mgr inż., dr n.h. Inez Okulska

Ile treści jest w semantyce, czyli jak bardzo można przekształcać wektory słów, by nie stracić na jakości uczenia?

Seminarium „Przetwarzanie Języka Naturalnego” IPI PAN
2 listopada 2020, godz. 10.15
Online na Zoom

System reagujący na zagrożenia bezpieczeństwa dzieci w cyberprzestrzeni ze szczególnym uwzględnieniem pornografii dziecięcej

#15lat NASK dyżurnet.pl

Finansowanie:
Narodowe Centrum Badań i Rozwoju

Skład konsorcjum:
NASK Państwowy Instytut Badawczy
Politechnika Warszawska
Enamor International Sp. Z. o. o.

Czas trwania:
36 miesięcy

Koordynator naukowy:
Prof. dr hab. inż. Andrzej Pacut

W CZYM PROBLEM?

Konieczność klasyfikacji tekstu, do którego badacze opracowujący model nie mogą mieć dostępu (materiały pornograficzne) wymaga zastosowania takiego **maskowania**, które:

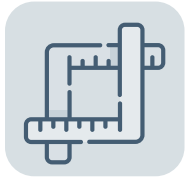
- **zagwarantuje niemożność odczytania treści tekstu** (rekonstrukcji znanych wektorów zanurzeń słów)

ALE JEDNOCZEŚNIE

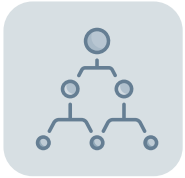
- **zachowa unikalne cechy semantyczne**, które pozwolą na poprawną klasyfikację

ZANURZENIA SŁÓW TYPU WORD2VEC ORAZ GLOVE

kodują pozycję słowa (ciągu znaków) w systemie języka



umożliwiają pewne operacje arytmetyczne (King – Man + Woman = Queen)



żyją **w niezwykle rzadkiej przestrzeni** wektorowej

miarą ich podobieństwa (choć niekoniecznie synonimii) jest **miara kosinusowa**

ETYKIETOWANE ZBIORY DANYCH DO EKSPERYMENTÓW

01

Maile od klientów

- 7 kategorii tematycznych
- 5000 próbek po 100 słów
- język angielski

02

Twitty

- 33 500 próbek po 10 słów
- 3 typy nacechowani (hate speech, offensive, neutral)
- język angielski

03

Literatura piękna

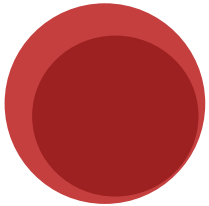
- 2464 próbek po 80 słów
- 4 kategorie autorów (Barańczak, Krasicki, Mickiewicz, Różewicz)
- język polski

01 Próbka tekstu jako macierz

długość wektora zanurzenia
x ilość słów w próbce

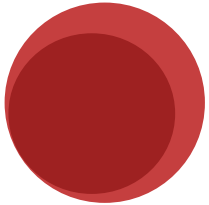
(NIE-DOBRA) ZMIANA REPREZENTACJI

SVD



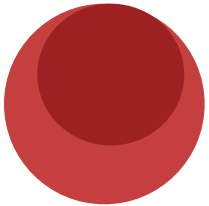
70%
trafności

PCA



69%
trafności

LAPLACIAN EIGENMAPS



65%
trafności

Belkin, Mikhail, and Partha Niyogi. "Laplacian eigenmaps for dimensionality reduction and data representation." *Neural computation* 15.6 (2003): 1373-1396.

KING
[0.34 0.15 -0.37 3.76 ...]

02 Pojedyncze wektory

Każde zanurzenie słowa traktowane osobno

DROGA DO CELU

REDUKCJA

Test na redundancję: ile elementów wektora możemy odciąć?

