Semantyka dystrybucyjna

Zaawansowane Przetwarzanie Języka Naturalnego

Mateusz Lango

Zakład Inteligentnych Systemów Wspomagania Decyzji
Wydział Informatyki i Telekomunikacji
Politechnika Poznańska

„Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”,
projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

Agenda

1. Czym jest znaczenie wyrazu?
2. Modelowanie znaczenia wyrazu sieciami semantycznymi
3. Semantyka dystrybucyjna

Czym jest znaczenie wyrazu?

- Pytanie NLP: w jaki sposób reprezentować znaczenie wyrazów na komputerze?
- Jednak czym jest znaczenie? Jak zamodelować znaczenie: „dobra", „piękna", „sprawiedliwości"?
- Semantyka – dział językoznawstwa zajmujący się znaczeniem
- Wg. teorii de Saussure'a znak składa się z dwóch elementów:
  - elementu znaczącego (signifiant)
  - elementu znaczonego (signifié)

Określenie znaczenia słów jest trudne...[1]

Formalne określenie znaczenia niektórych słów jak przyimków wydaje się z pozoru proste np. odwołując się do geometrii.

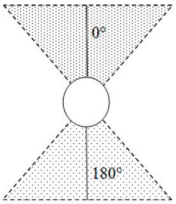


Figure 7. Above and below as regions.

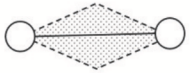


Figure 6. The region for *between*.

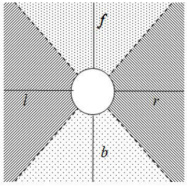


Figure 4. *In front of, behind, to the left of,* and *to the right of* as regions of the angle ϕ around a landmark.

Jak zdefiniować „na" albo „w"?

- „na" określa element znajdujący się w regionie „above" i dodatkowo stykając się z obiektem referencyjnym.
- „w" określa obiekt znajdujący się w regionie określonym przez objętość obiektu referencyjnego.

[1] Peter Gärdenfors: The Geometry Of Preposition Meanings

Określenie znaczenia słów jest trudne... [2]

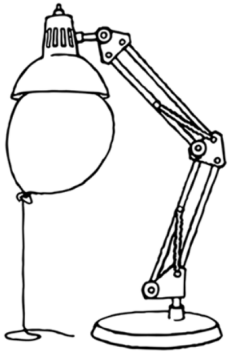


Figure 11. Is the lamp *on* the balloon?

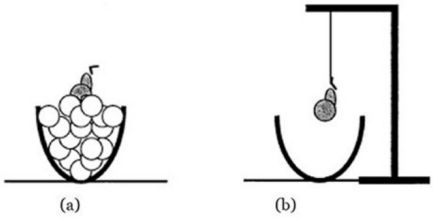


Figure 9. (a) The pear is in the bowl. (b) The pear is not in the bowl (from Garrod et al. 1999, p. 168).

[2] Peter Gärdenfors: The Geometry Of Preposition Meanings

Czym jest znaczenie?

- Znaczenie wyrazu – idea, obraz w umyśle związany z wyrazem
  - może mieć charakter indywidualny
  - zespół skojarzeń w całej wspólnocie językowej
  - odniesienie do przedmiotów w rzeczywistości

Problem w modelowaniu słów: polisemia

- Polisemia czyli wieloznaczność słów
- Najpopularniejsze czasowniki w języku angielskim (20% wystąpień) mają średnio 12 znaczeń, analogicznie częste rzeczowniki mają 7.8 znaczeń
- Przykłady:
  - jeść jabłko vs jeść zupę
  - „Budka z lodami będzie czynna w sezonie" vs „Wyciąg narciarski będzie czynny w sezonie"
  - „Doroszewski był polskim leksykografem" vs „Na półce stoi Doroszewski"
- Każde słowo jest wieloznaczne ze względu na użycie – polisemia pozorna
- Ludwig Wittgenstein „Nie szukajcie znaczenia wyrazu, szukajcie jego użyć"

Hipoteza dystrybucyjna

You shall know a word by the company it keeps

J. R. Firth, 1957.

Hipoteza dystrybucyjna

You shall know a word by the company it keeps

J. R. Firth, 1957.

Przykład

czarny ? miałknął

ukochany ? pił mleko

Hipoteza dystrybucyjna

You shall know a word by the company it keeps

J. R. Firth, 1957.

Przykład

czarny kotek miałknął

ukochany kotek pił mleko

Hipoteza dystrybucyjna

You shall know a word by the company it keeps

J. R. Firth, 1957.

Przykład

czarny kotek miałknął
ukochany kotek pił mleko
wyleniały kot miałknął
mój czarny kot zamruczał z zadowoleniem

Hipoteza dystrybucyjna

You shall know a word by the company it keeps
J. R. Firth, 1957.

