Inżynieria danych i modelowanie interdyscyplinarne

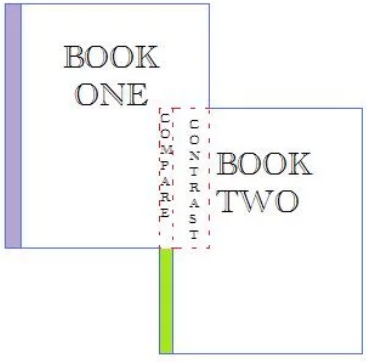
Wykład 10: text mining, podstawowe koncpecje, zastosowania

dr hab. inż. Julian Sienkiewicz, prof. ucz.

27 maja 2026

W jaki sposób możemy przedstawić dokument?

- w “codziennym życiu” zwykle nie mamy z tym większych problemów – używamy **opisu**, określając, mniej lub bardziej składnie, zawartość dokumentu,
- w ten sam sposób jesteśmy w stanie dokonać **porównania** dwóch dokumentów



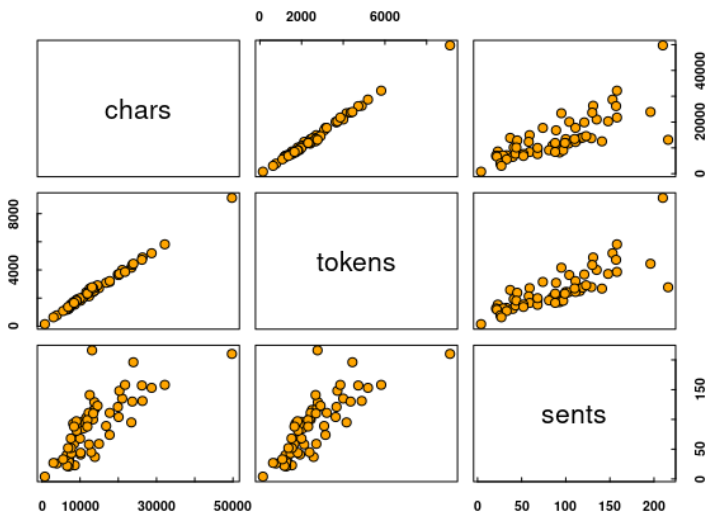
- w przypadku automatycznym, chcielibyśmy otrzymać jakąś określoną **reprezentację dokumentu**,
- dobrze byłoby, aby taka reprezentacja umożliwiała wyznaczanie **statystyk**, czyli była taka sama dla rożnych tekstów,
- potrzebne będzie również wykonywanie **porównań**

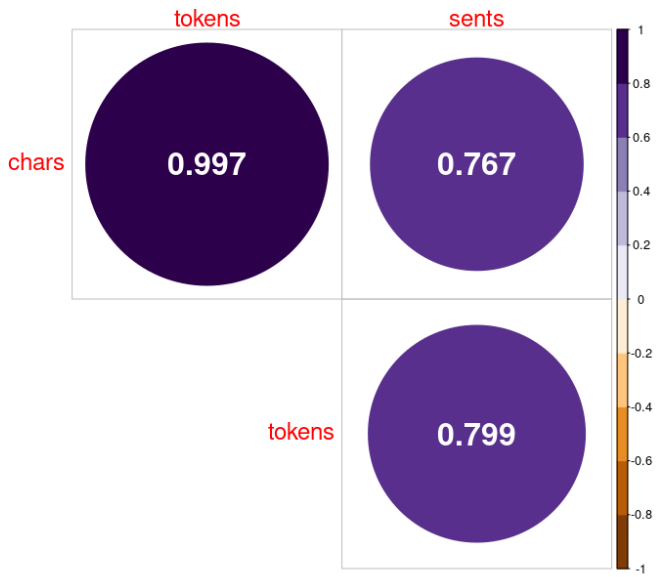
Miary długości

Najprostszymi do wyznaczenia i często też bardzo wygodnymi statystykami są:

- liczba **znaków** (liter i znaków interpunkcyjnych lub samych liter),
- liczba słów (często używa się określenia **token**),
- liczby **zdań**.

Te zmienne są zwykle mocno skorelowane ze sobą i ciężko uważać je za odrębne.





[Dane dotyczące mów inauguracyjnych prezydentów USA, pakiet quanteda.]