PERMUTACJA

Szyfrowanie z pomocą klucza w postaci macierzy mieszającej

FILTRACJA

Pierwszy etap potraktowania wektorów zanurzeń jak sygnałów

TRANSFORMATA

Dalsza eksploracja operacji sygnałowych – podstawowych oraz zmodyfikowanych

REDUKCJA KOLUMN

- Wektory GLOVE WIKI GIGAWORD 100 zostały obcięte o kolejne kolumny:
  10 pierwszych, 10 ostatnich, 10 pierwszych i ostatnich, co 2, co 5 lub co 7 kolumna

- Obcięcie nawet aż 50% kolumn wektora przyniosło pogorszenie klasyfikacji
  średnio tylko o 3 do 5%...

Trafność	Klasa 1	Klasa 2	Klasa 3	Klasa 4	Klasa 5	Klasa 6	Klasa 7
Glove R^{80}	**100%**	82%	**92%**	81%	92%	92%	**95%**
Glove R^{90}	95%	81%	89%	**84%**	90%	88%	94%
Glove R^{100}	**100%**	**87%**	89%	80%	**97%**	**94%**	92%

REDUKCJA KOLUMN

- ...ale wciąż zapewniało pełną odwracalność wartości:

wektorem najbliższym do tak obciętego (jeśli wypełnić brakujące miejsca zerami) i tak **był nasz wektor oryginalny**!

- kąt się zmieniał, ale wciąż zbyt mało:

np. dla wyrazu „see”

wektor zredukowany i wypełniony zerami osiąga miarę podobieństwa **0.7635252**, ale wciąż jest **na pierwszym miejscu** w rankingu „most_similar”

DLACZEGO TAK?

Odpytywanie słownika zanurzeń polega na szukaniu wektora najbliższego w sensie miary kosinusowej:

$$\cos\theta = \frac{a^T b}{\|a\|\|b\|}$$

Wstawianie kolejnych zer zmniejszało licznik,
ale i mianownik, wobec czego zmiana kąta była niewielka

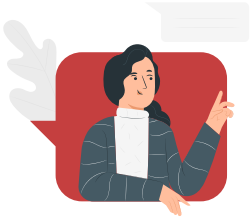
...trzeba więc zmodyfikować wektor tak, by zmieniła się wartość iloczynu skalarnego, ale nie zmieniła się norma

MACIERZ PERMUTACJI

- ✓ **zachowuje** oryginalne wartości elementów, ale miesza ich porządek,
- ✓ **zachowuje** bardzo wysoką jakość klasyfikacji (92% i 90% dla zbioru 01 i 02),
- ✓ **zmienia** wynik iloczynu skalarnego, podczas gdy norma zostaje ta sama,
- ✓ **zmienia** więc kosinus kąta...

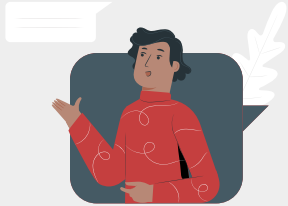
...ale istnieje gdzieś klucz w postaci wybranej macierzy permutacji

LOST IN TRANSLATION



Oryginalny tekst kodowany za pomocą GLOVE

„I think really hot I wanna see more picture please send it to me directly”