Przykład

czarny kotek miałknął
ukochany kotek pił mleko
wyleniały kot miałknął
mój czarny kot zamruczał z zadowoleniem
duży czarny kot/ kotek ??

WordNet: graf relacji pomiędzy słowami

- Różne znaczenia słów próbuje się zdyskretyzować w jednostki leksykalne
- Wierzchołkami grafu są tzw. word senses które reprezentują jeden aspekt znaczenia wyrazu
- Pozwala to na modelowanie homonimów (słowa które mają różne znaczenia):
  - zamek [PL]
  - bank [EN] (finanse/brzeg)

WordNet: graf relacji pomiędzy słowami

Relacje semantyczne pomiędzy jednostkami leksykalnymi:

- Synonimia – łączy wyrazy bliskoznaczne np. matka, mama, mamusia.
  - Nie ma wyrazów które znaczą dokładnie to samo
  - Wyrazy bliskoznaczne to takie które w pewnych sytuacjach mogą zostać zastąpione

WordNet: graf relacji pomiędzy słowami

Relacje semantyczne pomiędzy jednostkami leksykalnymi:

- Synonimia – łączy wyrazy bliskoznaczne np. matka, mama, mamusia.
  - Nie ma wyrazów które znaczą dokładnie to samo
  - Wyrazy bliskoznaczne to takie które w pewnych sytuacjach mogą zostać zastąpione
- Hiponimia – łączy węższe pojęcie z szerszym
- Hiperonimia – łączy szersze pojęcie z węższym

WordNet: graf relacji pomiędzy słowami

Relacje semantyczne pomiędzy jednostkami leksykalnymi:

- Synonimia – łączy wyrazy bliskoznaczne np. matka, mama, mamusia.
  - Nie ma wyrazów które znaczą dokładnie to samo
  - Wyrazy bliskoznaczne to takie które w pewnych sytuacjach mogą zostać zastąpione
- Hiponimia – łączy węższe pojęcie z szerszym
- Hiperonimia – łączy szersze pojęcie z węższym
- Meronimia – część całości
- Holonimia – całość z części

WordNet: graf relacji pomiędzy słowami

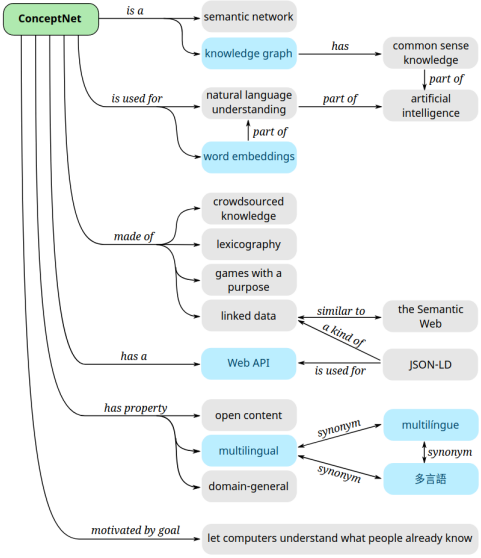
Relacje semantyczne pomiędzy jednostkami leksykalnymi:

- Synonimia – łączy wyrazy bliskoznaczne np. matka, mama, mamusia.
  - Nie ma wyrazów które znaczą dokładnie to samo
  - Wyrazy bliskoznaczne to takie które w pewnych sytuacjach mogą zostać zastąpione
- Hiponimia – łączy węższe pojęcie z szerszym
- Hiperonimia – łączy szersze pojęcie z węższym
- Meronimia – część całości
- Holonimia – całość z części
- Antonimia, Mieszkaniec (Szczecin-szczecinianin), Stopniowanie (dobry-lepszy),...

Problem

Jak policzyć podobieństwo między jednostkami leksykalnymi w WordNet?

Również inne podobne zasoby...



Semantyka dystrybucyjna: co to znaczy że słowa są podobne?