Miary czytelności

Innym rodzajem miary stosowanym do oceny tekstu jego jego **czytelność** (*readability*). W większości przypadków są to wskaźniki opierające się na udziale liczby słów, znaków oraz sylab, oceniające stopień skomplikowania tekstu. Popularne miary:

- Flesch-Kincaid Readability Score,

$$0.39 \frac{\text{liczba słów}}{\text{liczba zdań}} + 11.8 \frac{\text{liczba sylab}}{\text{liczba słów}} - 15.9$$

- Gunning's Fog Index,

$$0.4 \left(\frac{\text{liczba słów}}{\text{liczba zdań}} + 100 \frac{\text{liczba złożonych słów}}{\text{liczba słów}} \right)$$

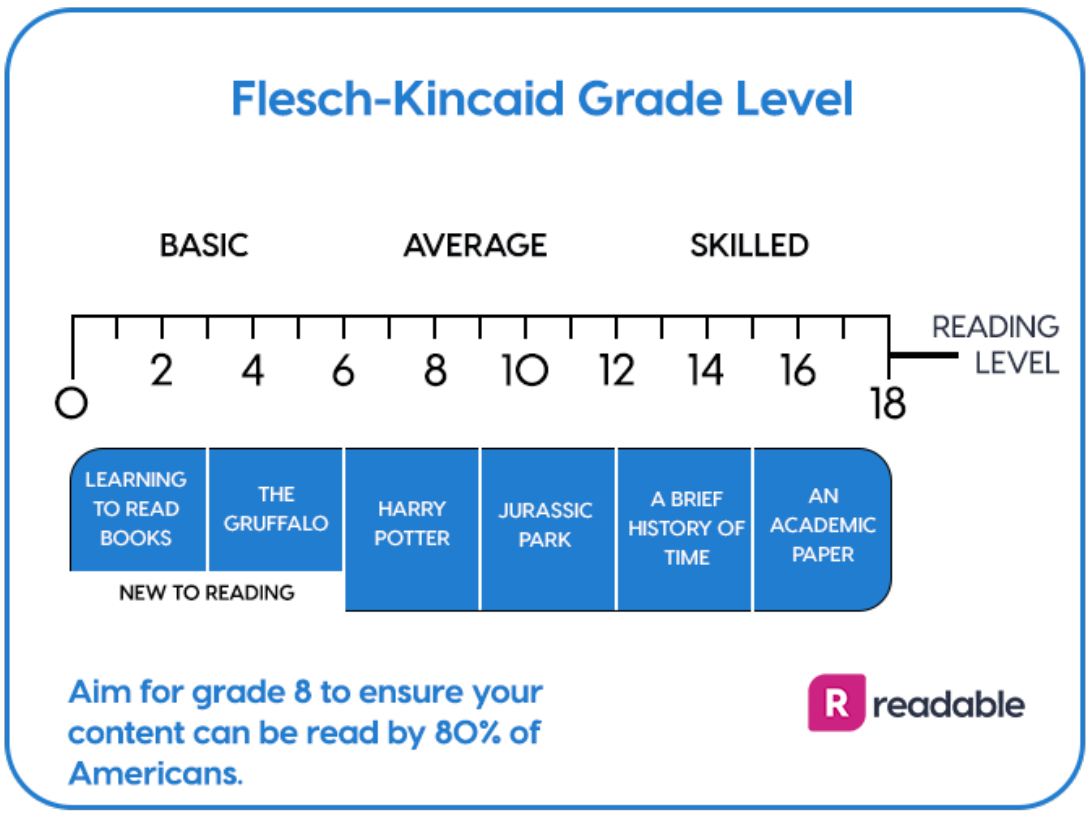
- Simple Measure of Gobbledygook (SMOG)

$$3 + \sqrt{\text{liczba złożonych słów} \frac{30}{\text{liczba zdań}}}$$

Złożone słowa to takie (w jęz. angielskim), które mają ponad dwie sylaby. Problemy: nie zawsze złożone słowa są trudne.

Interpretacja wskaźników czytelności

- liczba lat formalnej edukacji potrzebnej do rozumienia tekstu
- np. teksty dla szerokiej publiczności powinny mieć indeks Fog co najwyżej 12 – ocena to poziom maturzysty



[Grafika ze strony `https://readable.com/`]

Bag-of-words

- **bag-of-words** (BOW) jest chyba najprostszą reprezentacją tekstu,
- jak sama nazwa wskazuje, zakładamy w nim, że zawartośc dokumentu to po prostu poszczególne słowa, bez względu na ich kolejność pojawiania się w tekście,

Przykład

- John likes to watch movies,
- Mary likes movies too
- John also likes football,

John	likes	to	watch	movies	Mary	too	also	football
1	1	1	1	1	0	0	0	0
0	1	0	0	1	1	1	0	0
1	1	0	0	0	0	0	1	1

Innymi słowy tworzymy po prostu słownik słów, zaznaczając w nim ilość ich występowania.