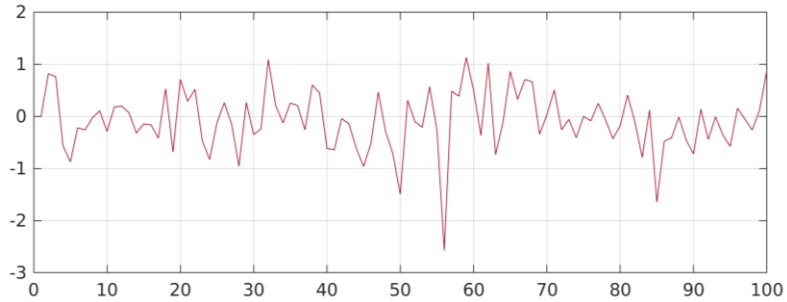


Najbliższe wektory GLOVE do tych uzyskanych po permutacji elementów wektora

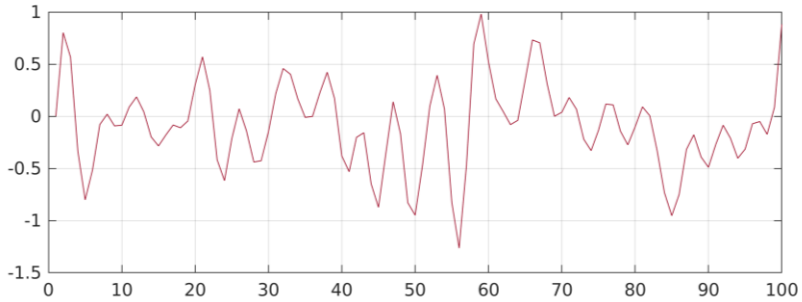
„27.90 94.98 overharvesting ibish 27.90 porn inbreeding skateboard erosion dogger conformably”

FILTRACJA DOLNO-, GÓRNO- I ŚREDNIOPRZEPUSTOWA

- ✓ zachowuje bardzo wysoką jakość klasyfikacji,
- zachowuje również pozycję w przestrzeni wektorów, wobec czego wektor najbliższy do otrzymanego to wciąż ten oryginalny.

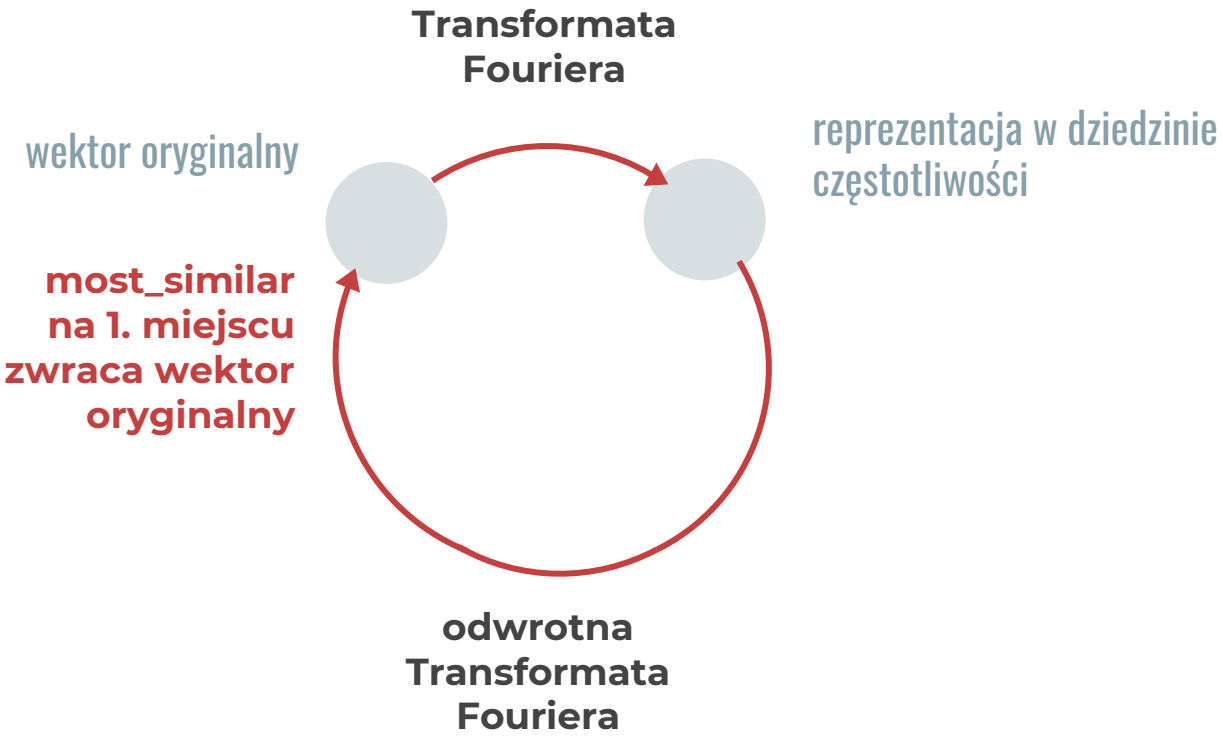


Oryginalny wektor dla słowa „think”

