W09. Przetwarzanie języka naturalnego (NLP) w psychologii

Psychoinformatyka | Wykład 09

2026-11-04

Cele i pytania badawcze

- Zdefiniować tekst jako dane obserwacyjne o określonym pochodzeniu i kontekście.
- Zastosować tokenizację, Bag-of-Words i TF-IDF do krótkich wypowiedzi.
- Rozpoznać ograniczenia interpretacji psychologicznej i walidacji modeli NLP.

Język jako wskaźnik procesów psychicznych

- Wypowiedzi odzwierciedlają wybór słów, organizację narracji i kontekst społeczny.
- Cechy językowe mogą korelować z poznaniem lub afektem, lecz nie są ich bezpośrednim pomiarem.
- Interpretacja zależy od tematu, języka, sytuacji i odbiorcy.

Zakres NLP w psychologii

- NLP obejmuje wstępne przetwarzanie, reprezentację, klasyfikację i generowanie tekstu.
- Jednostką analizy może być token, wypowiedź, osoba albo rozmowa.
- Metoda powinna odpowiadać konstrukcji pytania badawczego.

Źródła danych językowych

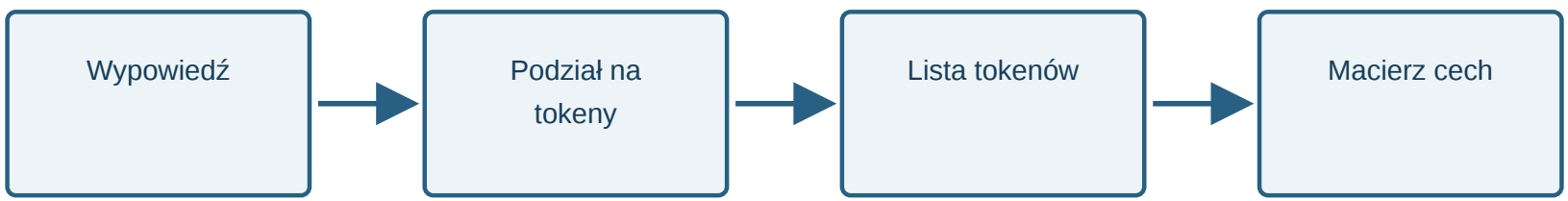
- Dzienniczki EMA, wywiady, ankiety otwarte i transkrypcje rejestrują różne aspekty języka.
- Platformy społecznościowe wprowadzają silną selekcję osób i treści.
- Zgoda i ograniczenie celu przetwarzania są częścią protokołu badania.

Budowa korpusu

- Rejestruj identyfikator uczestnika, czas, instrukcję i źródło wypowiedzi.
- Zachowaj wersję anotacji, schemat etykiet i rozdzielenie zbiorów danych.
- Kontroluj duplikaty i rozbieżności między anotatorami.

Tokenizacja: jednostki tekstu

- Tokenizator wydziela słowa, liczby, znaki i czasem interpunkcję.
- W języku polskim forma wyrazowa nie jest równoznaczna z leksemem.
- Błędy tokenizacji zmieniają słownik modelu i liczbę cech.



Rysunek 1: Tokenizacja: jednostki tekstu: wektorowy schemat metody

Normalizacja i jej koszty

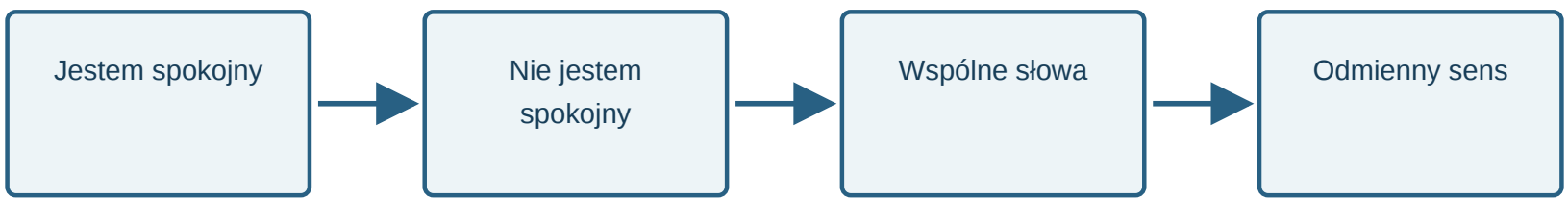
- Przekształcenie na małe litery zmniejsza liczbę wariantów leksykalnych.
- Usuwanie interpunkcji, cyfr i słów funkcyjnych może usuwać sygnał psychologiczny.
- Przed wdrożeniem dokumentuj dokładną kolejność operacji.