- podobieństwo właściwości – mają wspólne właściwości, atrybuty
  - szczególny przypadek: podobieństwo semantyczne: jeśli słowa współdzielą hiperonim
- podobieństwo relacyjne pomiędzy parami słów
  - słowa semantycznie powiązane, takie które występują obok siebie w tekście

Patrząc na powyższe możemy wyróżnić dwa typy dystrybucji słów

- współwystępowanie (semantycznie powiązane)
- występowanie na podobnych pozycjach w zadaniu (semantycznie podobne)

Przestrzeń semantyczna

Przestrzeń semantyczna to przestrzeń (najczęściej wielowymiarowa) w której słowa lub koncepty są reprezentowane jako punkty, a ich pozycja wzdłuż każdej z osi jest powiązana ze znaczeniem słowa.

Przestrzeń semantyczna jest użyteczna do określenia relacji między słowami – można je skwantyfikować funkcją odległości.
Podejścia tradycyjne: zdefiniujmy znaczenie słowa i umieśćmy (ręcznie) słowo na osi.

Hyperspace Analogue to Language (HAL)

- Problem metod tradycyjnych: jak wybrać i zdefiniować wymiary?
- Metoda HAL: taka przestrzeń powstaje poprzez proste zliczanie
  - parametr metody: wielkość okienka (w oryginalnej terminologii: promień)
  - zliczamy słowa które mieszczą się w okienku z wagami
  - waga maleje w miarę jak się oddalamy od analizowanego słowa

Rozszerzenia: symetryczne okienko

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	0	0	0	0	0	0	0
like	0	0	0	0	0	0	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	0	0	0	0	0	0	0
Ret.	0	0	0	0	0	0	0	0
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	0	0	0	0	0	0	0	0

Rozszerzenia: symetryczne okienko

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	1	0	0	0	0	0	0
like	0	0	0	0	0	0	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	0	0	0	0	0	0	0
Ret.	0	0	0	0	0	0	0	0
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	0	0	0	0	0	0	0	0

Rozszerzenia: symetryczne okienko

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	1	0	0	0	0	0	0
like	1	0	0	1	0	0	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	0	0	0	0	0	0	0
Ret.	0	0	0	0	0	0	0	0
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	0	0	0	0	0	0	0	0

Rozszerzenia: symetryczne okienko

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	1	0	0	0	0	0	0
like	1	0	0	1	0	0	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	0	0	0	0	0
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	0	0	0	0	0	0	0	0

Rozszerzenia: symetryczne okienko

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	1	0	0	0	0	0	0
like	1	0	0	1	0	0	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	1	0	0	0	1
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	0	0	0	0	0	0	0	0

Rozszerzenia: symetryczne okienko

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	1	0	0	0	0	0	0
like	1	0	0	1	0	0	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	1	0	0	0	1
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	1	0	0	0	1	0	0	0

Rozszerzenia: symetryczne okienko

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	2	0	0	0	0	0	1
like	1	0	0	1	0	0	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	1	0	0	0	1
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	1	0	0	0	1	0	0	0

Rozszerzenia: symetryczne okienko

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	2	0	0	0	0	0	1
like	2	0	0	1	0	1	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	1	0	0	0	1
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	1	0	0	0	1	0	0	0

Rozszerzenia: symetryczne okienko

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	2	0	0	0	0	0	1
like	2	0	0	1	0	1	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	1	0	0	0	1
Stats.	0	1	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	1	0	0	0	1	0	0	0

Rozszerzenia: symetryczne okienko

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	2	1	0	0	0	0	1
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	1	0	0	0	1
Stats.	0	1	0	0	0	0	0	0
flying	0	0	1	0	0	0	0	0
and	1	0	0	0	1	0	0	0

Zastosowanie PCA na tej macierzy i poprzez wyszukiwanie najbliższego sąsiada odpowiadanie na pytania testowe o wybór synonimów na TOEFL.
Semantyka dystrybucyjna: 92.5% vs. Przeciętny egzaminowany: 64.5%

Rozszerzenia: symetryczne okienko

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	2	1	0	0	0	0	1
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	1	0	0	0	1
Stats.	0	1	0	0	0	0	0	0
flying	0	0	1	0	0	0	0	0
and	1	0	0	0	1	0	0	0

Zastosowanie PCA na tej macierzy i poprzez wyszukiwanie najbliższego sąsiada odpowiadanie na pytania testowe o wybór synonimów na TOEFL.
Semantyka dystrybucyjna: 92.5% vs. Przeciętny egzaminowany: 64.5%
Co wystarcza by dostać się na wiele uniwersytetów w US!

Czym wypełnić macierz kontekst-dokument?

Pojawia się pytanie czy uzupełnianie macierzy słowo-kontekst zliczeniami jest dobrym rozwiązaniem? Czy nie powinno się zastąpić tych liczności jakimiś miarami asocjacji?