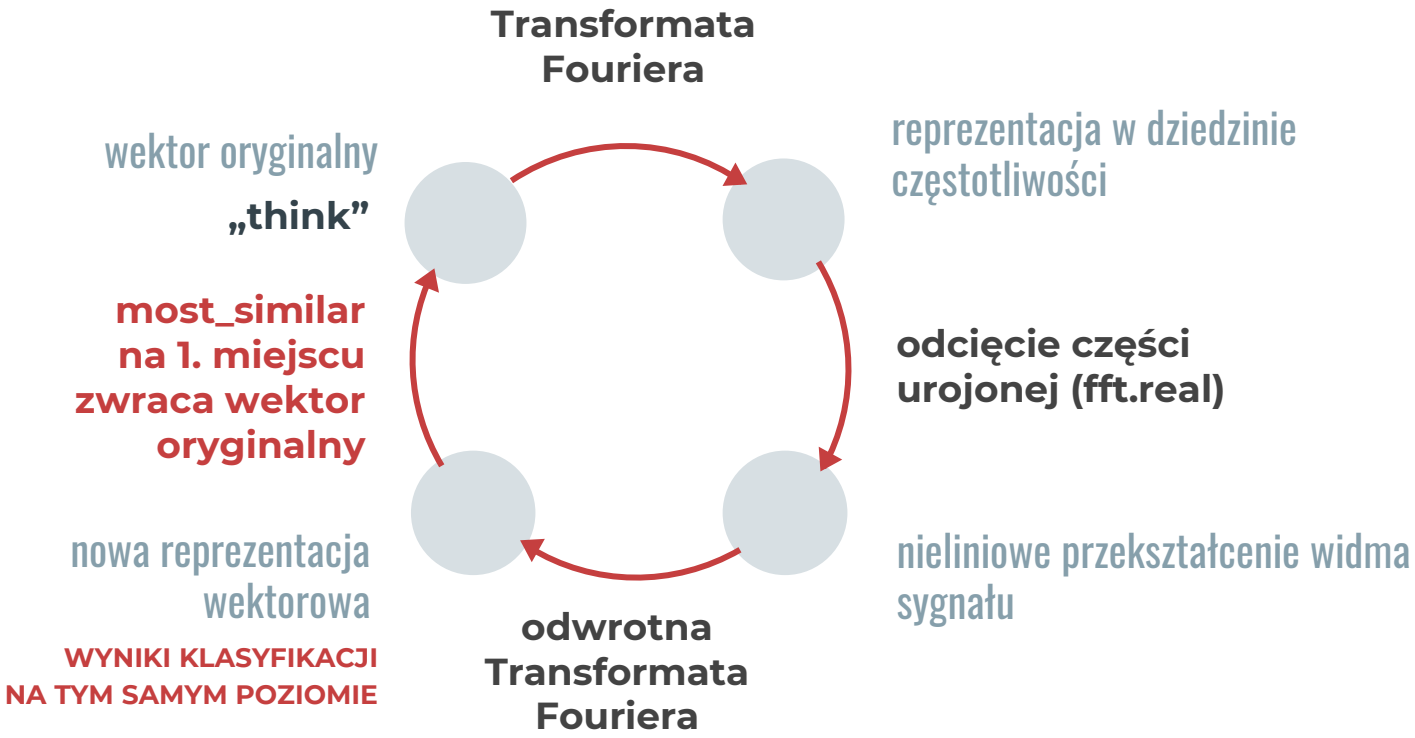


Filtr dolnoprzepustowy (0.5 pasma)

WYKORZYSTANIE TRANSFORMATY FOURIERA



WYKORZYSTANIE TRANSFORMATY FOURIERA



CIEKAWOSTKA

Jak bardzo odporne są nasze wektory na wsteczną lokalizację?
Wykonanie następujących przekształceń:

