



OŚRODEK
PRZETWARZANIA
INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Jak spróbować w NLP *)

* i się przy tym nie połamać

Maciej Kowalski

31.05.2022



OŚRODEK
PRZETWARZANIA
INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Metody detekcji planet spoza Układu Słonecznego

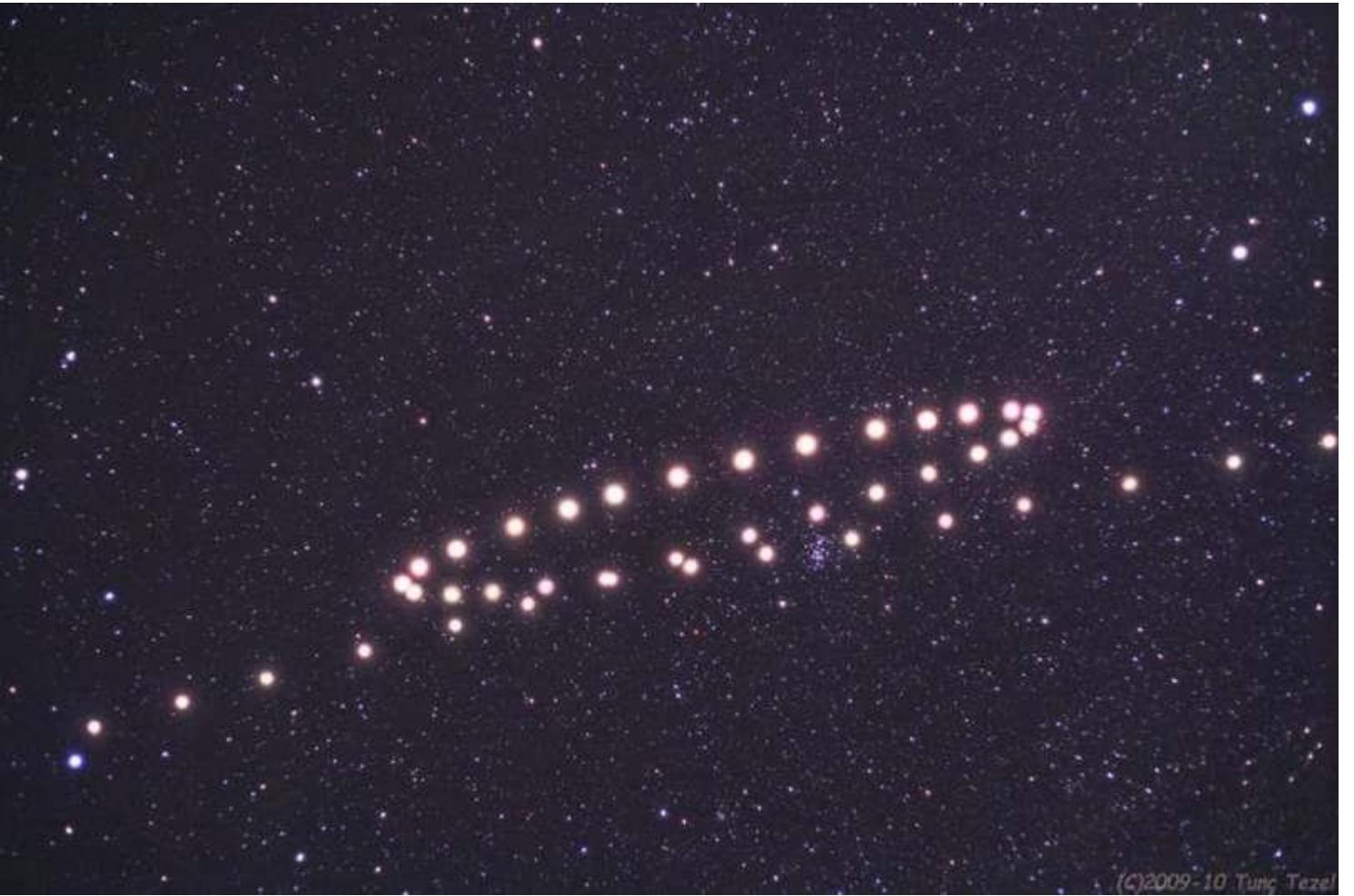
Maciej Kowalski

31.05.2022

O czym będziemy mówić...

- Astrometria,
- Soczewkowanie (mikro) grawitacyjne,
- Obrazowanie (koronografia),
- Metoda tranzytowa.

A na początku było...



<https://phys.org/news/2018-08-aboriginal-traditions-complex-motions-planets.html>



OŚRODEK
PRZETWARZANIA
INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY



Jak spróbować w NLP *)

* i się przy tym nie połamać

Maciej Kowalski

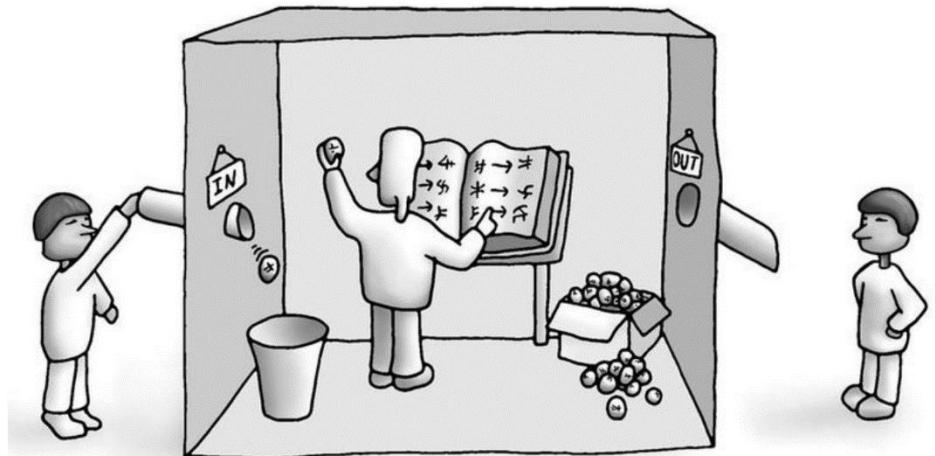
31.05.2022

Dla Was

- <https://github.com/dotjabber/alk-nlp-intro>
- <https://colab.research.google.com>

NLP?

- **NLP** – *Natural Language Processing* - Właściwie wszystko, co jest związane z przetwarzaniem informacji zapisanej w języku naturalnym,
- **NLU** – *Natural Language Understanding* - Semantyka i logika, Rozumienie nie zawsze okazuje się niezbędne (Chiński Pokój),
- **NLG** – *Natural Language Generation* - Generowanie tekstu (w języku naturalnym).



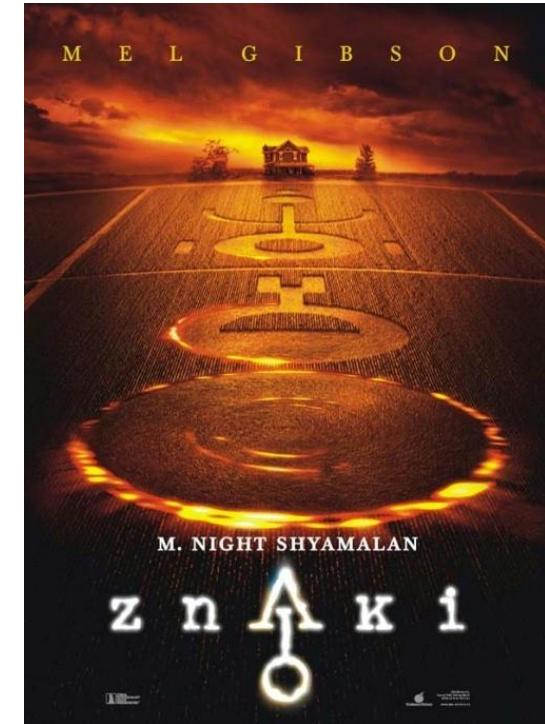
Przykłady

The collage consists of several screenshots from different applications:

- Google Wiadomości:** A screenshot of the Google News interface showing the 'Najważniejsze artykuły' (Most important articles) section. Headlines include news about COVID-19 and political events in Warsaw and Ukraine.
- Pogoda lokalna:** A weather forecast for Warsaw, showing a high of 13°C with rain on Saturday and Sunday.
- JSA (JEDNOLITY SYSTEM ANTYPLAGIATOWY):** A screenshot of the plagiarism detection system's dashboard. It shows a 'Statystyka' (Statistics) section with counts for various types of matches (Znaki, Wyrazy, Nierozpoznane wyrazy, Elementy graficzne, Fragmenty innego stylu) and a 'Rozkład długości wyrazów' (Length distribution of words) chart.
- InPost - Paczkomaty, Kurier:** A screenshot of a mobile application or website for InPost delivery services. It includes a text input field for tracking packages ('Wprowadź tekst') and a message from the service provider.

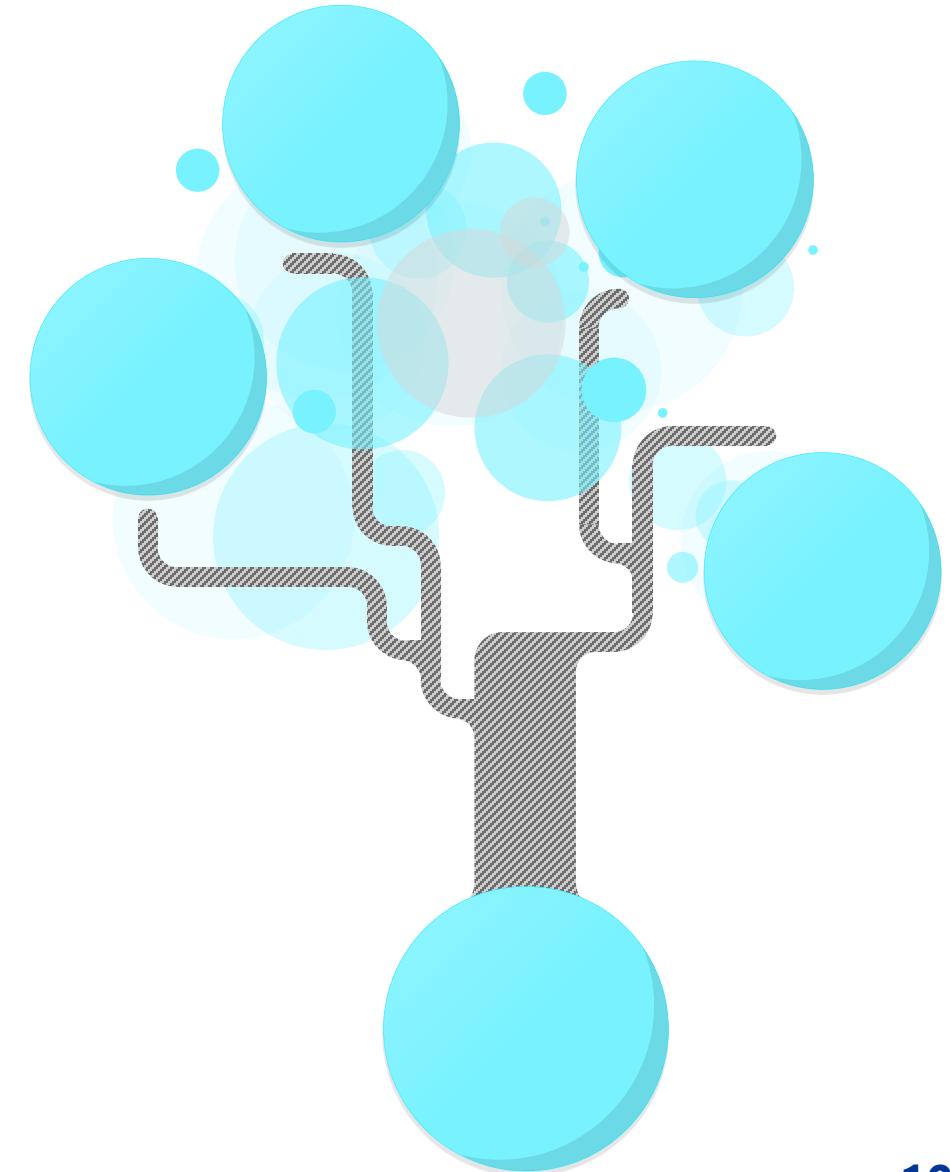
Składnik: tokenizacja

- Na początku były znaki...
- Ciągi liter/cyfr z punktu widzenia komputera wszystkie równorzędne,
- Dzielenie tekstu na mniejsze fragmenty, które zawierają w sobie spójną informację.