The Bag of Words Representation

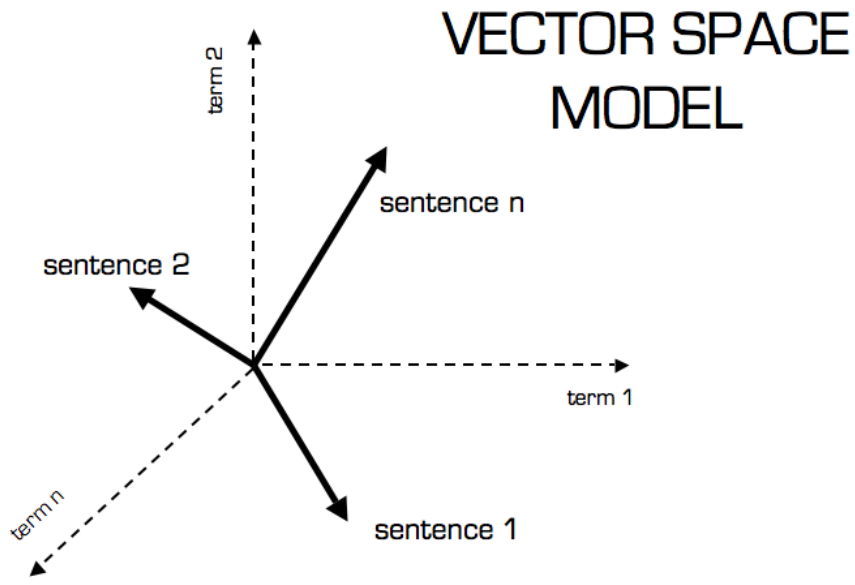
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

15

- oczywiście z takiej reprezentacji nie da się stworzyć sensownego tekstu (brak informacji o pozycji słów i gramatyce),
- BOW jest jednak przydatne do tworzenia **klasyfikatorów** opartych na czynnikach (features)

Model Przestrzeni Wektorowej

- Model Przestrzeni Wektorowej (Vector Space Model - VSM) jest pojęciem związanym z BOW, ale nie tożsamym z nim,
- odwołujemy się tu do pojęcia wektora

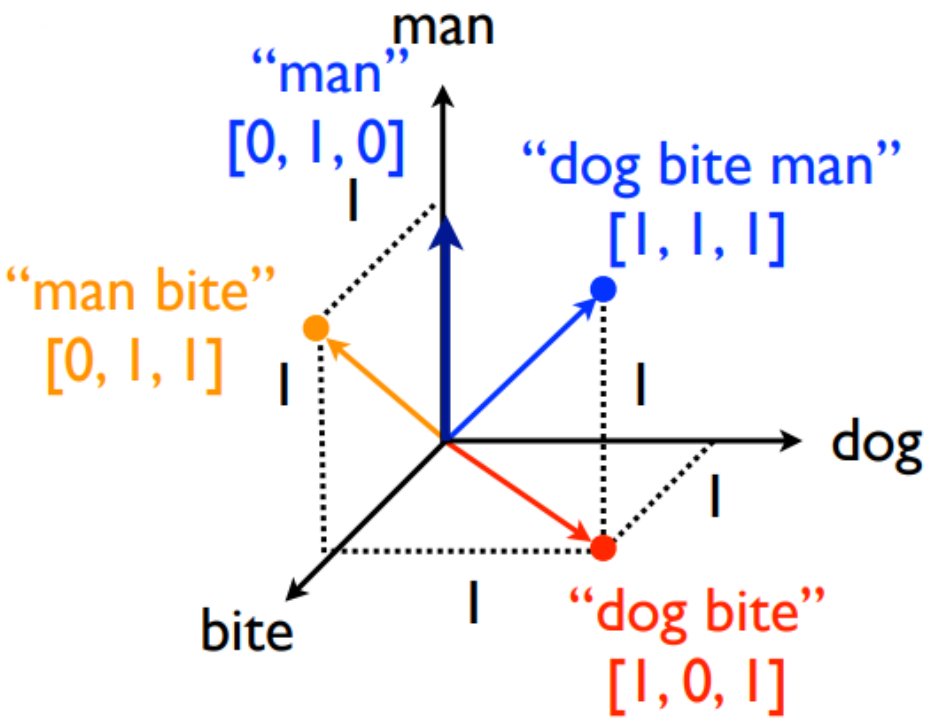


Model Przestrzeni Wektorowej

- zakładamy, że każde słowo to **oś** (czyli jest kierunkiem) w przestrzeni,
- w efekcie każdy obiekt (np. dokument albo słowo) może być reprezentowane przez wektor w tak skonstrukowanej przestrzeni

Prosty przykład

	dog	*man*	*bite*
doc_1	1	1	1
doc_2	1	0	1
doc_3	0	1	1
doc_4	0	1	0



Jak wyznaczyć podobieństwo dwóch dokumentów?

Najprościej za pomocą iloczynu skalarnego:

$$D(A, B) = \sum_{i=1}^{i=N} A_i B_i$$

gdzie N to liczba słów w słowniku, a A_i i B_i to informacja, czy słowo i wystąpiło, odpowiednio, w dokumecie A i B.

- pokaźnym problemem takiego podejścia jest to, że nie zakłada ono, iż dokumenty mogą mieć różną długość.
- w przypadku np. wyszukiwania, jest większa sznasa na to, że dłuższy dokuemnt będzie zawierał dane słowo

Odległość

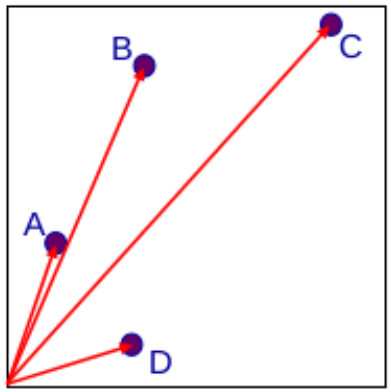
Innym sposobem jest wyznaczenie odległości (Euklidesowej) pomiędzy poszczególnymi dokumentami / zdaniami

$$D(A, B) = \sqrt{\sum_{i=1}^{i=N} (A_i - B_i)^2}$$

Podobieństwo Cosinusowe

Można również wyznaczyć tzw. podobieństwo cosinusowe

$$D(A, B) = \cos(A, B) = \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2} \sqrt{\sum_i B_i^2}}$$



- dokumenty “w tym samym kierunku” są do siebie podobne,
- w odróżnieniu od odległości ograniczona miara $D \in \langle 0; 1 \rangle$

Weźmy realny przykład

- Movie: Rocky (1976)
- Plot:

Rocky Balboa is a struggling boxer trying to make the big time. Working in a meat factory in Philadelphia for a pittance, he also earns extra cash as a debt collector. When heavyweight champion Apollo Creed visits Philadelphia, his managers want to set up an exhibition match between Creed and a struggling boxer, touting the fight as a chance for a "nobody" to become a "somebody". The match is supposed to be easily won by Creed, but someone forgot to tell Rocky, who sees this as his only shot at the big time. Rocky Balboa is a small-time boxer who lives in an apartment in Philadelphia, Pennsylvania, and his career has so far not gotten off the canvas. Rocky earns a living by collecting debts for a loan shark named Gazzo, but Gazzo doesn't think Rocky has the viciousness it takes to beat up deadbeats. Rocky still boxes every once in a while to keep his boxing skills sharp, and his ex-trainer, Mickey, believes he could've made it to the top if he was willing to work for it. Rocky, goes to a pet store that sells pet supplies, and this is where he meets a young woman named Adrian, who is extremely shy, with no ability to talk to men. Rocky befriends her. Adrain later surprised Rocky with a dog from the pet shop that Rocky had befriended. Adrian's brother Paulie, who works for a meat packing company, is thrilled that someone has become interested in Adrian, and Adrian spends Thanksgiving with Rocky. Later, they go to Rocky's apartment, where Adrian explains that she has never been in a man's apartment before. Rocky sets her mind at ease, and they become lovers. Current world heavyweight boxing champion Apollo Creed comes up with the idea of giving an unknown a shot at the title. Apollo checks out the Philadelphia boxing scene, and chooses Rocky. Fight promoter Jergens gets things in gear, and Rocky starts training with Mickey. After a lot of training, Rocky is ready for the match, and he wants to prove that he can go the distance with Apollo. The 'Italian Stallion', Rocky Balboa, is an aspiring boxer in downtown Philadelphia. His one chance to make a better life for himself is through his boxing and Adrian, a girl who works in the local pet store. Through a publicity stunt, Rocky is set up to fight Apollo Creed, the current heavyweight champion who is already set to win. But Rocky really needs to triumph, against all the odds...

Co nam da zliczanie słów:

rank	term	freq.	rank	term	freq.
1	a	22	16	creed	5
2	rocky	19	17	philadelphia	5
3	to	18	18	has	4
4	the	17	19	pet	4
5	is	11	20	boxing	4
6	and	10	21	up	4
7	in	10	22	an	4
8	for	7	23	boxer	4
9	his	7	24	s	3
10	he	6	25	balboa	3
11	adrian	6	26	it	3
12	with	6	27	heavyweigh	3
13	who	6	28	champion	3
14	that	5	29	fight	3
15	apollo	5	30	become	3

Ale wszystkie one są tak samo dla nas ważne?

rank	term	freq.	rank	term	freq.
1	a	22	16	creed	5
2	rocky	19	17	philadelphia	5
3	to	18	18	has	4
4	the	17	19	pet	4
5	is	11	20	boxing	4
6	and	10	21	up	4
7	in	10	22	an	4
8	for	7	23	boxer	4
9	his	7	24	s	3
10	he	6	25	balboa	3
11	adrian	6	26	it	3
12	with	6	27	heavyweigh	3
13	who	6	28	champion	3
14	that	5	29	fight	3
15	apollo	5	30	become	3

Potrzebujemy więc sposobu, aby jakoś “wyróżnić” te słowa, które faktycznie odnoszą się do istotnej treści.