Stemming a lematyzacja

- Stemming przycina formy wyrazów algorytmicznie, często bez prawidłowego hasła słownikowego.
- Lematyzacja przybliża podstawową formę z uwzględnieniem cech morfosyntaktycznych.
- Wybór metody może zmieniać wyniki w języku fleksyjnym.

Język polski: fleksja i negacja

- Szyk zdania jest stosunkowo swobodny, a odmiana bogata.
- Negacja zmienia znaczenie całej frazy, chociaż wiele wyrazów pozostaje wspólnych.
- Korpusu polskiego nie należy bez walidacji traktować jak korpusu angielskiego.



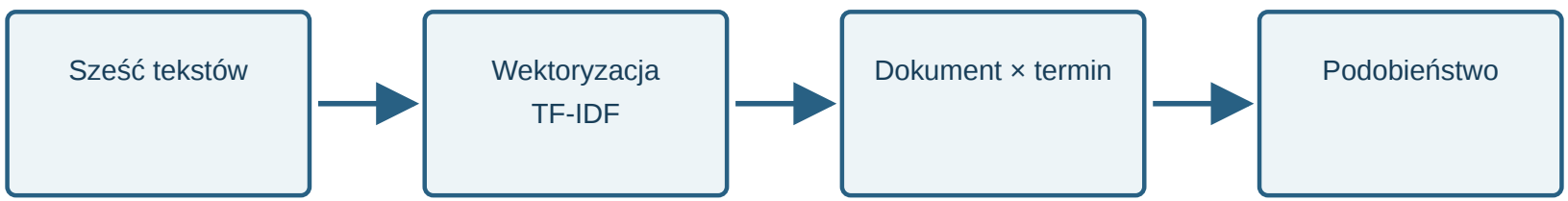
Rysunek 2: Python: wpływ bigramów: wektorowy schemat metody

Bag-of-Words

- Reprezentacja BOW zlicza wystąpienia tokenów w każdym dokumencie.
- Macierz dokument × termin jest rzadka i ignoruje większość kolejności słów.
- Zliczanie nie rozróżnia ironii, negacji i znaczenia w kontekście.

TF-IDF i znaczenie częstości

- TF mierzy udział terminu w dokumencie; IDF obniża wagę terminów powszechnych w korpusie.
- Wektor TF-IDF opisuje względną charakterystyczność słów, nie ich kliniczne znaczenie.
- Wektoryzator ucz tylko na zbiorze treningowym.



Rysunek 3: TF-IDF i znaczenie częstości: wektorowy schemat metody

LIWC i analiza słownikowa

- Słowniki grupują wyrażenia w kategorie językowe i psychologiczne.
- Częstość kategorii jest wskaźnikiem operacyjnym zależnym od walidacji słownika.
- Tłumaczenie kategorii na język polski wymaga ponownej oceny trafności.

Wskaźniki stylu językowego

- Zaimki, długość zdań i różnorodność leksykalna mogą opisywać styl wypowiedzi.
- Wartości zależą od długości tekstu i zadanej formy odpowiedzi.
- Raportuj normalizację wskaźników oraz jednostkę analizy.

Analiza sentymentu

- Klasyfikator polaryzacji ocenia ton wypowiedzi w ramach przyjętej definicji etykiet.
- Sentyment tekstu nie jest tożsamy z afektem autora.
- Sarkazm, cytat i negacja są szczególnie trudne.

Studium przypadku: dzienniczki stresu

- Porównaj sześć syntetycznych wypowiedzi o nauce, śnie i egzaminie.
- Najpierw analizuj podobieństwa leksykalne, a dopiero potem rozważ hipotezy psychologiczne.
- Wielokrotne wpisy tej samej osoby wymagają walidacji grupowej.

5 minut przerwy

Pytania do przemyślenia:

1. Jak tokenizacja różni się od lematyzacji?
2. Dlaczego słowo „nie„ zmienia interpretację tekstu?
3. Co oznacza wysoka waga TF-IDF?
4. Dlaczego nie należy utożsamiać sentymentu z emocją autora?