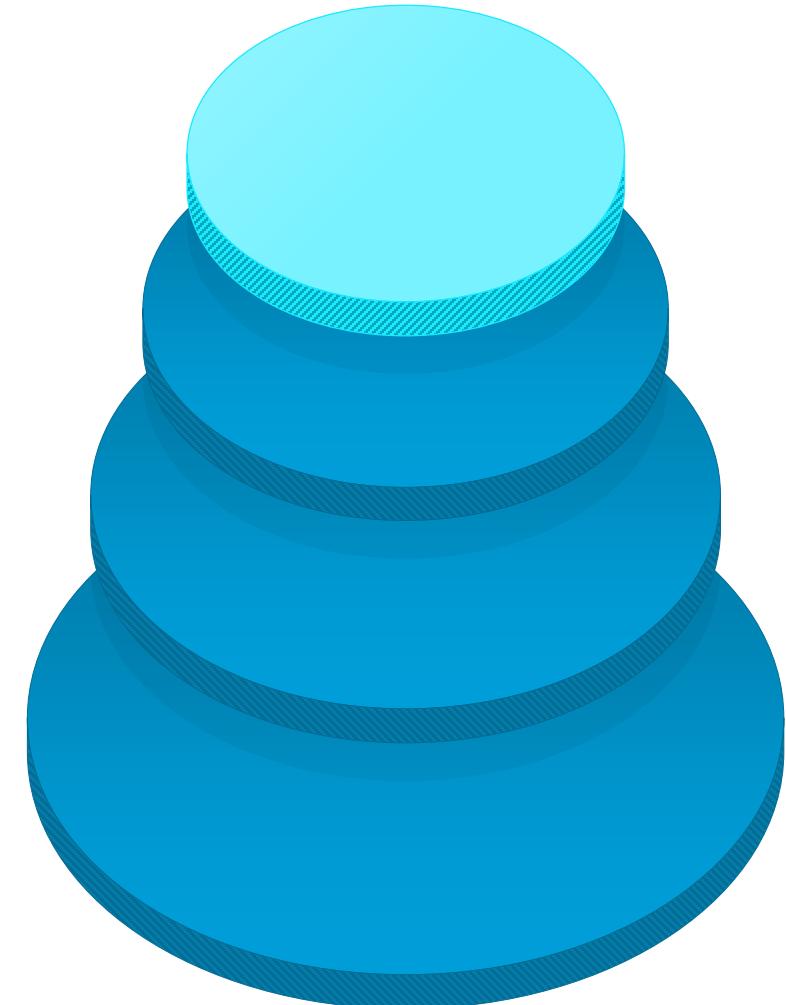
Składnik: stemming/lematyzacja

- Leksem/Lemma - oznaczenie wszystkich form fleksyjnych danego słowa/podstawowa forma leksemu.
- Stemming – automatyczne odnajdywanie rdzeni wyrazów. Większość stemmerów nie zapewnia tego, iż tworzone przez nie ciągi liter to rzeczywiście rdzenie – nie jest to jednak istotne, tak długo jak dla wszystkich wyrazów należących do danego leksemu otrzymujemy taki sam rdzeń.



Składnik: PoS tagging

- Lingwistyka grupuje słowa w zbiory, według ich podobnego zachowania w zdaniach (składni) i podobieństwa cech morfologicznych.
- Nazwy – części mowy (part of speech – PoS), kategorie syntaktyczne, kategorie gramatyczno-leksykalne itp.
 - rzeczownik – opis rzeczy (przedmiotów, pojęć itp.),
 - czasownik – opis działania, akcji,
 - przyimkownik – opis cech rzeczowników.

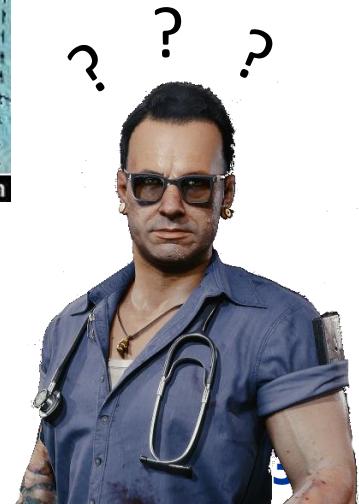
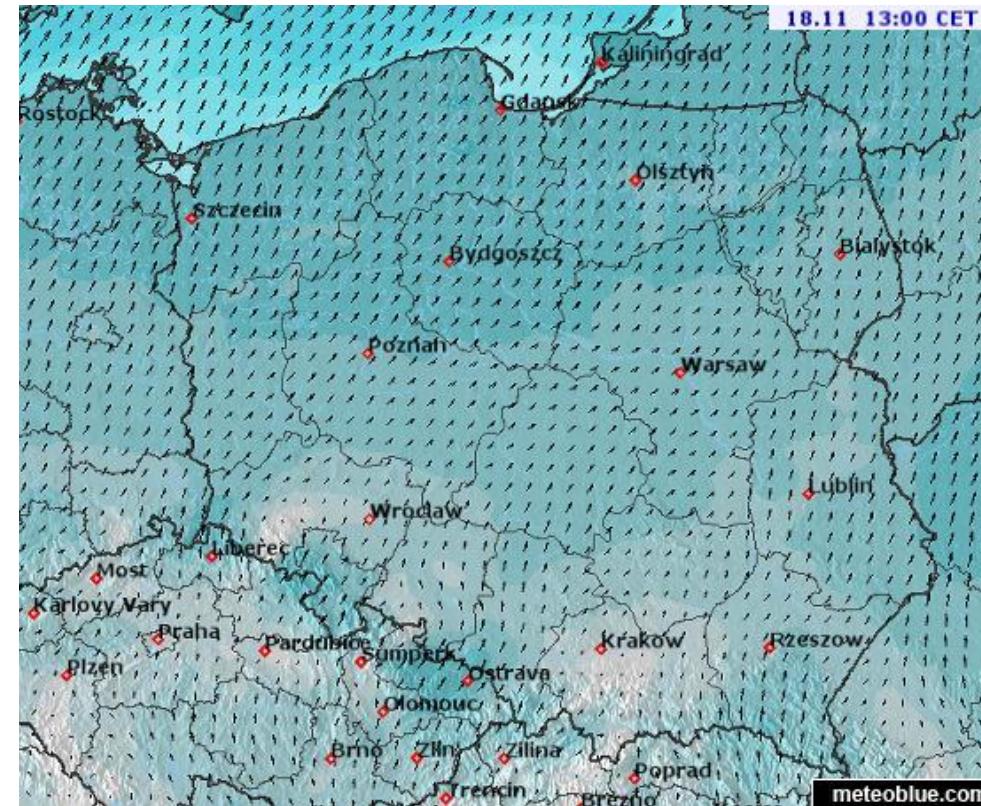
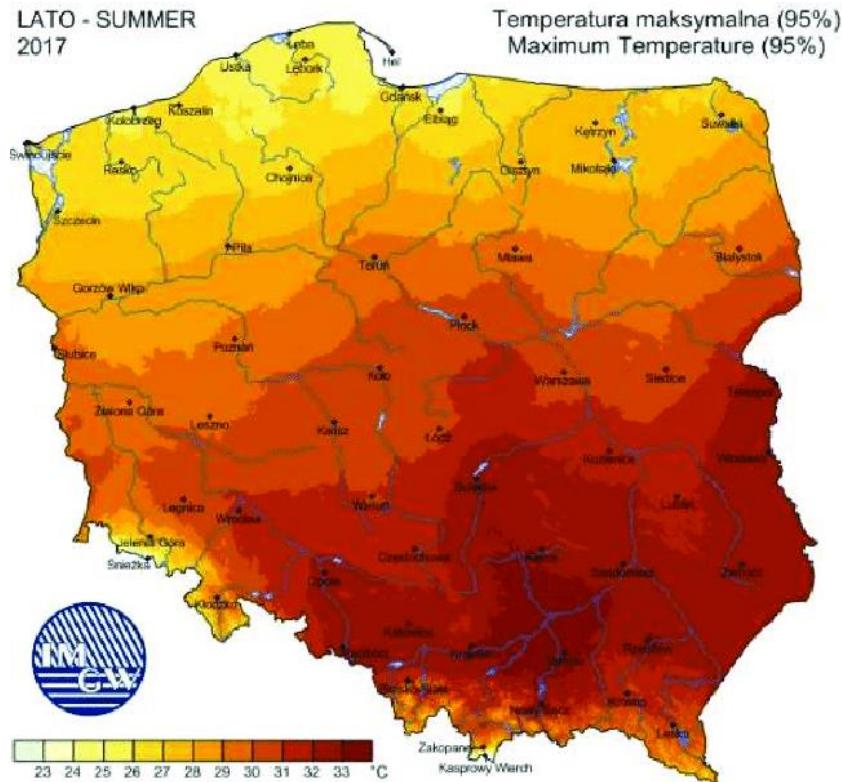