Można oczywiście dokonać *stopword reduction*, czyli pozbyć się takic słów jak a, the, is... traktując je jako funkcyjne, ale to nie do końca jest rozwiązanie “systemowe”.

Można ocenić jak relatywnie ważne jest słowo, w odniesieniu do innych dokumentów.

$$idf_i = \log\left(\frac{M}{df_i}\right)$$

gdzie M to liczba dokumentów, które rozpatrujemy, a df_i to liczba dokumentów, w których występuje słowo i.

Daje to tzw. **odwrotną częstość w dokumentach** (IDF – inverse term frequency).

Daje to następujący efekt:

rank	term	idf	rank	term	idf
1	doesn	11.66	16	creed	6.84
2	adrain	10.96	17	paulie	6.82
3	viciousness	9.95	18	packing	6.81
4	deadbeats	9.86	19	boxes	6.75
5	touting	9.64	20	forgot	6.72
6	jergens	9.35	21	ease	6.53
7	gazzo	9.21	22	thanksgivin	6.52
8	pittance	9.05	23	earns	6.51
9	balboa	8.61	24	pennsylvani	6.50
10	heavyweigh	7.18	25	promoter	6.43
11	stallion	7.17	26	befriended	6.38
12	canvas	7.10	27	exhibition	6.31
13	ve	6.96	28	collecting	6.23
14	managers	6.88	29	philadelphia	6.19
15	apollo	6.84	30	gear	6.18

- czynnik IDF jest tym większy im dany wyraz **rzadziej** występuje w całym zbiorze dokumentów
- nie musi to jednak oznaczać, że jest bardzo istotny dla tego konkretnego dokumentu – może to być np. błąd lub jakieś mało popularne, wyszukane słowo
- potrzebny jest jeszcze czynnik, który odnosi się do częstości występowania słowa w danym tekście, czyli **częstość termów**

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

gdzie $n_{i,j}$ to liczba wystąpień wyrazu i w dokumencie j

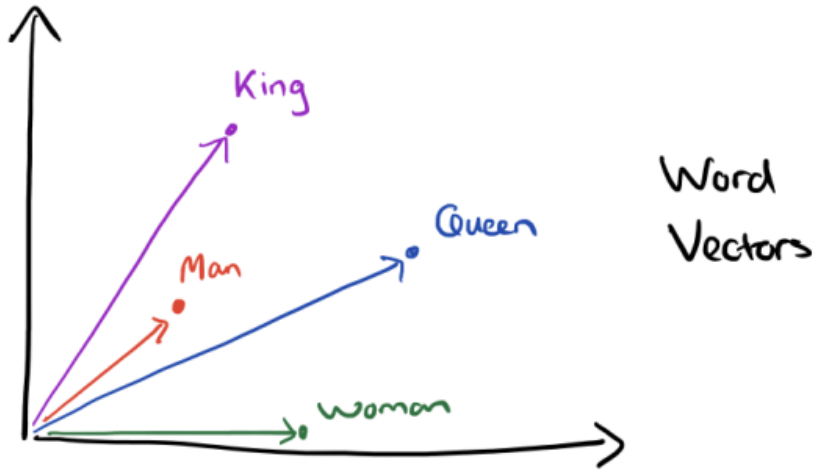
- razem daje to tzw. **TF-IDF**

$$(\mathrm{tf} - \mathrm{idf})_{i,j} = tf_{i,j} \times idf_i$$

I końcowo mamy:

rank	term	idf	rank	term	idf
1	rocky	96.72	16	meat	11.76
2	apollo	34.20	17	doesn	11.66
3	creed	34.18	18	adrain	10.96
4	philadelphia	30.95	19	fight	10.02
5	adrian	26.44	20	viciousness	9.95
6	balboa	25.83	21	deadbeats	9.86
7	boxing	22.37	22	touting	9.64
8	boxer	22.19	23	current	9.57
9	heavyweigh	21.54	24	jergens	9.35
10	pet	21.17	25	s	9.29
11	gazzo	18.43	26	struggling	9.21
12	champion	15.08	27	training	9.17
13	match	13.96	28	pittance	9.05
14	earns	13.01	29	become	8.96
15	apartment	11.82	30	mickey	8.96

Jednym z częstych zagadnień związanych z text mining jest budowa tzw. **word embedding**, czyli sposobu reprezentacji wyrazów w postaci liczbowej. W pewien sposób mówiliśmy już o tym, kiedy rozpatrywaliśmy **model przestrzeni wektorowej** - jest to jednak bardzo ograniczona reprezentacja.



