

Raport - Przetwarzanie tekstu w języku Python

1. Wprowadzenie

W ramach laboratorium przeanalizowano dwie biblioteki służące do przetwarzania tekstu w języku Python: TextBlob oraz spaCy. Obie biblioteki umożliwiają wykonywanie zadań związanych z przetwarzaniem języka naturalnego (NLP), takich jak tokenizacja, analiza sentymentu czy rozpoznawanie części mowy.

2. Wybrane biblioteki

2.1 TextBlob

TextBlob to prosta w użyciu biblioteka Python, która umożliwia przeprowadzanie podstawowej analizy tekstu. Pozwala na tokenizację, analizę sentymentu, rozpoznawanie części mowy oraz korektę gramatyczną. Jest idealna dla początkujących użytkowników.

Link do dokumentacji: <https://textblob.readthedocs.io/en/dev/>

2.2 spaCy

spaCy to nowoczesna i szybka biblioteka NLP, używana w projektach produkcyjnych. Umożliwia m.in. tokenizację, analizę składniową, rozpoznawanie nazw własnych (NER), lematyzację oraz analizę zależności gramatycznych. Jest bardziej zaawansowana i dokładna niż TextBlob, ale wymaga większej wiedzy i konfiguracji.

Link do dokumentacji: <https://spacy.io/usage>

3. Porównanie bibliotek

Cecha	TextBlob	spaCy
Prostota użycia	Bardzo prosta	Wymaga konfiguracji
Dokładność	Niska	Wysoka
Szybkość działania	Wolniejsza	Bardzo szybka
Analiza sentymentu	Tak	Brak domyślnie

NER	Nie	Tak
(rozpoznawanie nazw własnych)		

4. Wnioski

TextBlob to doskonałe narzędzie dla osób rozpoczynających pracę z NLP, oferujące prostotę i podstawowe funkcje analizy tekstu. Natomiast spaCy jest bardziej odpowiedni do zaawansowanych zastosowań, gdzie liczy się wydajność i dokładność. W zależności od potrzeb projektu, można wybrać odpowiednie narzędzie lub nawet używać obu równolegle.

4.4 Raport podsumowujący

TextBlob

Przeznaczenie: Proste przetwarzanie języka naturalnego w Pythonie.

Główne funkcje:

- - Tokenizacja
- - Analiza sentymentu
- - Rozpoznawanie części mowy (POS tagging)
- - Korekta gramatyczna

Zalety:

- - Bardzo prosta składnia i szybki start
- - Integracja z NLTK
- - Duża liczba tutoriali i przykładów

Ograniczenia:

- - Mniej precyzyjna analiza niż spaCy
- - Wolniejsze działanie na dużych zbiorach danych

Linki:

- Dokumentacja: <https://textblob.readthedocs.io/en/dev/>
- Repozytorium: <https://github.com/sloria/TextBlob>

spaCy

Przeznaczenie: Wydajne i zaawansowane przetwarzanie języka naturalnego, stosowane komercyjnie.

Główne funkcje:

- - Tokenizacja
- - Rozpoznawanie nazw własnych (NER)
- - Analiza składniowa (dependency parsing)
- - Lematyzacja
- - POS tagging

Zalety:

- - Bardzo szybka i zoptymalizowana
- - Dokładna analiza składniowa
- - Regularnie aktualizowana i rozwijana

Ograniczenia:

- - Brak wbudowanej analizy sentymentu
- - Większy rozmiar i więcej zależności

Linki:

- Dokumentacja: <https://spacy.io/usage>
- Repozytorium: <https://github.com/explosion/spaCy>

Link do github:

[python-intro/Zadanie 5 at main · rarity96/python-intro](https://github.com/rarity96/python-intro/blob/main/Zadanie%205.py)