Składnik: Stop-words

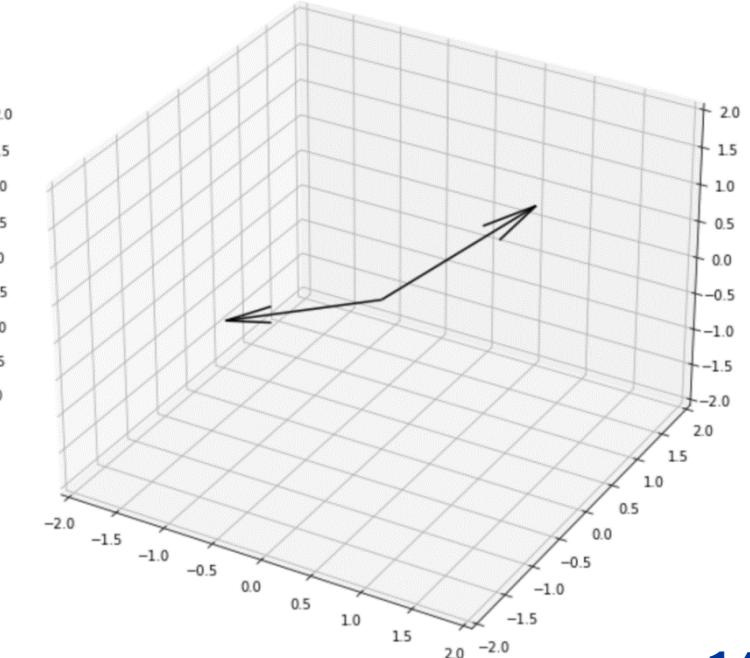
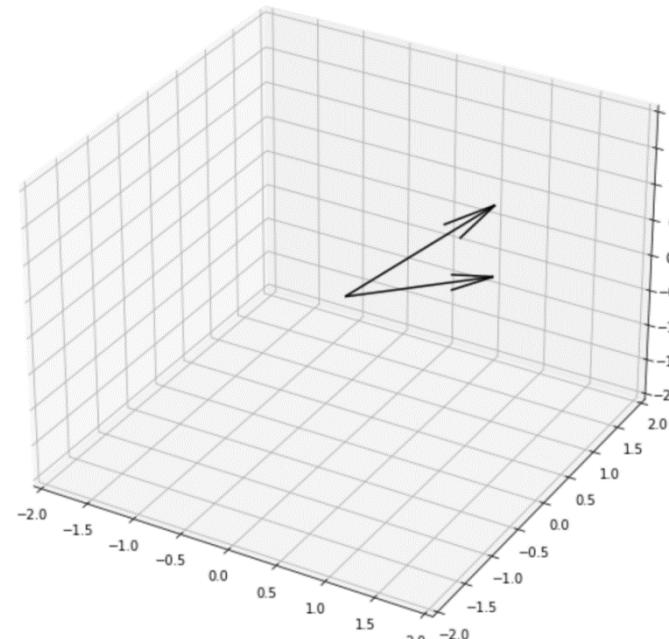
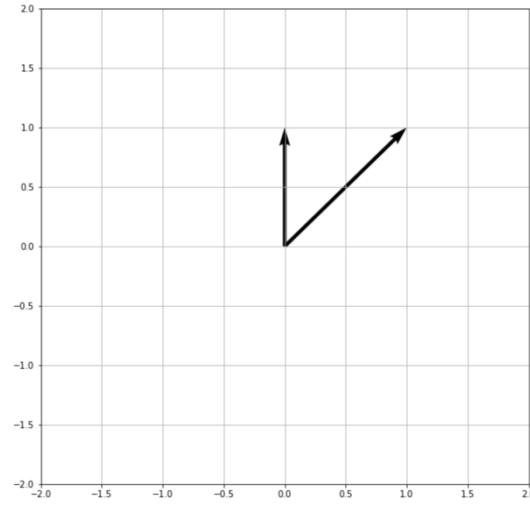
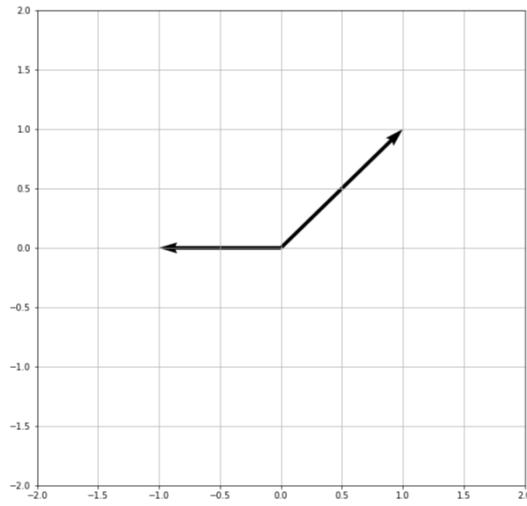
- Czy są słowa, które nie dają (prawie) żadnych informacji?
 - <https://pl.wiktionary.org/wiki/Indeks:Polski - Najpopularniejsze słowa 1-1000 wersja Jerzego Kazojskiego>

a	bardzo	ci	dla	którym	nasze	pomimo
aby	beda	cie	dlaczego	którzy	naszego	ponad
ach	bedzie	ciebie	dlatego	ku	naszych	poniewaz
acz	bez	cię	do	lat	natomiaszt	ponieważ
aczkolwiek	deda	co	dobrze	lecz	natychmiast	powinien
aj	będą	cokolwiek	iz	lub	nawet	powinna
albo	bede	cos	iż	ma	nia	powinni
ale	będę	coś	ja	mają	nią	powinno
alez	będzie	czasami	jak	mało	nic	poza

Wektry



Jak komputery mierzą prawdopodobieństwo?

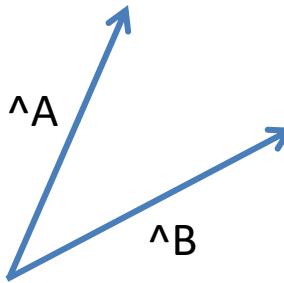


Jak komputery mierzą prawdopodobieństwo?

- Miary odległości: Manhattan, Euklides



- Miary podobieństwa: Cosinus



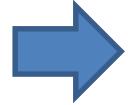
$$\cos(A, B) = \frac{A \cdot B}{|A| * |B|}$$

Dobra, ale po co nam to?

NLP – w kontekście nowych technologii oznacza przetwarzanie języka komunikacyjnego, odpowiedzącego za kolejny etap komunikacji. NLU – czyli rozumienie języka naturalnego odpowiada za dane maszynowe, na podstawie których NLG – czyli tworzenie naturalnego języka. Na podstawie zapotrzebowania i danych, które wprowadziły się do systemu, musi być określony sposób, jakiego bezpieczeństwa dla użytkownika. Proces dekodowania i kodowania informacji na dalszych krokach.

NLU – czyli rozumienie języka naturalnego odpowiada za dane maszynowe, na podstawie których NLG – czyli tworzenie naturalnego języka. Na podstawie zapotrzebowania i danych, które wprowadziły się do systemu, musi być określony sposób, jakiego bezpieczeństwa dla użytkownika. Proces dekodowania i kodowania informacji na dalszych krokach.

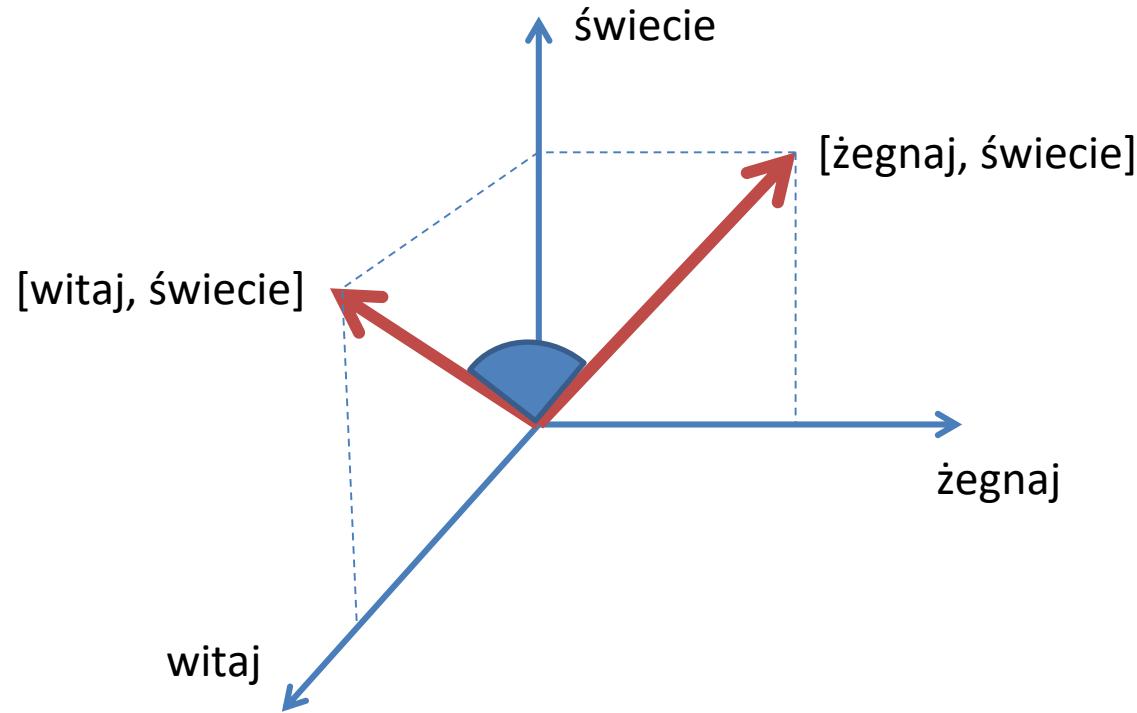
Polska i Węgry nie wyrażają zgody na mechanizm uzależnienia wypłaty unijnych funduszy od przestrzegania zasad praworządności. To komplikuje kwestię przyjęcia budżetu UE oraz Funduszu Odbudowy.



	bot	budżet	czytelny	człowiek
D0	0	0	0	1
D1	0	0	0	1
D2	1	0	1	0
D3	0	1	0	0

Term Document Matrix (TDM)

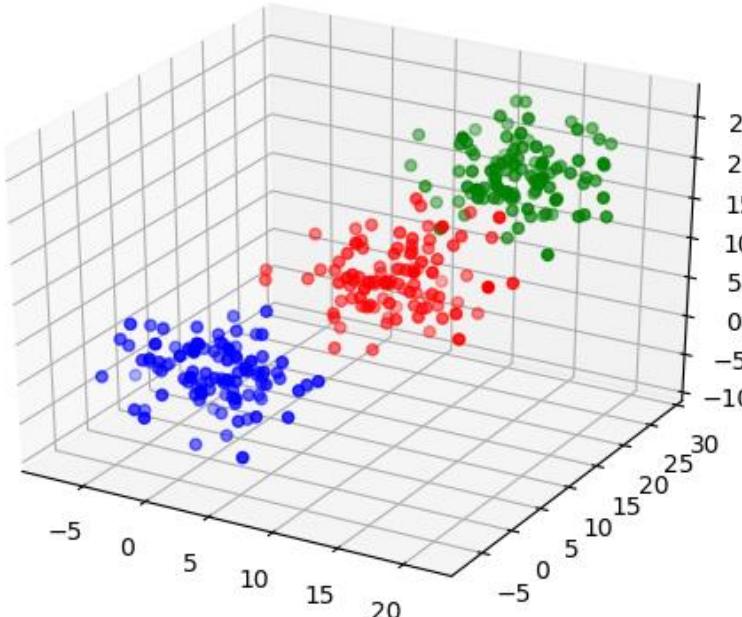
Czyli?



$$\text{Podobieństwo} = \cos([\text{witaj}, \text{świecie}], [\text{żegnaj}, \text{świecie}])$$

Aaaaaa... 😊

- Mając słownik (najlepiej stemmowany/ lematyzowany), możemy każdy dokument zapisać/ zakodować w formie wektorowej.
- Tak zbudowane wektory możemy do siebie porównywać jak zwykłe matematyczne obiekty (na przykład grupować).



Jest coś równie prostego co można jeszcze wykorzystać?

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(A|b_1, b_2) = \frac{P(b_1|A)P(b_2|A)P(A)}{P(b_1)P(b_2)}$$

- Reguła Bayes'a (prawdopodobieństwo warunkowe) – naiwny klasyfikator Bayes'a (Naive Bayes).

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	N
Rainy	Hot	High	True	N
Overcast	Hot	High	False	Y
Sunny	Mild	High	False	Y
Sunny	Cool	Normal	False	Y
Sunny	Cool	Normal	True	N
Overcast	Cool	Normal	True	Y
Rainy	Mild	High	False	N
Rainy	Cool	Normal	False	Y
Sunny	Mild	Normal	False	Y
Rainy	Mild	Normal	True	???
Overcast	Mild	High	True	Y
Overcast	Hot	Normal	False	Y
Sunny	Mild	High	True	N

Troszkę matematyki

Outlook	Y	N	P(Y)	P(N)
Sunny	3	2	3/8	2/5
Overcast	4	0	4/8	0/5
Rainy	1	3	1/8	3/5
Suma	8	5	1	1

Humidity	Y	N	P(Y)	P(N)
High	3	4	3/8	4/5
Normal	5	1	5/8	1/5
Suma	8	5	1	1

Temperature	Y	N	P(Y)	P(N)
Hot	2	2	2/8	2/5
Mild	3	2	3/8	2/5
Cold	3	1	3/8	1/5
Suma	8	5	1	1

Windy	Y	N	P(Y)	P(N)
True	2	3	2/8	3/5
False	6	2	6/8	2/5
Suma	8	5	1	1

Golf?	#	P
Y	8	8/13
N	5	5/13
Suma	13	1

Nie polubicie mnie za to ☺

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Mild	Normal	True	???

$$P(\text{Rainy}|Y) = \frac{1}{8} \quad P(\text{Mild}|Y) = \frac{3}{8} \quad P(\text{Normal}|Y) = \frac{5}{8} \quad P(\text{True}|Y) = \frac{2}{8}$$

$$P(\text{Rainy}|N) = \frac{3}{5} \quad P(\text{Mild}|N) = \frac{2}{5} \quad P(\text{Normal}|N) = \frac{1}{5} \quad P(\text{True}|N) = \frac{3}{5}$$

$$P(Y) = \frac{8}{13} \quad P(N) = \frac{5}{13}$$



$$P(Y|\text{Rainy, Mild, Normal, True}) \cong P(\text{Rainy}|Y)P(\text{Mild}|Y)P(\text{Normal}|Y)P(\text{True}|Y)P(Y)$$

$$P(N|\text{Rainy, Mild, Normal, True}) \cong P(\text{Rainy}|N)P(\text{Mild}|N)P(\text{Normal}|N)P(\text{True}|N)P(N)$$

$$P(Y|\text{Rainy, Mild, Normal, True}) \cong \frac{1}{8} * \frac{3}{8} * \frac{5}{8} * \frac{2}{8} * \frac{8}{13} = 0,00451$$

$$P(N|\text{Rainy, Mild, Normal, True}) \cong \frac{3}{5} * \frac{2}{5} * \frac{1}{5} * \frac{3}{5} * \frac{5}{13} = 0,01108$$

Zasnąłłem, po co to wszystko?