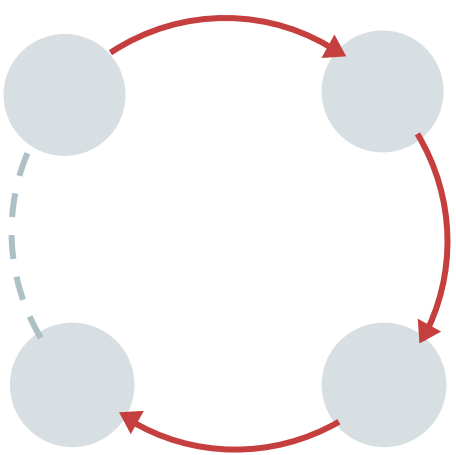
1. transformata Fouriera wektora oryginalnego,
2. odcięcie części urojonej z wyniku,
3. zastąpienie co drugiej wartości współrzędną sąsiednią,

```
[-0.046539  0.61966  0.56647  -0.46584  -1.189  0.44599  …]
[-0.046539 -0.046539  0.56647  0.56647  -1.189  -1.189  …]
```

4. transformata odwrotna z tak otrzymanego wektora,

sprawia, że wektorem najbliższym do tak „zdeformowanego” wektora jest wciąż ten oryginalny!

WYKORZYSTANIE TRANSFORMATY FOURIERA



Transformata Fouriera

wektor oryginalny

reprezentacja w dziedzinie częstotliwości

dostatecznie duża odległość wektorów

1. odcięcie części urojonej (fft.real)
2. wartość bezwzględna (abs(fft.real))

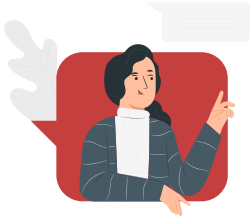
nowa reprezentacja wektorowa

nieliniowe przekształcenie widma sygnału

odwrotna Transformata Fouriera

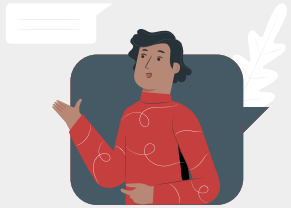
BARDZO DOBRE LUB NIEWIELE GORSZE WYNIKI KLASYFIKACJI

LOST IN TRANSLATION



Oryginalny tekst kodowany
za pomocą GLOVE

*„I think really hot
I wanna see more
picture please send
it to me directly”*



Najbliższe wektory GLOVE do tych uzyskanych
po zmodyfikowanej Transformacie Fouriera

*„sugarhill sugarhill sugarhill
gtb sugarhill wends
sugarhill chicas sugarhill
sugarhill boeken”*

KLASYFIKACJA PRZY UŻYCIU MASKOWANYCH WEKTORÓW GLOVE R^{100} (EN)

Trafność	Klasa 1	Klasa 2	Klasa 3	Klasa 4	Klasa 5	Klasa 6	Klasa 7
Oryginalne Glove	93%	96%	78%	85%	81%	92%	93%
Maskowane Wektory GLOVE	97%	89%	84%	83%	81%	92%	84%

KLASYFIKACJA PRZY UŻYCIU MASKOWANYCH WEKTORÓW WORD2VEC^{100} (PL)

Trafność	Klasa 1	Klasa 2	Klasa 3	Klasa 4
Oryginalne word2vec	93%	93%	89%	93%
Maskowane Wektory word2vec	86%	86%	75%	84%

- ✓ Niemożliwa rekonstrukcja oryginalnego sygnału
- ✓ Wystarczająco zmieniona odległość kosinusowa
- ✓ Zachowana akceptowalna jakość klasyfikacji tekstu

DZIĘKUJĘ ZA UWAGĘ

https://www.researchgate.net/profile/Inez_Okulska
https://www.linkedin.com/in/inezokulska/

© szablon prezentacji Slidesgo, ikony Flaticon, infografiki Freepik oraz ilustracje Stories