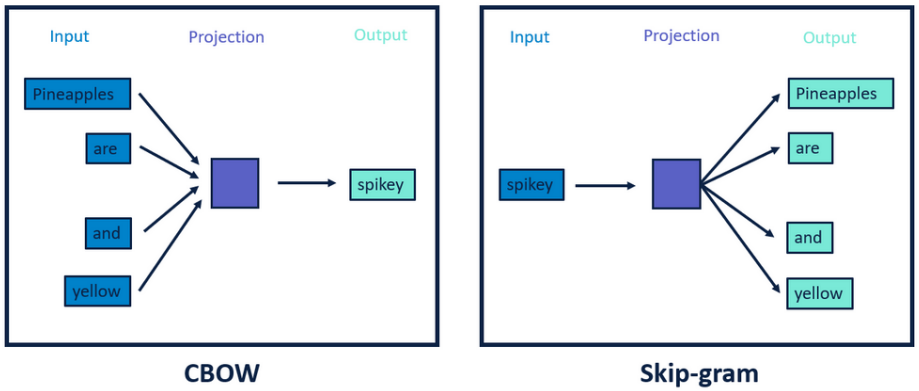
... and the *cute* **kitten** *purred* and then ...
... the *cute furry* **cat** *purred* and *miaowed* ...
... that the *small* **kitten** *miaowed* and she ...
... the *loud furry* **dog** *ran* and *bit* ...

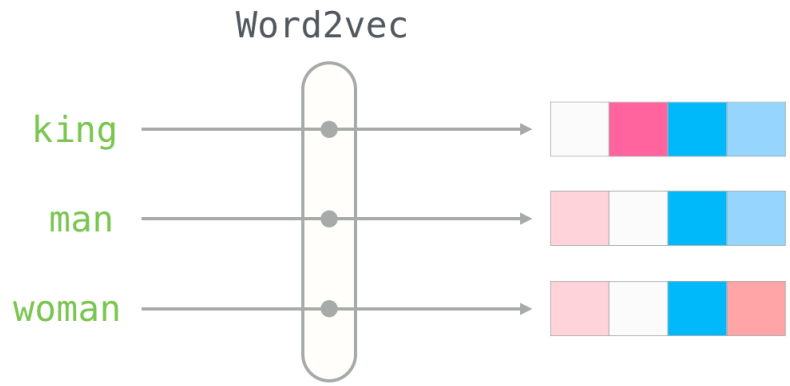
Example **basis vocabulary**: *{bit, cute, furry, loud, miaowed, purred, ran, small}*.

kitten context words: {cute, purred, small, miaowed}.
cat context words: {cute, furry, miaowed}.
dog context words: {loud, furry, ran, bit}.

Semantyka dystrybucyjna ma szerokie zastosowanie w rozwiązywaniu szeregu zadań związanych z przetwarzaniem języka naturalnego. U jej podstaw leży hipoteza, że słowa występujące w **podobnych kontekstach** w dużych zbiorach danych tekstowych mają **podobne znaczenie**.

Jednym z bardziej popularnych przykładów, należącym do tej grupy metod jest **word2vec**, stworzony przez Tomasa Mikolova z Google ok. 5 lat temu. Podejście to może wykorzystywać dwie różne techniki: **CBOW** - Continuous Bag of Words oraz **Skip-Gram**. Teoretycznie metody są algorytmicznie identyczne, oprócz tego, że CBOW przewiduje kluczowe słowo na podstawie kontekstu a Skip-Gram odwrotnie.