- A co jeżeli mamy taki zbiór?

Usman polecam ★★★★★ 5/5 Wystawiono 5 miesięcy temu
Polecam ceneo.pl, ponieważ jest łatwy w użyciu i umożliwia porównanie dostępnych cen. Karta podarunkowa to niesamowity wybór.

Forma dostawy:
Wysyłka elektroniczna / email ★★★★★ 5/5
To było natychmiastowe przez e-mail.

Usman polecam ★★★★★ 5/5 Wystawiono 5 miesięcy temu
Polecam ceneo.pl, ponieważ jest łatwy w użyciu i umożliwia porównanie dostępnych cen. Karta podarunkowa to niesamowity wybór.

Forma dostawy:
Wysyłka elektroniczna / email ★★★★★ 5/5
To było natychmiastowe przez e-mail.

arek polecam ★★★★★ 5/5 Wystawiono 6 miesięcy temu
Dobra cena.

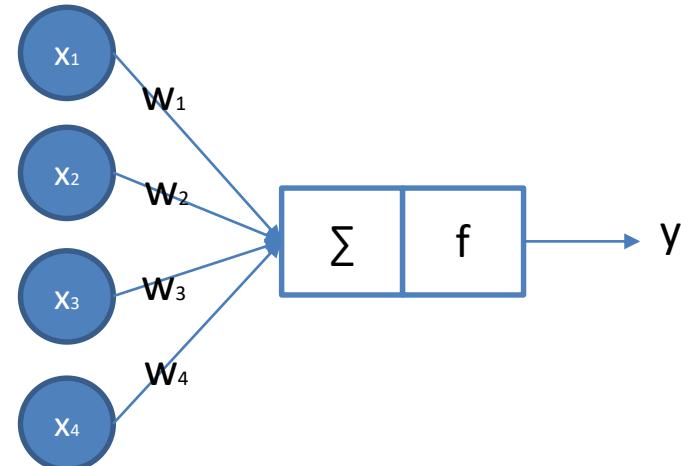
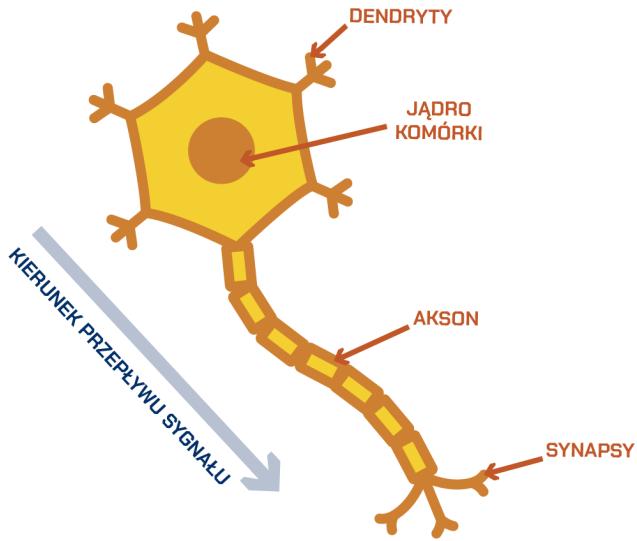
Szymon polecam ★★★★☆ 4/5 Wystawiono 6 miesięcy temu
Ok, ale słaby kontakt. Moje zgłoszenie zostało przyjęte->zweryfikowane->zamknięte, ale bez odpowiedzi, więc musiałem pisać drugi raz. Wtedy dopiero dostałem odpowiedź, że zwroty nie są przyjmowane dla kart podarunkowych, mimo że w e-mailu potwierdzającym zakup była informacja, że zwrot do 14 dni...

Forma dostawy:

Analiza sentymentu

- Posiadając ocenione komentarze możemy zbudować model oparty o klasyfikator Bayes'a, który oceni nam czy dana odpowiedź była pozytywna czy negatywna.
- Prawdopodobieństwo wystąpienia słowa w pozytywnym/negatywnym kontekście – który kontekst przeważa, taki jest sentyment.

Pan Neuron i Pan Wektor

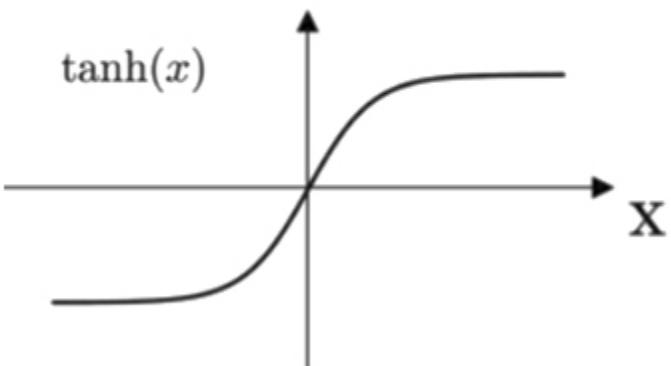


- Czyli możemy zapisać, że:

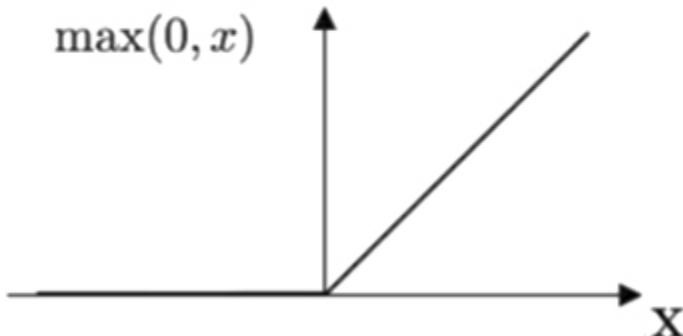
$$y = f(x_1 * w_1 + x_2 * w_2 + x_3 * w_3 + x_4 * w_4) \Leftrightarrow y = f\left(\sum_{i=1}^n x_i * w_i\right)$$