Czym wypełnić macierz kontekst-dokument?

Pojawia się pytanie czy uzupełnianie macierzy słowo-kontekst zliczeniami jest dobrym rozwiązaniem? Czy nie powinno się zastąpić tych liczności jakimiś miarami asocjacji?

$P(c|\text{computer})$

the 0.032
a 0.019
is 0.014
we 0.008
...
text 0.00018
...
program 0.00013
software 0.0001

Adapted from a presentation by ChengXiang Zhai „Probabilistic Retrieval Model: Statistical Language Model"

Czym wypełnić macierz kontekst-dokument?

Pojawia się pytanie czy uzupełnianie macierzy słowo-kontekst zliczeniami jest dobrym rozwiązaniem? Czy nie powinno się zastąpić tych liczności jakimiś miarami asocjacji?

$P(c\|\text{computer})$
the 0.032
a 0.019
is 0.014
we 0.008
...
text 0.00018
...
program 0.00013
software 0.0001

$P(c)$
the 0.03
a 0.02
is 0.015
we 0.01
..
text 0.00006
...
program 0.00000125
software 0.000000667

Adapted from a presentation by ChengXiang Zhai „Probabilistic Retrieval Model: Statistical Language Model"

Czym wypełnić macierz kontekst-dokument?

Pojawia się pytanie czy uzupełnianie macierzy słowo-kontekst zliczeniami jest dobrym rozwiązaniem? Czy nie powinno się zastąpić tych liczności jakimiś miarami asocjacji?

$P(c\|\text{computer})$
the 0.032
a 0.019
is 0.014
we 0.008
...
text 0.00018
...
program 0.00013
software 0.0001

$P(c)$
the 0.03
a 0.02
is 0.015
we 0.01
..
text 0.00006
...
program 0.00000125
software 0.000000667

$\frac{P(c\|\text{computer})}{P(c)}$
software 150
program 104
...
text 3.0
..
the 1.1
a 0.99
is 0.9
we 0.8

Adapted from a presentation by ChengXiang Zhai „Probabilistic Retrieval Model: Statistical Language Model"

Pointwise mutual information – miara zależności między słowami

Otrzymujemy więc miarę asocjacji między słowami:

$$Association(w, c) = \frac{P(c|w)}{P(c)}$$

Pointwise mutual information – miara zależności między słowami

Otrzymujemy więc miarę asocjacji między słowami:

$$Association(w, c) = \frac{P(c|w)}{P(c)} = \frac{P(w, c)}{P(w)P(c)}$$

Pointwise mutual information – miara zależności między słowami

Otrzymujemy więc miarę asocjacji między słowami:

$$Association(w, c) = \frac{P(c|w)}{P(c)} = \frac{P(w, c)}{P(w)P(c)}$$

Która po zlogarytmowaniu przyjmuje postać punktowej wzajemnej informacji:

$$PMI(w, c) = log_2 \frac{P(w, c)}{P(w)P(c)}$$

Pointwise mutual information – miara zależności między słowami

Otrzymujemy więc miarę asocjacji między słowami:

$$Association(w, c) = \frac{P(c|w)}{P(c)} = \frac{P(w, c)}{P(w)P(c)}$$

Która po zlogarytmowaniu przyjmuje postać punktowej wzajemnej informacji:

$$PMI(w, c) = log_2 \frac{P(w, c)}{P(w)P(c)}$$

PMI jest w praktyce kłopotliwe w użyciu: co się dzieje jak $P(w, c) = 0$?

Pointwise mutual information – miara zależności między słowami

Otrzymujemy więc miarę asocjacji między słowami:

$$Association(w, c) = \frac{P(c|w)}{P(c)} = \frac{P(w, c)}{P(w)P(c)}$$

Która po zlogarytmowaniu przyjmuje postać punktowej wzajemnej informacji:

$$PMI(w, c) = log_2 \frac{P(w, c)}{P(w)P(c)}$$

PMI jest w praktyce kłopotliwe w użyciu: co się dzieje jak $P(w, c) = 0$?

Częstą praktyką w IL jest założenie że w takiej sytuacji $PMI = 0$, jednak rodzi to niespójność:

- Mamy dwa niepowiązane ze sobą słowa, które razem wystąpiły tylko raz w korpusie
- Mamy dwa niepowiązane ze sobą słowa, które nigdy nie wystąpiły razem w korpusie
- Ile wynosi PMI w tych sytuacjach?

