

Dane Bez Twarz

Automatyczna Anonimizacja Danych Osobowych

dr inż. Sławomir Słowik

dr inż. Paweł Łosiński

Python • RODO/GDPR • Open Source

Biblioteka + CLI + Standalone EXE

Problem i Rozwiązanie

- **Problem**

- Dane osobowe w dokumentach wymagają ochrony zgodnie z RODO/GDPR
- Ręczna anonimizacja setek plików jest czasochłonna i błędogenna
- Brak uniwersalnego narzędzia dla różnych formatów (DOCX, PDF, XLSX)

- **Rozwiązanie**

- Automatyczne wykrywanie 25+ typów danych osobowych
- Wsparcie dla 5+ formatów plików (TXT, DOCX, PDF, XLSX, CSV)
- 3 interfejsy: Biblioteka Python, CLI, Standalone EXE

Architektura i Technologie

- **4 Detektory Działające Równolegle**
 - PlaceholderDetector: Wykrywa [name], [email], [pesel] - pewność 1.0
 - RegexDetector: PESEL, NIP, email, telefon - pewność 0.8-0.95
 - PolishDetector: Polskie wzorce: adresy, kody pocztowe - pewność 0.7-0.9
 - NLPDetector: spaCy NER dla imion/nazwisk - pewność 0.6-0.95
 - LLMDetector: PLLUM (12B parametrów) - najwyższa dokładność
- **Stack Technologiczny**
 - Python 3.12 • spaCy 3.7 • langchain-openai
 - PyPDF2, python-docx, openpyxl (procesory formatów)
 - matplotlib, plotly (wizualizacja raportów)
 - Nuitka (kompilacja do standalone EXE)

Kluczowe Funkcjonalności

- **6 Metod Anonimizacji**
 - Maskowanie (**), Pseudonimizacja (Osoba_A)
 - Entity ([name] [surname]), Haszowanie, Szyfrowanie, Redakcja
- **Przetwarzanie Wsadowe**
 - Pojedyncze pliki lub całe katalogi (rekurencyjnie)
 - Wzorce plików: *.txt, *.docx, *.xlsx, *.pdf
- **Raportowanie**
 - 3 formaty: JSON (dane), HTML (interaktywne wykresy), PDF (druk)
 - Statystyki: liczba encji, typy, pewność, czas wykonania
 - Automatyczna rotacja logów (5 × 10 MB)
- **Chunking dla LLM**
 - Automatyczny podział dużych plików (3000 znaków/fragment)
 - Nakładanie fragmentów (200 znaków) - eliminacja utraty kontekstu

Demo i Wyniki

```
dane-bez-twarz anonymize cv.docx -o cv_anon.docx \
--method entity --use-nlp --use-llm \
--llm-api-key "klucz" --add-report report --report-format all -v
```

- ## Osiągnięte Wyniki

- ✓ Wykryto 25+ typów danych osobowych (PESEL, NIP, email, telefon, imiona)
- ✓ Chunking LLM obsługuje pliki >100 KB bez błędów 413/400
- ✓ Raport HTML z wykresami Plotly (interaktywny) i PDF (printable)
- ✓ Standalone EXE (~50 MB LITE, ~600 MB FULL z NLP)
- ✓ Format ENTITY zgodny z nask_train (orig.txt)