Funkcja aktywacji

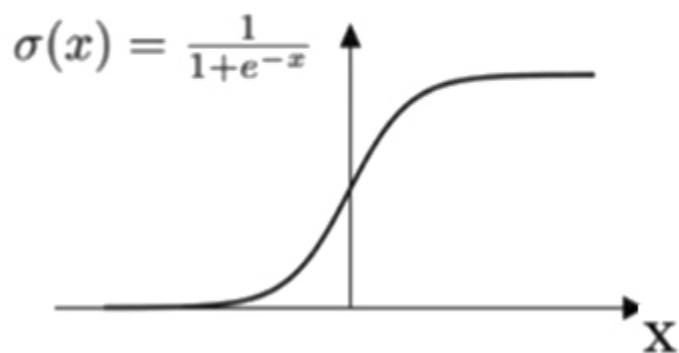
Tanh



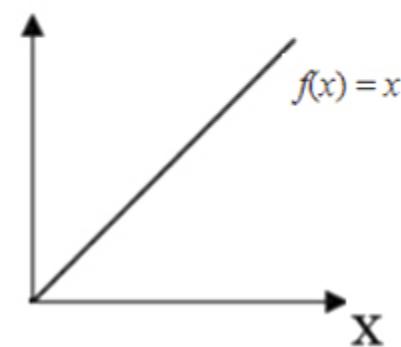
ReLU



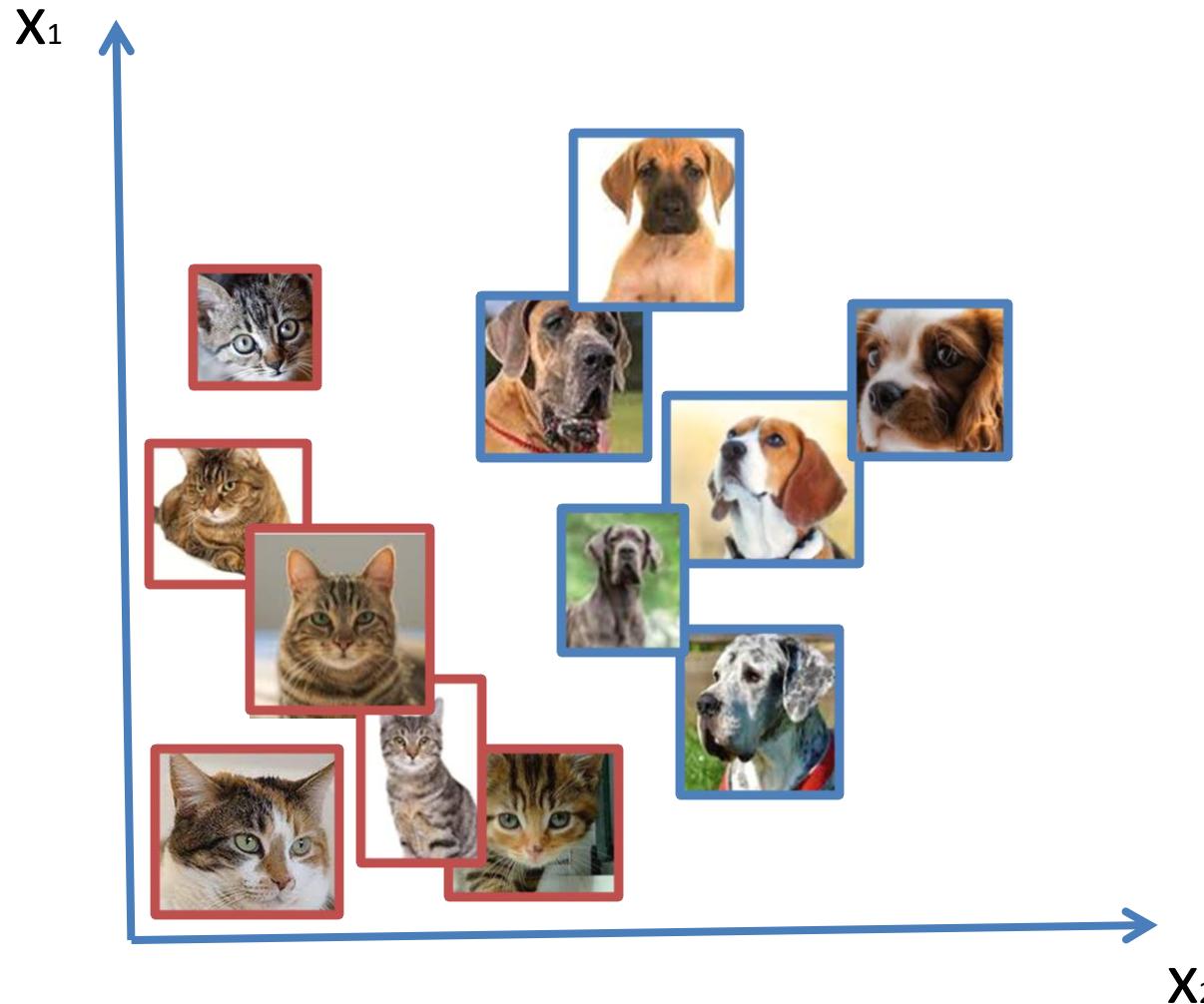
Sigmoid



Linear

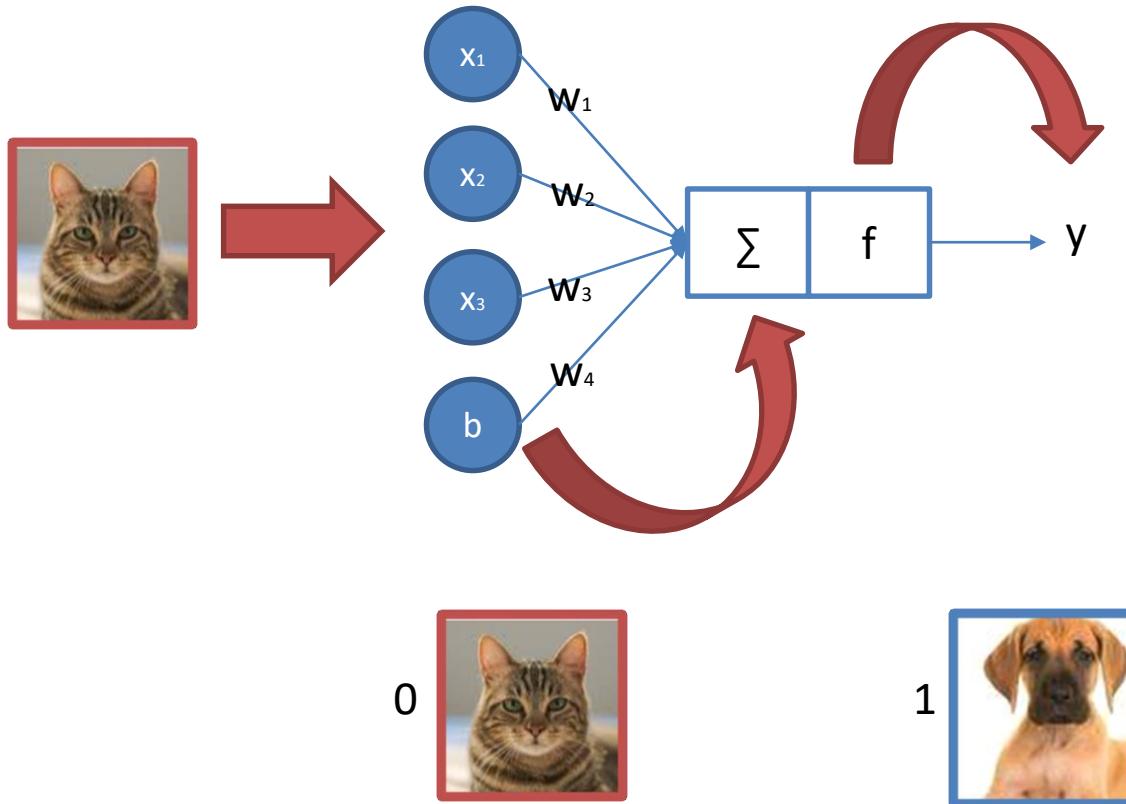


Pan Wektor i Pan Neuron idą do baru...



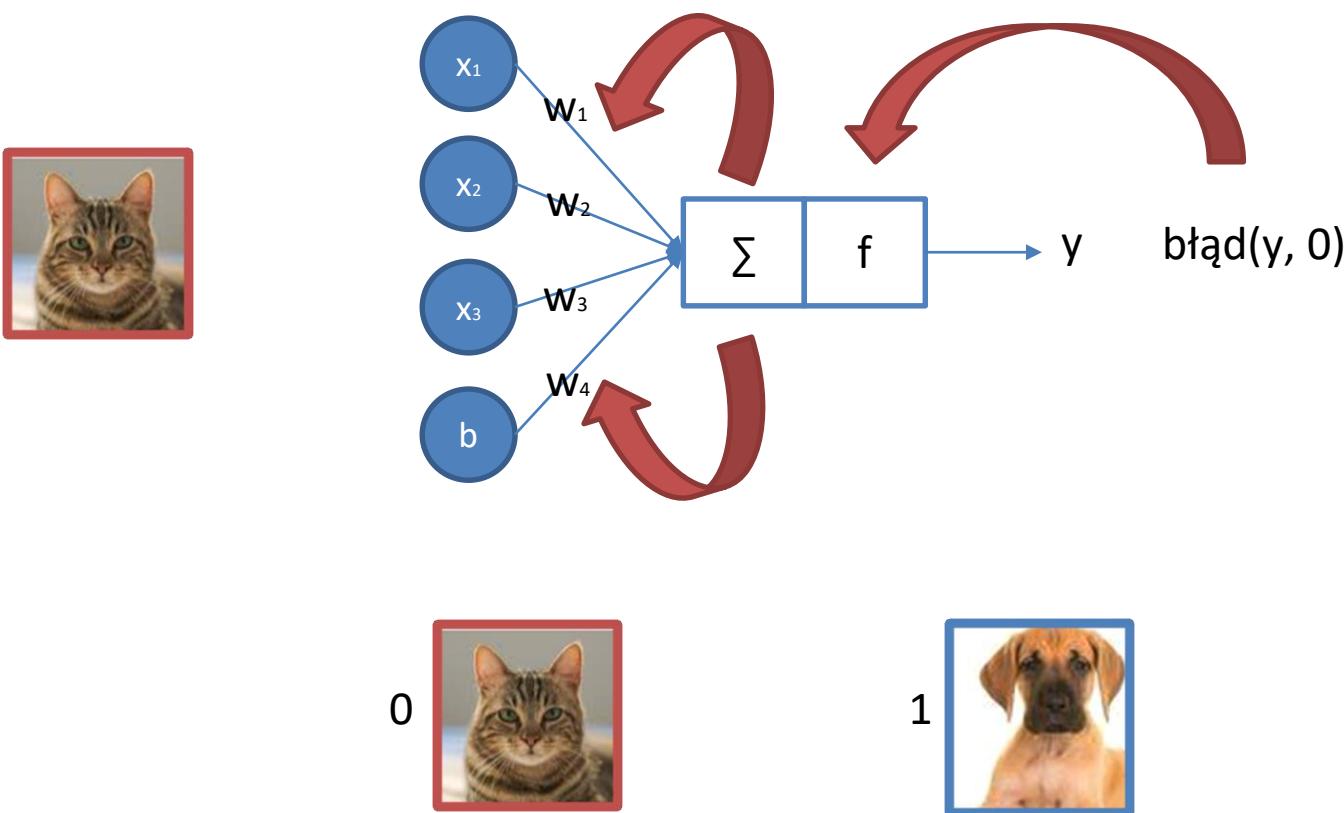
Nagle po drodze spotykają gradient...

- Liczymy wartość

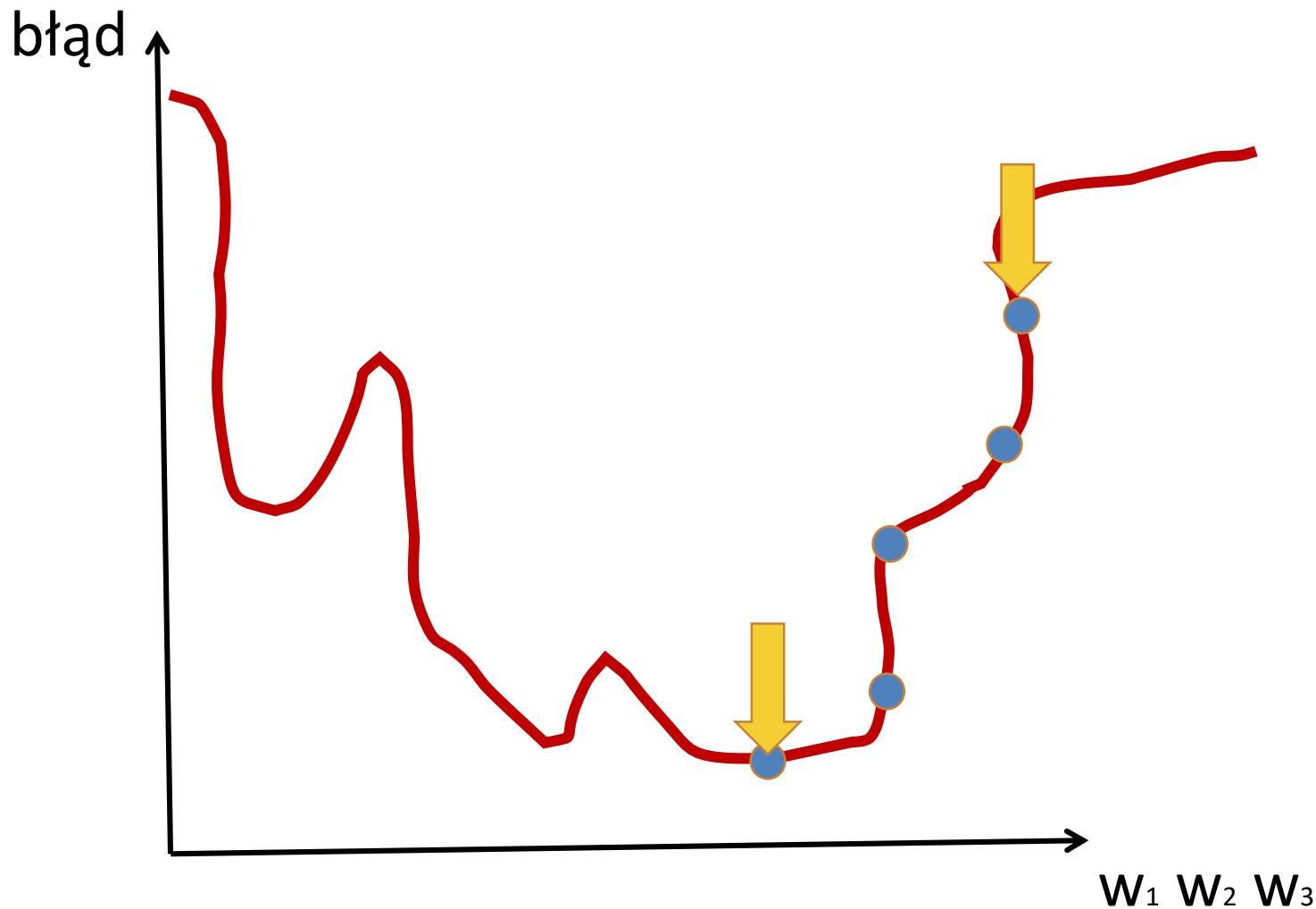


I gradient mówi – „bar jest tam na dole” ;)