Składnia, semantyka, pragmatyka

- Składnia opisuje strukturę wypowiedzi, semantyka znaczenie, a pragmatyka użycie w sytuacji.
- Zależności między słowami mogą zmieniać znaczenie zdania.
- Nie istnieje pojedynczy wskaźnik tekstowy, który bez kontekstu mierzy emocję.

N-gramy i lokalny kontekst

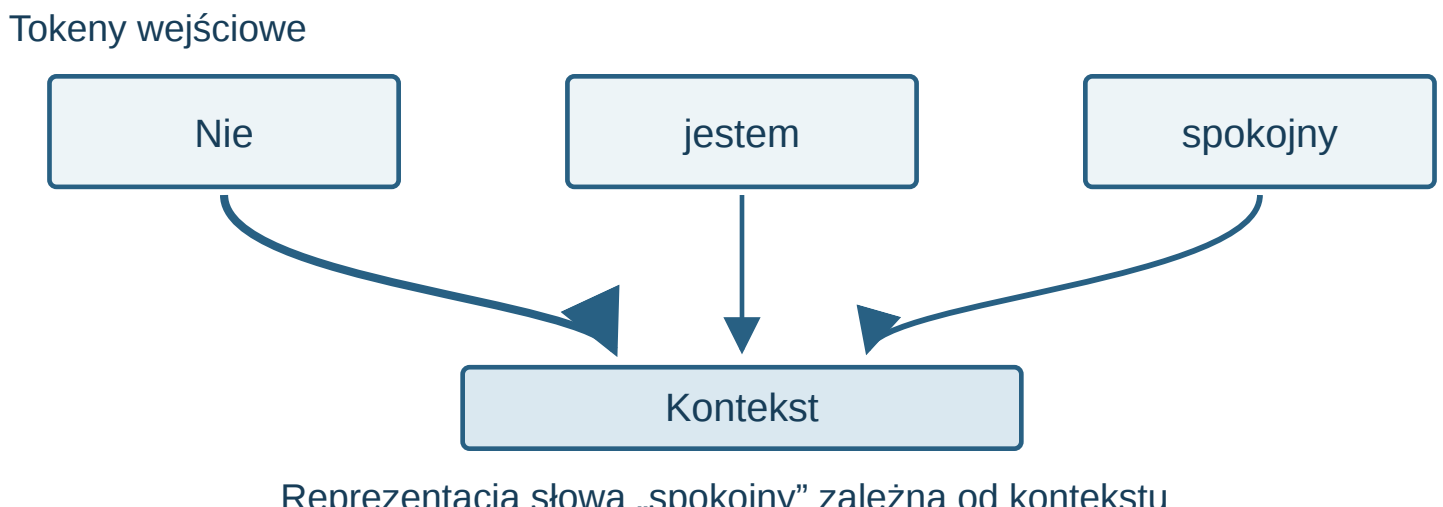
- Unigram obejmuje jedno słowo, bigram dwa kolejne, trigram trzy.
- Fraza „nie jestem„ zachowuje więcej informacji niż oddzielne tokeny.
- Większe n-gramy zwiększają rzadkość i wymiarowość cech.

Osadzenia semantyczne

- Embedding odwzorowuje token lub tekst na wektor w przestrzeni numerycznej.
- Podobieństwo wektorów może odzwierciedlać współwystępowanie i użycie słów.
- Bliskość embeddingów nie oznacza równoważności stanów psychicznych.

Transformer i mechanizm uwagi

- Mechanizm self-attention łączy informację z różnych pozycji sekwencji.
- Reprezentacje kontekstowe zmieniają się wraz z otaczającymi słowami.
- Mechanizm uwagi oblicza wagi relacji między pozycjami; nie jest automatycznym wyjaśnieniem decyzji.
- Duże modele nadal mogą tworzyć nieprawdziwe interpretacje i uogólnienia.



Rysunek 4: Tokeny i reprezentacja kontekstowa w uproszczonym mechanizmie self-attention

LLM jako narzędzia badawcze

- Modele generatywne wspierają kodowanie jakościowe i robocze streszczenia materiału.
- Wyniki zależą od promptu, wersji modelu i próby tekstów.
- Wymagaj recenzji ludzkiej, walidacji trafności oraz ochrony danych.

Klasyfikacja tekstu

- Definicja etykiety powinna wynikać z protokołu anotacji, a nie z intuicji modelu.
- Porównuj klasyfikator z prostym modelem bazowym.
- Raportuj czułość, precyzję i błędy dla poszczególnych klas.

