co jest błędem krytycznym).

---

7. Dostępne zasoby

Uczestnikom zostanie udostępniony na Discordzie **zbiór danych (dataset)**. Będzie to paczka syntetycznych tekstów konwersacyjnych w języku polskim wraz z ich zanonimizowaną wersją, odzwierciedlająca charakter danych (format: JSONL lub CSV).

---

8. Kryteria oceny

Przykładowe kryteria:

- **Skuteczność anonimizacji (F1-score)** — 35% (Najważniejsze kryterium: bezpieczeństwo danych).
- **Wydajność, jakość kodu oraz łatwość wdrożenia** — 20% (Czas przetwarzania próbki danych, czytelność, dokumentacja, konteneryzacja).
- **Skuteczny moduł do generacji danych syntetycznych do zanonimizowanych tekstów** – 20% (dodatkowy moduł umożliwiający podmianę zanonimizowanego obiektu danej kategorii na inny obiekt z tej samej kategorii zachowując morfologię: Warszawskiej -> {city} -> Krakowskiej)
- **Poprawność rozróżniania kontekstu** — 15% (Szczególnie rozróżnienie `{city}` vs `{address}` oraz poprawna detekcja `{name}`/`{surname}` w odmianie fleksyjnej).
- **Pomysłowość podejścia** — 10% (Np. hybrydowe łączenie RegEx z modelami ML).

---

9. Dodatkowe uwagi / kontekst wdrożeniowy

Zwycięskie rozwiązanie ma szansę stać się oficjalnym elementem pipeline'u przetwarzania danych dla modelu PLLuM. Jest to unikalna okazja, aby mieć realny wkład w rozwój najważniejszego polskiego projektu AI. Autorzy najlepszych rozwiązań zostaną wymienieni w dokumentacji technicznej modelu jako kontrybutorzy.

---

10. Kontakt

W sprawach merytorycznych dotyczących definicji klas i struktury danych prosimy o kontakt na kanale Discord: `#dane-bez-twarzy` lub zgłaszanie się do mentorów przy stanowisku oznaczonym logo NASK. Kontakt mailowy do mentorek zadania: aleksandra.krasnodebska@nask.pl, katarzyna.dziewulska@nask.pl

---