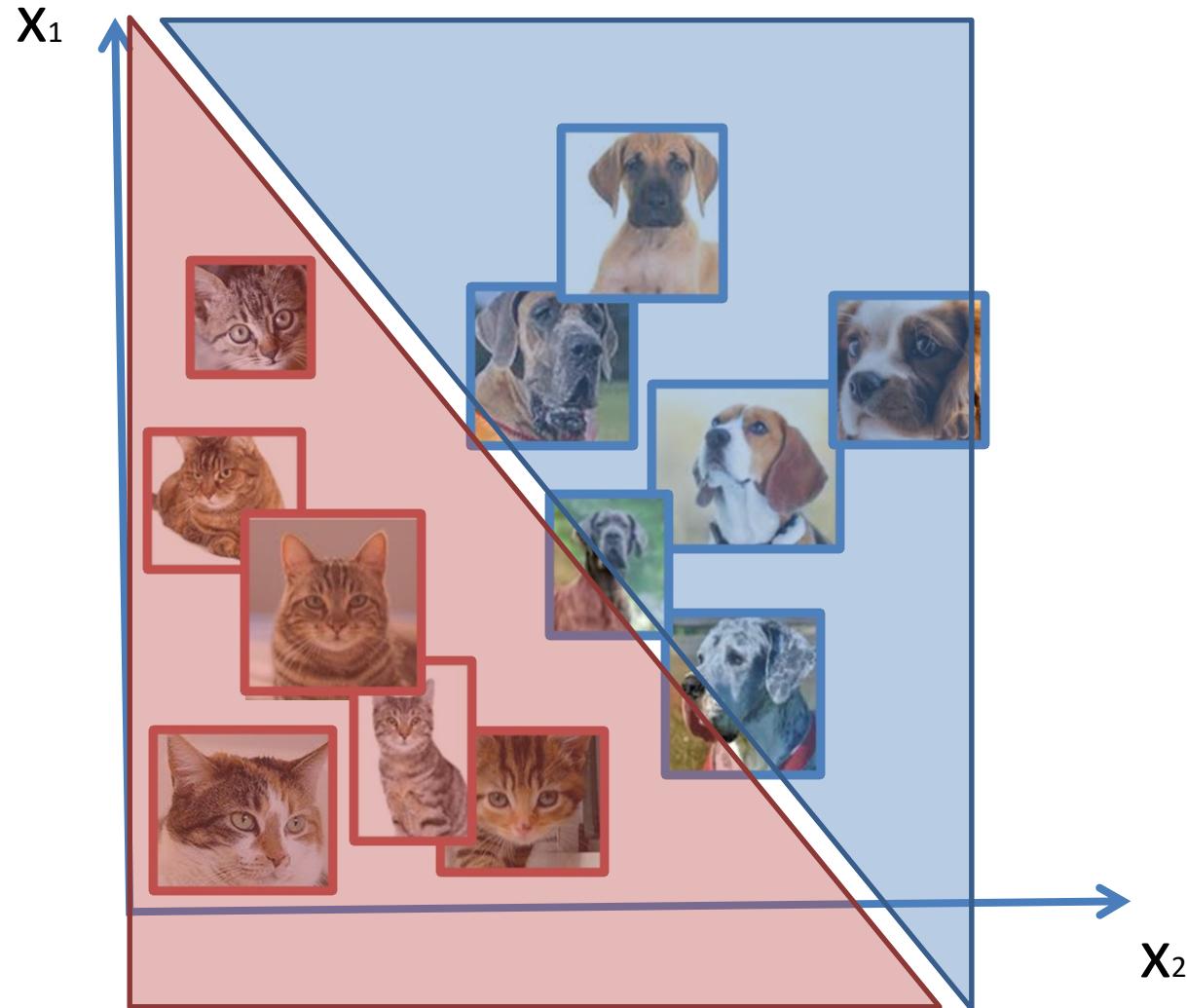
- Liczymy błąd (i na jego podstawie określamy, która waga jak się zmienia)



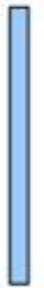
Wyjaśnienie sucharka z poprzednich slajdów



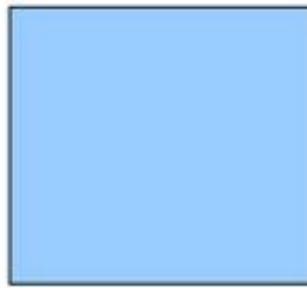
I to będzie nasz neuronowy klasyfikator 😊



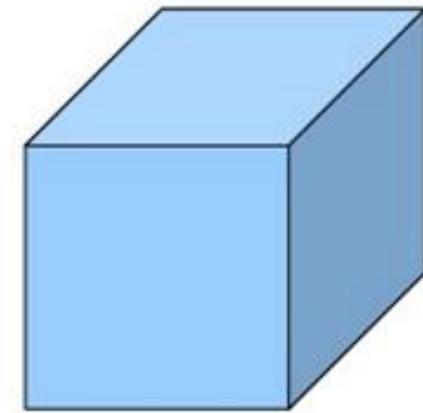
Wektory, macierze, tensory...



1d-tensor



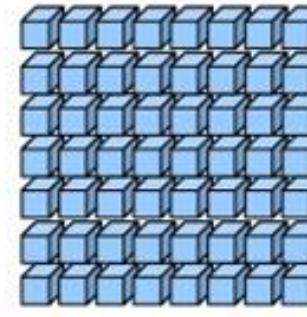
2d-tensor



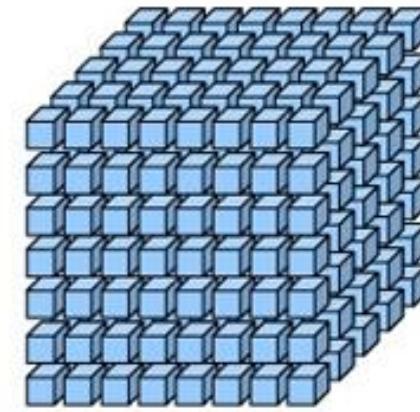
3d-tensor



4d-tensor

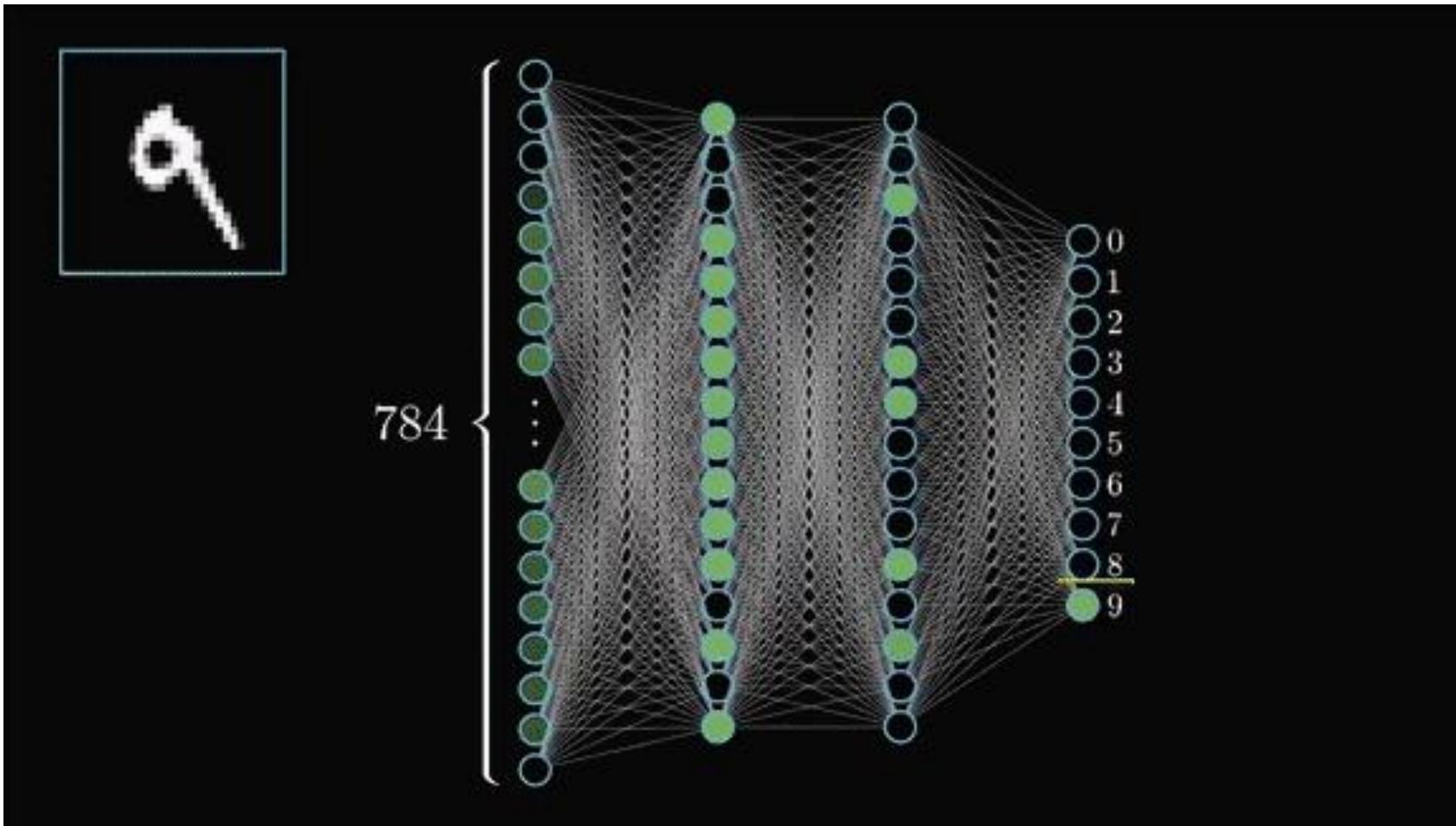


5d-tensor

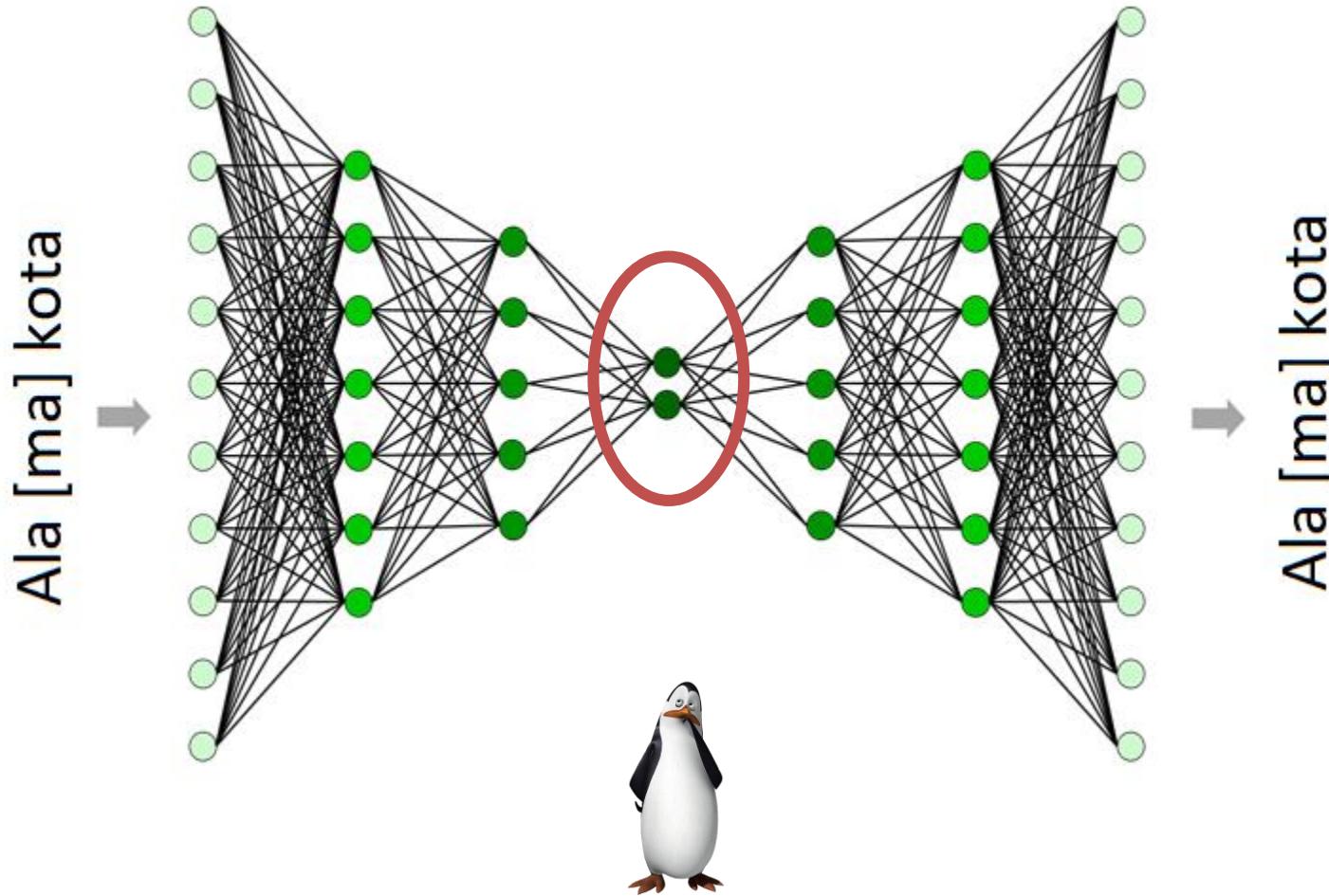


6d-tensor

Sieci wielowarstwowe

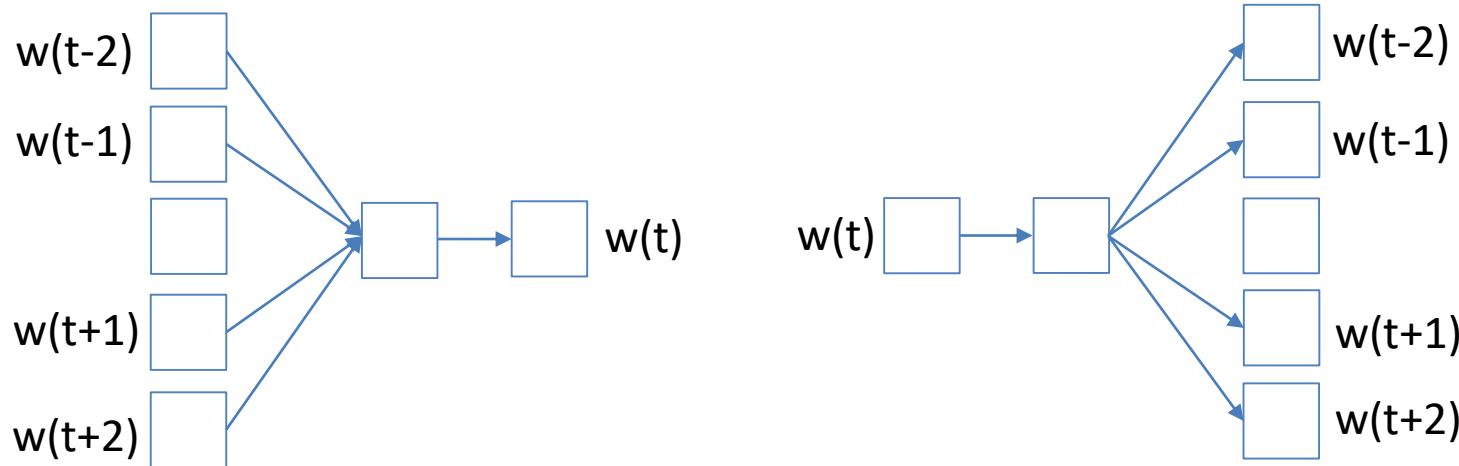


Autoencoder

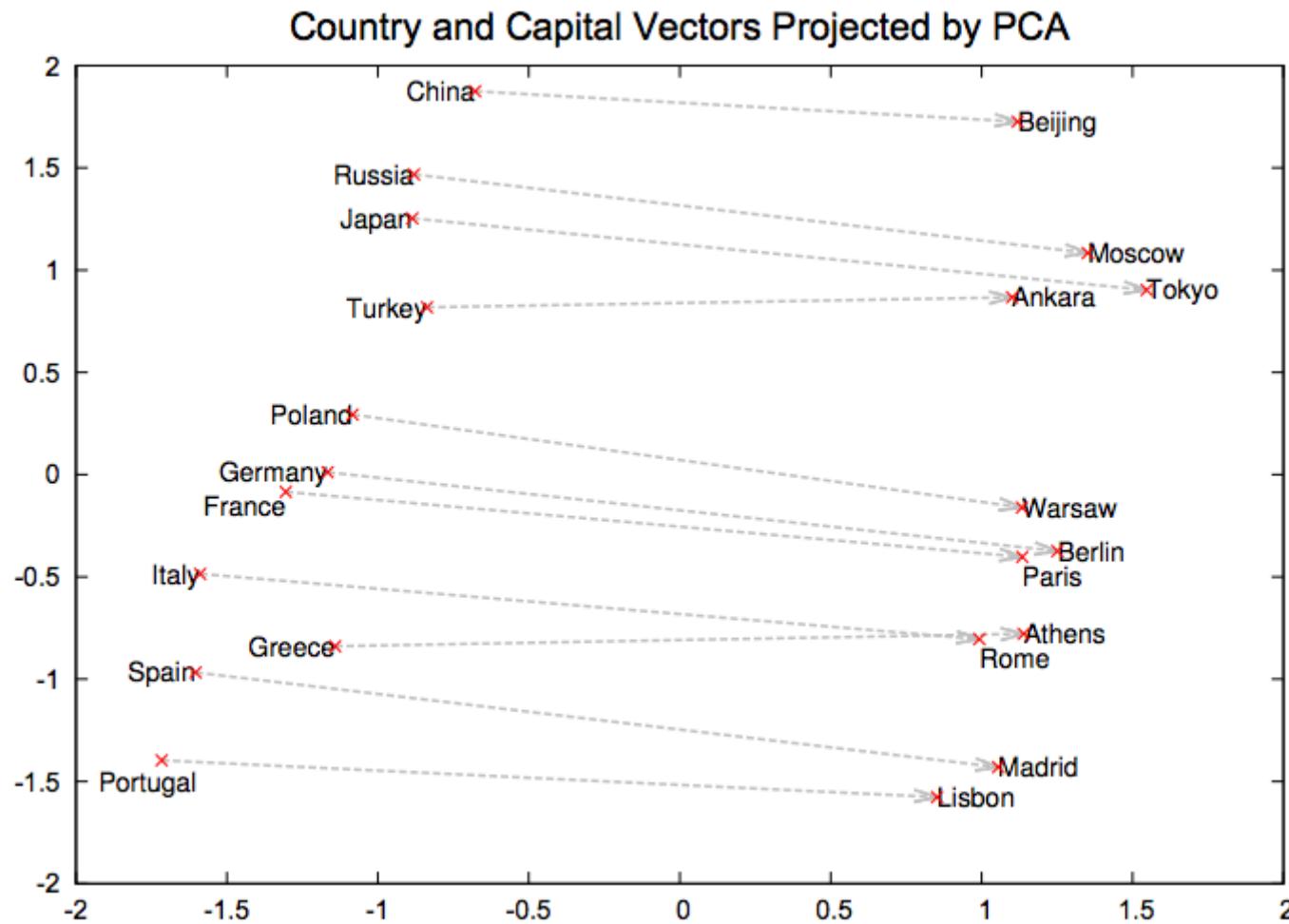


Word2Vec

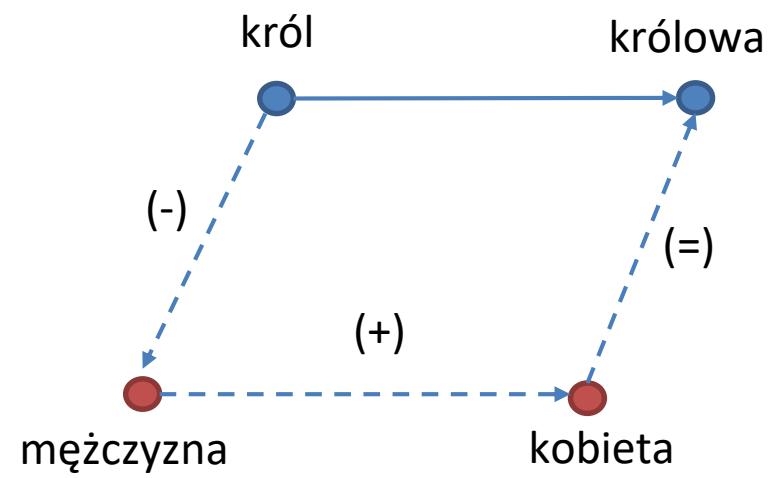
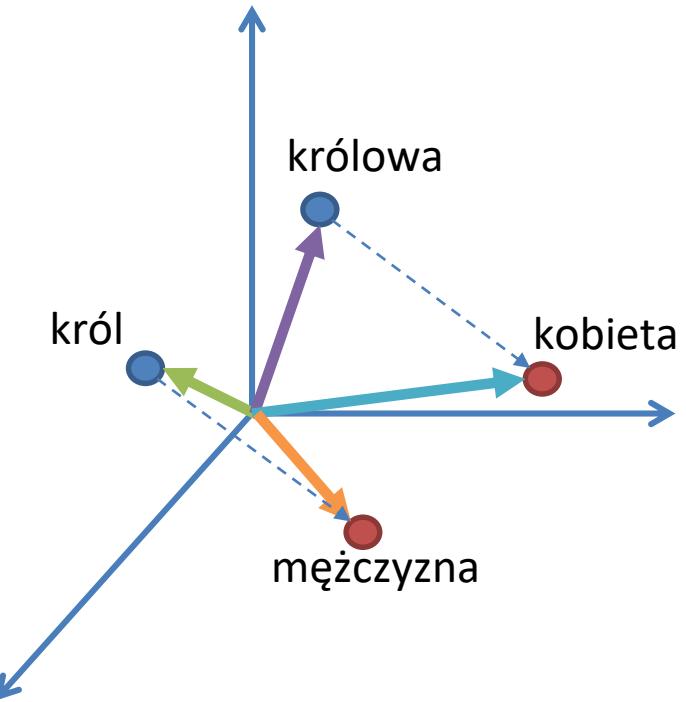
- CBOW (Continuous Bag of Words) – na podstawie okalających słów ustalamy to, którego brakuje.
- Skip-gram – na podstawie jednego słowa, ustalamy słowa okalające.