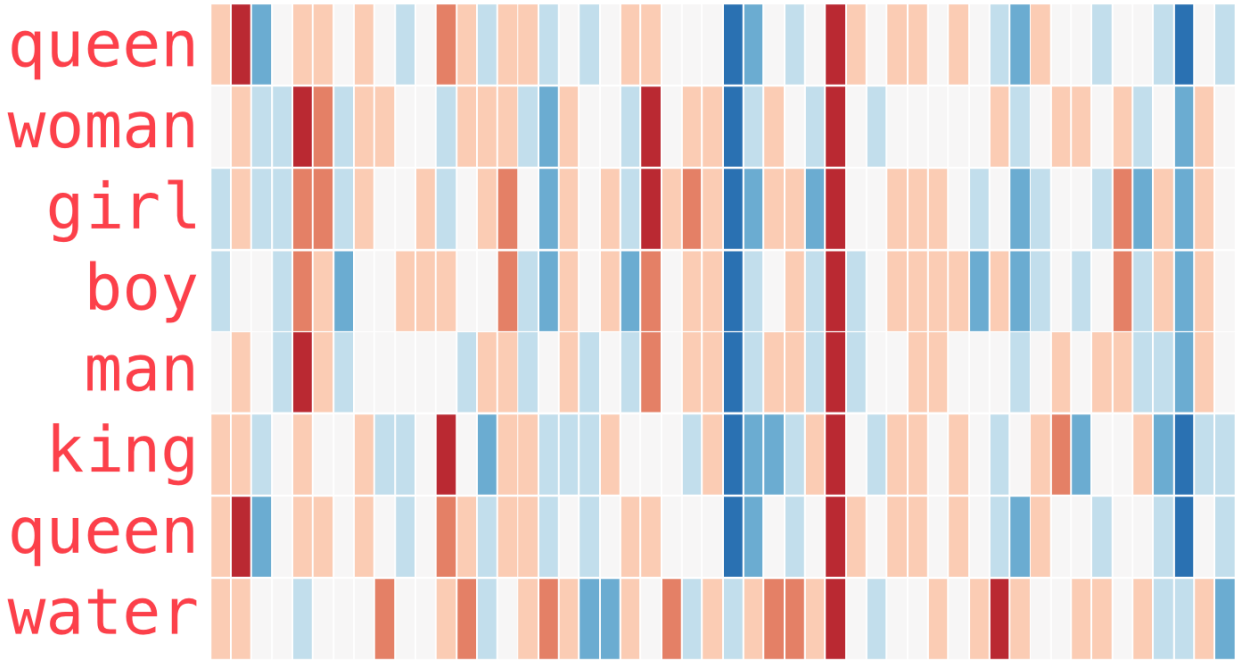
Word2vec używa sieci neuronowej z pojedynczą warstwą jako podstawowej architektury, ale jako ogólny wskaźnik używane jest po prostu prawdopodobieństwo $p(w_t|h)$ słowa w_t pod warunkiem historii (otoczenia) h, które jest wyznaczane przy użyciu metody największej wiarygodności.

GloVe, czyli **Global Vectors for Word Representation** jest datującą się na rok 2014 techniką zaproponowaną przez naukowców ze Stanford University i opierającą się na następującym algorytmie:

- wykonaj statystykę współwystępowania słów i zapisz ją w postaci macierzy $\mathbf{X}$; zwykle korpus jest skanowany w taki sposób, że szukamy słów kontekstowych w pewnym oknie zarówno przed jak i po interesującym nas słowie; zwykle też dalsze słowa wchodzą z mniejszą wagą np. $w = 1/przesuniecie$
- zdefiniuj więzy dla każdej pary słów występującej w macierzy $\mathbf{X}$ jako $\mathbf{w}_i^T \mathbf{w}_j + b_i + b = \log(\mathbf{X}_{ij})$, gdzie $\mathbf{w}_i$ to wektor głównego słowa, a $\mathbf{w}_j$ to wektor słowa z kontekstu, natomiast b_i, b to skalary
- zdefiniuj funkcję kosztu $J = \sum_i \sum_j (\mathbf{w}_i^T \mathbf{w}_j + b_i + b - \log(\mathbf{X}_{ij}))^2$

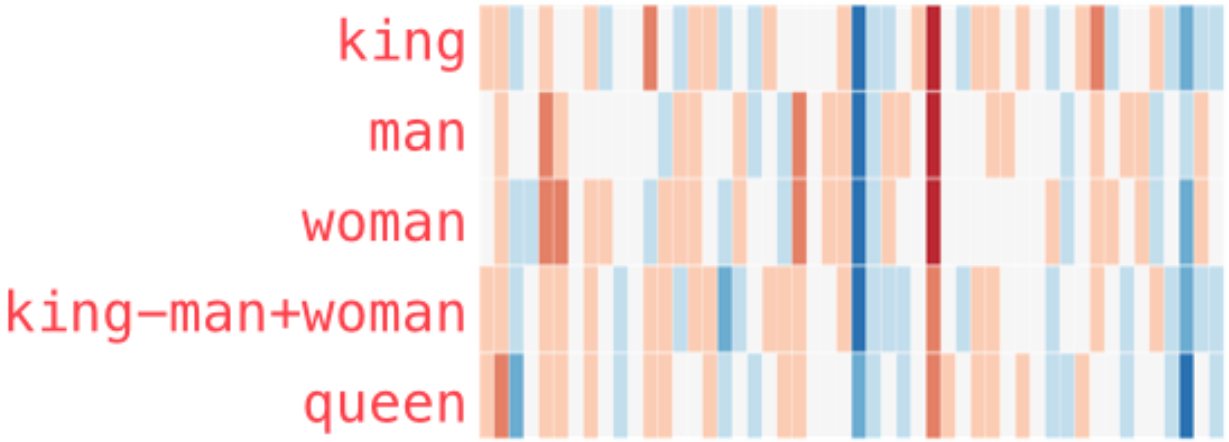
przy czym $f(\mathbf{X}_{ij})$ jest pewną funkcją ważącą, zapobiegającą uczeniu się jedynie od najczęściej wystepujących słów. Funkcja zostało zaproponowana jako $f(\mathbf{X}_{ij}) = (\mathbf{X}_{ij}/x_{max})^{\alpha}$ dla $\mathbf{X}_{ij} < x_{max}$ i $f(\mathbf{X}_{ij}) = 1$ w przeciwnym razie.