Pointwise mutual information – miara zależności między słowami

Otrzymujemy więc miarę asocjacji między słowami:

$$Association(w, c) = \frac{P(c|w)}{P(c)} = \frac{P(w, c)}{P(w)P(c)}$$

Która po zlogarytmowaniu przyjmuje postać punktowej wzajemnej informacji:

$$PMI(w, c) = log_2 \frac{P(w, c)}{P(w)P(c)}$$

PMI jest w praktyce kłopotliwe w użyciu: co się dzieje jak $P(w, c) = 0$?
Częstą praktyką w IL jest założenie że w takiej sytuacji $PMI = 0$, jednak rodzi to niespójność:

- Mamy dwa niepowiązane ze sobą słowa, które razem wystąpiły tylko raz w korpusie
- Mamy dwa niepowiązane ze sobą słowa, które nigdy nie wystąpiły razem w korpusie
- Ile wynosi PMI w tych sytuacjach?
- Ludzie łatwo wskazują powiązane słowa: Grenlandia-śnieg, a negatywnie powiązane?

Pointwise mutual information – miara zależności między słowami

Otrzymujemy więc miarę asocjacji między słowami:

$$Association(w, c) = \frac{P(c|w)}{P(c)} = \frac{P(w, c)}{P(w)P(c)}$$

Która po zlogarytmowaniu przyjmuje postać punktowej wzajemnej informacji:

$$PMI(w, c) = log_2 \frac{P(w, c)}{P(w)P(c)}$$

PMI jest w praktyce kłopotliwe w użyciu: co się dzieje jak $P(w, c) = 0$?
Częstą praktyką w IL jest założenie że w takiej sytuacji $PMI = 0$, jednak rodzi to niespójność:

- Mamy dwa niepowiązane ze sobą słowa, które razem wystąpiły tylko raz w korpusie
- Mamy dwa niepowiązane ze sobą słowa, które nigdy nie wystąpiły razem w korpusie
- Ile wynosi PMI w tych sytuacjach?
- Ludzie łatwo wskazują powiązane słowa: Grenlandia-śnieg, a negatywnie powiązane?

Rozwiązanie:

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right)$$

Obliczenia PPMI – przykład

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = \textit{information}, c = \textit{data}) = \frac{6}{19}$$

Obliczenia PPMI – przykład

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = \textit{information}, c = \textit{data}) = \frac{6}{19}$$

$$P(w = \textit{information}) = \frac{11}{19}$$

Obliczenia PPMI – przykład

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = information, c = data) = \frac{6}{19}$$

$$P(w = information) = \frac{11}{19}$$

$$P(c = data) = \frac{7}{19}$$

Obliczenia PPMI – przykład

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = information, c = data) = \frac{6}{19}$$

$$P(w = information) = \frac{11}{19}$$

$$P(c = data) = \frac{7}{19}$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) = \max\left(0, \log_2 \frac{\frac{6}{19}}{\frac{11}{19} \cdot \frac{7}{19}}\right) = 0.568$$

Obliczenia PPMI – przykład

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = digital, c = data) = ?$$

$$P(w = digital) = ?$$

$$P(c = data) = ?$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) =$$

Obliczenia PPMI – przykład

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = digital, c = data) = \frac{1}{19}$$

$$P(w = digital) = ?$$

$$P(c = data) = ?$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) =$$

Obliczenia PPMI – przykład

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = digital, c = data) = \frac{1}{19}$$

$$P(w = digital) = \frac{4}{19}$$

$$P(c = data) = ?$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) =$$

Obliczenia PPMI – przykład

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = digital, c = data) = \frac{1}{19}$$

$$P(w = digital) = \frac{4}{19}$$

$$P(c = data) = \frac{7}{19}$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) =$$

Obliczenia PPMI – przykład

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = digital, c = data) = \frac{1}{19}$$

$$P(w = digital) = \frac{4}{19}$$

$$P(c = data) = \frac{7}{19}$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) = \max\left(0, \log_2 \frac{\frac{1}{19}}{\frac{4}{19} \cdot \frac{7}{19}}\right) = \max(0, -0.26) = 0$$

PPMI – uwagi końcowe

- W metodach semantyki dystrybucyjnej wyznaczamy (P)PMI dla wszystkich par słowo-kontekst i uzupełniamy nimi całą macierz $\Rightarrow$ Macierz (P)PMI
- Podobieństwo między słowami możemy policzyć funkcją odległości pomiędzy wierszami tej macierzy
- Macierz tę możemy też przetworzyć metodami redukcji wymiarowości jak np. PCA

Do zobaczenia!

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego