Ćwiczenie laboratoryjne

Analiza tekstów medycznych: rozpoznawanie jednostek medycznych i klasyfikacja dokumentów za pomoca NLP

Katedra Informatyki i Automatyki

1 Cel ćwiczenia

Celem ćwiczenia jest zapoznanie studentów z podstawowymi technikami przetwarzania jezyka naturalnego (NLP) stosowanymi w analizie tekstów medycznych. Po ukończeniu zajeć student potrafi:

- wczytywać i wstepnie przetwarzać dane tekstowe (czyszczenie, tokenizacja, lematyzacja),
- rozpoznawać jednostki medyczne w tekście (ang. *Named Entity Recognition, NER*),
- klasyfikować dokumenty medyczne według kategorii (np. diagnozy, typy badań),
- oceniać skuteczność modeli NLP przy użyciu odpowiednich metryk.

2 Wprowadzenie teoretyczne

Przetwarzanie jezyka naturalnego (NLP)

NLP (*Natural Language Processing*) jest dziedzina sztucznej inteligencji zajmujaca sie automatycznym przetwarzaniem i analiza jezyka ludzkiego. W kontekście medycyny, NLP pozwala m.in. na:

- ekstrakcje informacji z opisów badań i raportów lekarskich,
- rozpoznawanie nazw chorób, leków, procedur i jednostek medycznych,
- automatyczna klasyfikacje dokumentacji medycznej,
- wspomaganie systemów diagnozujacych i wyszukiwania informacji klinicznych.

Rozpoznawanie jednostek nazwanych (NER)

Zadanie NER polega na automatycznym wykrywaniu i etykietowaniu fragmentów tekstu, które odnosza sie do konkretnych kategorii pojeciowych — np. **choroba**, **lek**, **procedura medyczna**. W kontekście medycyny stosuje sie czesto specjalizowane modele jezykowe, np. *scispaCy*, *BioBERT* czy *MedGPT*.

Klasyfikacja dokumentów tekstowych

Klasyfikacja polega na przypisaniu całemu dokumentowi etykiety tematycznej (np. „karta informacyjna", „opis RTG", „wynik laboratoryjny"). Stosuje sie tu modele takie jak:

- klasyczne: Naive Bayes, SVM, Logistic Regression z wektorami TF–IDF,
- głebokie: modele transformerowe (np. *BERT*, *BioBERT*, *ClinicalBERT*).

3 Stosowane technologie i biblioteki

Python – jezyk analizy danych i NLP.

SpaCy – biblioteka do tokenizacji, lematyzacji, rozpoznawania encji.

scispaCy – modele NLP trenowane na tekstach biomedycznych.

Scikit-learn – klasyfikacja tekstów (TF–IDF, modele liniowe, SVM).

Transformers (Hugging Face) – modele BERT/BioBERT do klasyfikacji dokumentów.

Pandas, NumPy – zarzadzanie danymi i obliczenia pomocnicze.

Matplotlib, Seaborn – wizualizacja wyników klasyfikacji.

4 Przebieg ćwiczenia

Etap 1: Przygotowanie danych tekstowych

- Wczytaj zbiór tekstów medycznych (np. opisy badań, streszczenia artykułów, notatki kliniczne).
- Wyczyść dane: usuń znaki specjalne, liczby, nadmiarowe białe znaki.
- Przeprowadź tokenizacje i lematyzacje.

Etap 2: Rozpoznawanie jednostek medycznych (NER)

- Zastosuj model *scispaCy* (np. `en_core_sci_sm`) do wykrycia encji medycznych.
- Wyodrebnij z tekstu nazwy chorób, leków i procedur.
- Zlicz czestość wystepowania najpopularniejszych jednostek.

Etap 3: Klasyfikacja dokumentów

- Utwórz wektory cech tekstu przy użyciu *TF–IDF*.
- Zastosuj klasyfikator (np. `LogisticRegression` lub `LinearSVC`) do rozpoznania typu dokumentu.
- Porównaj skuteczność modeli przy użyciu metryk: *accuracy*, *precision*, *recall*, *F1-score*.

Etap 4: Wizualizacja i interpretacja wyników

- Wyświetl fragmenty tekstu z oznaczonymi jednostkami medycznymi.
- Narysuj wykresy: czestość wystepowania encji, macierz pomyłek klasyfikacji.
- Omów błedy klasyfikacji i ich możliwe przyczyny.

5 Warianty zadań dla studentów

Zadania na osobnych zestawach danych obejmuja cztery etapy:

1. Przygotowanie danych tekstowych,
2. Rozpoznawanie jednostek medycznych (NER),
3. Klasyfikacja dokumentów,
4. Wizualizacja i interpretacja wyników.

6 Literatura i źródła

- Explosion AI, *SpaCy Documentation*: `https://spacy.io/`
- Allen Institute for AI, *scispaCy: Biomedical Models for SpaCy*: `https://allenai.github.io/scispacy/`
- Hugging Face, *Transformers Library*: `https://huggingface.co/docs/transformers`