Word2Vec

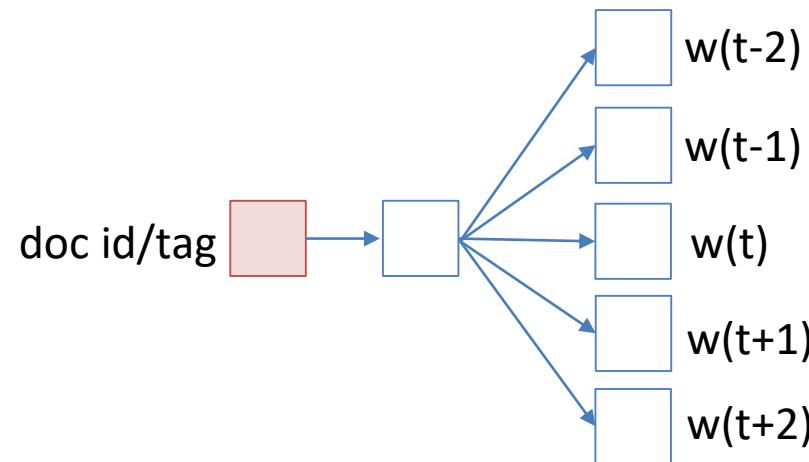
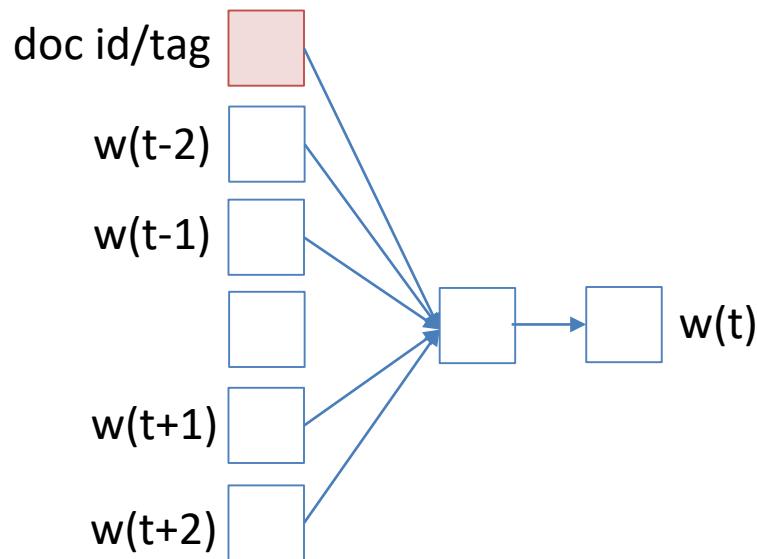


Word2Vec

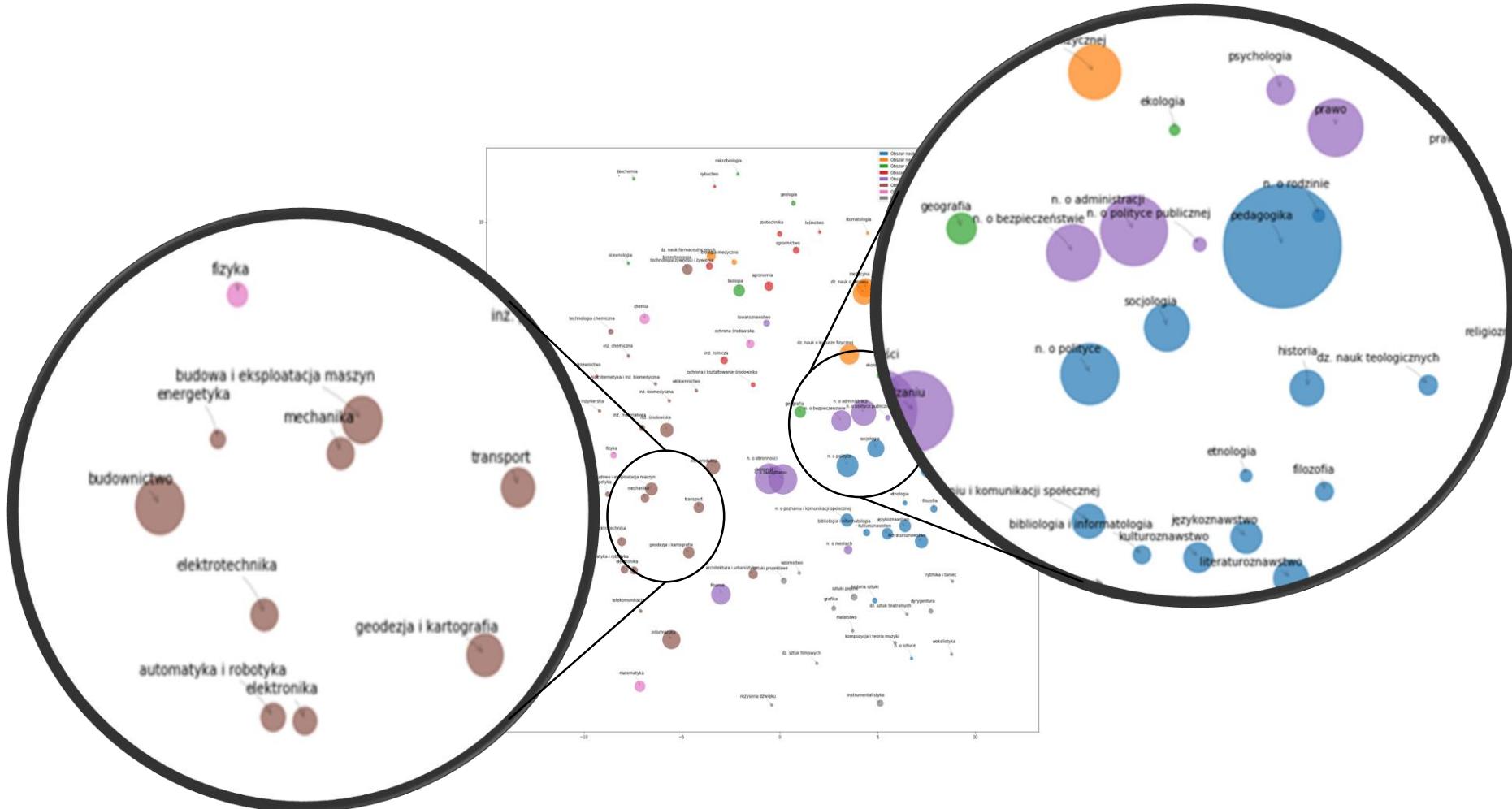


Doc2Vec

- PV-DM (Distributed Memory version of Paragraph Vector), dodajemy ID dokumentu.
- PV-DBOW (Distributed Bag of Words version of Paragraph Vector).

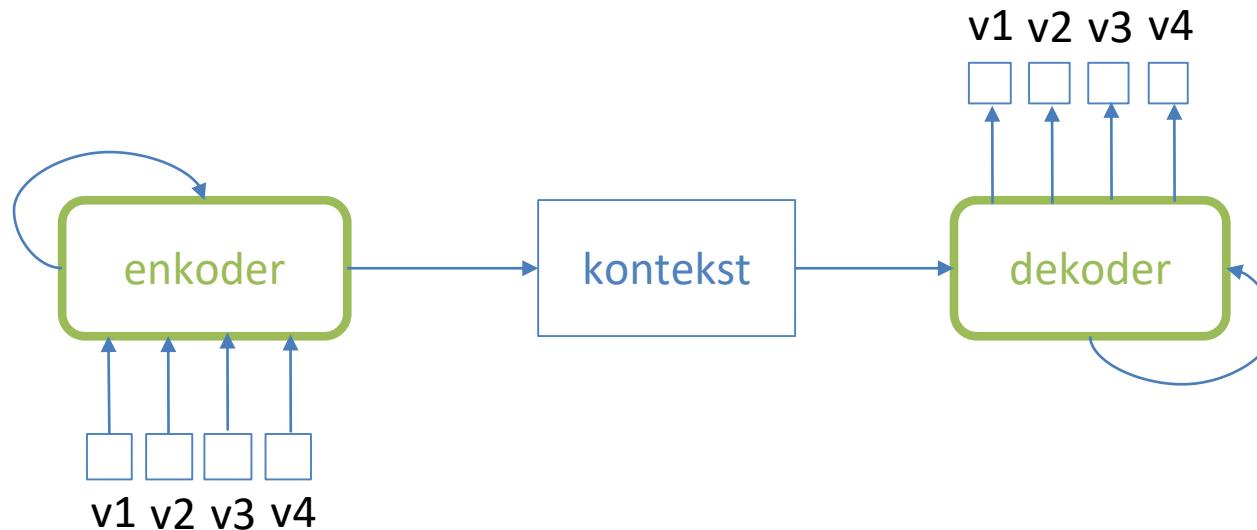


Doc2Vec

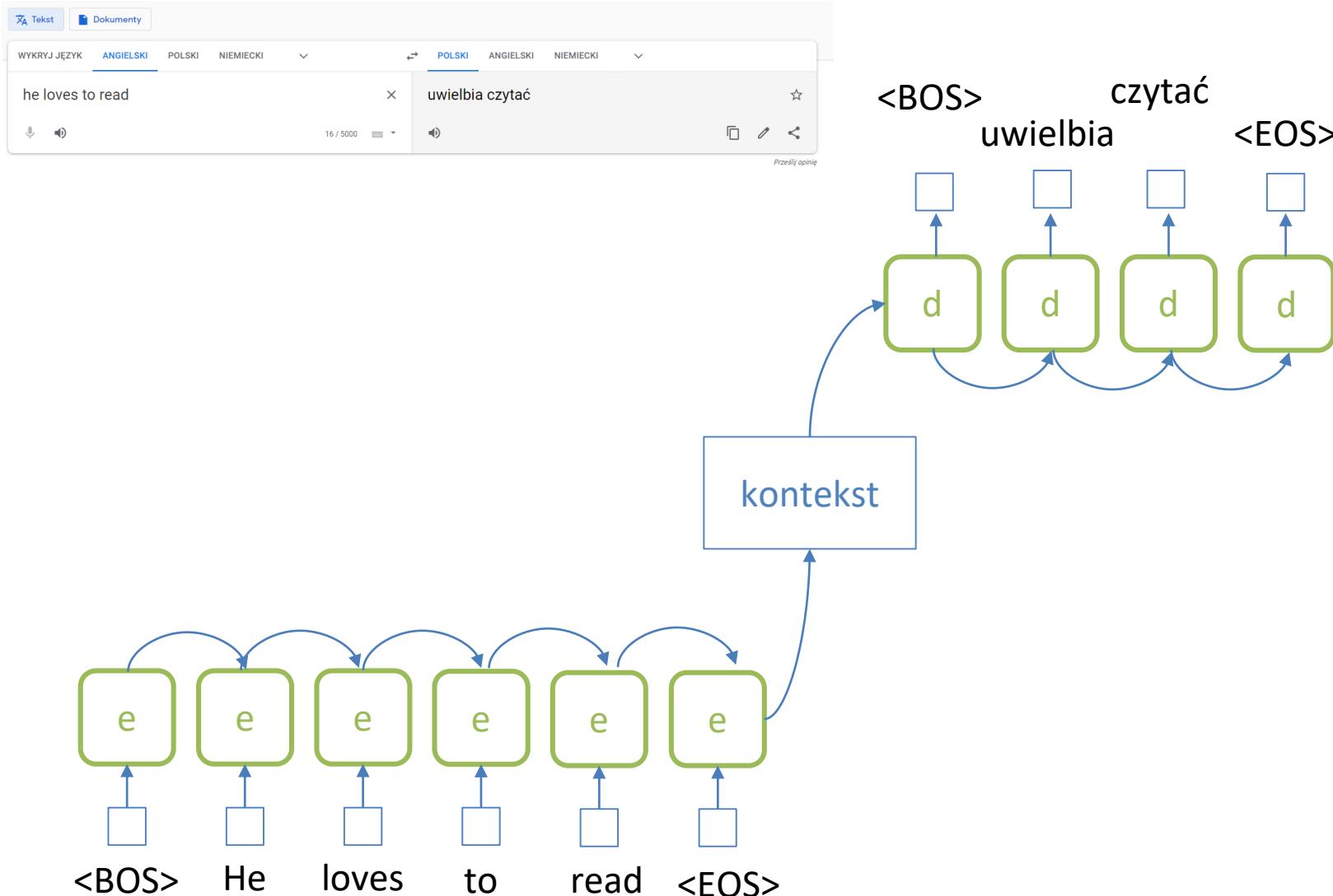


Seq2Seq

- Enkoder / Dekoder – rekurencyjne sieci neuronowe



Tłumaczenia maszynowe (i czatboty)



Czatboty 😊

- Uwaga!
- <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>





OŚRODEK
PRZETWARZANIA
INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY



Dziękujemy za uwagę

Ośrodek Przetwarzania Informacji
Państwowy Instytut Badawczy
al. Niepodległości 188 B
00-608 Warszawa

tel.: +48 22 570 14 00
faks: +48 22 825 33 19
e-mail: opi@opi.org.pl
www.opi.org.pl



OŚRODEK
PRZETWARZANIA
INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Dziękujemy za uwagę

Ośrodek Przetwarzania Informacji
Państwowy Instytut Badawczy
al. Niepodległości 188 B
00-608 Warszawa

tel.: +48 22 570 14 00
faks: +48 22 825 33 19
e-mail: opi@opi.org.pl
www.opi.org.pl