Walidacja i przeciek danych

- Rozdziel osoby i rozmowy przed dopasowaniem TF-IDF.
- Fragmenty tej samej wypowiedzi i duplikaty mogą sztucznie zawyżyć wyniki.
- Sprawdzaj zgodność wyników między językami i populacjami.

Python: teksty i biblioteki

- Syntetyczny korpus zawiera sześć polskich zdań.
- W przykładzie wykorzystano scikit-learn i NumPy.
- Identyfikatory dokumentów służą do podpisywania macierzy.

```
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
teksty = [
    "Jestem spokojny dzisiaj", "Nie jestem spokojny dzisiaj",
    "Czuję napięcie przed egzaminem", "Po egzaminie odczuwam ulgę",
    "Martwię się wynikami", "Dzisiaj spałem dobrze"
]
print("Liczba tekstów:", len(teksty))
```

Python: reprezentacja TF-IDF

- Dopasuj `TfidfVectorizer` do jawnie zdefiniowanego korpusu.
- Przedstaw kształt macierzy dokument × termin.
- Odczytaj kilka elementów słownika, pamiętając o formach fleksyjnych.

```
wektoryzator = TfidfVectorizer(token_pattern=r"(?u)\b\w+\b")
X = wektoryzator.fit_transform(teksty)
print("Macierz TF-IDF:", X.shape)
```

Python: podobieństwo cosinusowe

- Iloczyn skalarny znormalizowanych wektorów TF-IDF odpowiada podobieństwu cosinusowemu.
- Wartość bliska 1 oznacza podobne proporcje cech, a nie ten sam sens psychologiczny.
- Macierz jest symetryczna; przekątna odpowiada tekstom porównanym ze sobą.

```
podobienstwo = cosine_similarity(X)
print("Podobieństwo zdań 1 i 2:", round(podobienstwo[0, 1], 3))
```

Python: wpływ bigramów

- Porównaj macierze podobieństwa dla unigramów i unigramów z bigramami.
- Bigramy częściowo zachowują frazę „nie jestem„, ale nie rozwiązują semantyki negacji.
- Sam bigram nie rozwiązuje problemów składni i pragmatyki.

```
X2 = TfidfVectorizer(ngram_range=(1, 2)).fit_transform(teksty)
podobienstwo2 = cosine_similarity(X2)
print("Po dodaniu bigramów:", round(podobienstwo2[0, 1], 3))
import matplotlib.pyplot as plt
fig, ax = plt.subplots(figsize=(6, 4))
im = ax.imshow(podobienstwo2, vmin=0, vmax=1)
ax.set(xlabel="Wypowiedź", ylabel="Wypowiedź", title="TF-IDF: unigramy i bigramy")
fig.colorbar(im, ax=ax, label="Podobieństwo cosinusowe")
plt.tight_layout()
plt.show()
```

Zastosowania naukowe NLP

- Analiza narracji, tematów i wzorców językowych może wspierać badania podłużne.
- Predykcja wrażliwych cech wymaga szczególnych przesłanek i walidacji.
- Zawsze podawaj populację, kontekst i rodzaj etykiety referencyjnej.

Etyka i ochrona wypowiedzi

- Tekst może ujawniać tożsamość mimo usunięcia imienia i nazwiska.
- Ogranicz dostęp, retencję i wtórne przetwarzanie transkrypcji.
- Raportuj niepewność i nie używaj niezwalidowanych wyników do decyzji o osobach.

Literatura i dokumentacja

- Tausczik, Y. R., Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology, 29(1), 24–54. DOI.
- Chancellor, S., De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. npj Digital Medicine, 3, artykuł 43. DOI.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT, 4171–4186. DOI.
- Vaswani, A., i in. (2017). Attention Is All You Need. NeurIPS 30. Artykuł.

Dokumentacja narzędzi: scikit-learn: ekstrakcja cech tekstowych.

Wnioski i samodzielna praca

- BOW i TF-IDF opisują występowanie cech językowych, nie rozpoznają bezpośrednio emocji.
- Kontekst, anotacja, rozdzielenie uczestników i ochrona prywatności warunkują rzetelną analizę.
- Uruchom `pi_w09.ipynb` i porównaj podobieństwo dwóch zdań po dodaniu negacji.