OK, ale co faktycznie z tego otrzymamy? Otóż dzięki powyższym zabiegom możemy uzyskać reprezentację słowa w pewnej przestrzeni (wymiary tej przestrzeni nie są związane z żadnymi konkretnymi słowami).

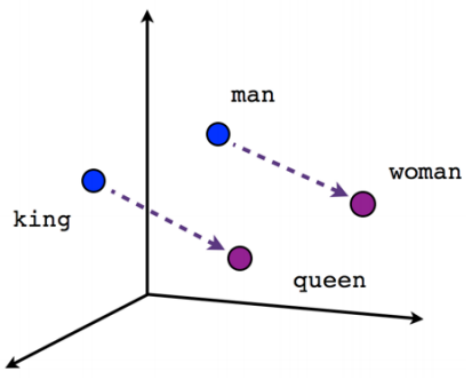


Co więcej, okazuje się, że w takiej przestrzeni dobrze działa wektorowa arytmetyka “semantyczna”.

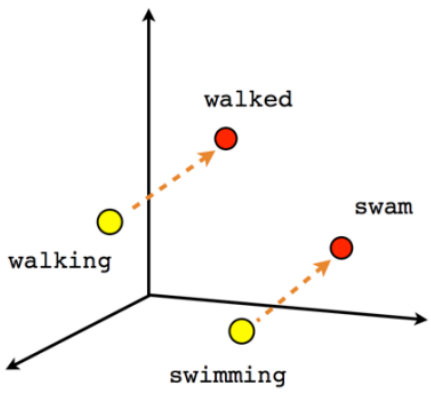
king − man + woman ~= queen



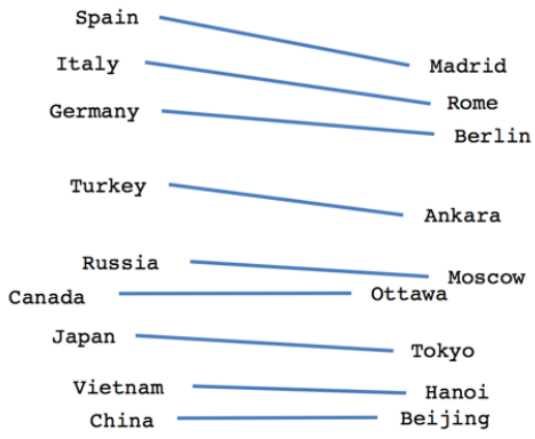
Jeśli natomiast dokonamy rzutowania za pomocą skalowania wielowymiarowego na przestrzeń dwuwymiarową, to można w łatwy sposób dostrzec analogie pomiędzy odpowiednimi parami słów (nie tylko rzeczowników).



Male-Female

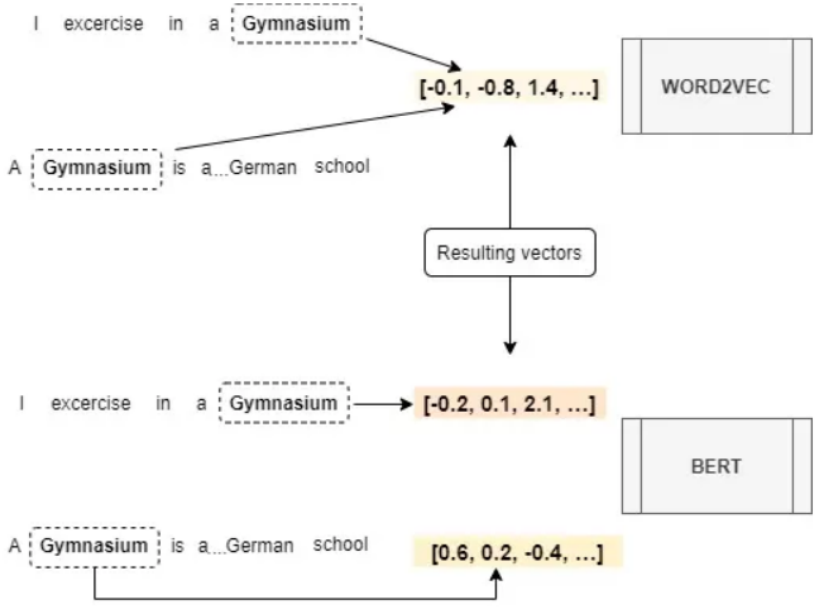


Verb tense



Country-Capital

BERT (Google, 2018) = Bidirectional Encoder Representations from Transformers

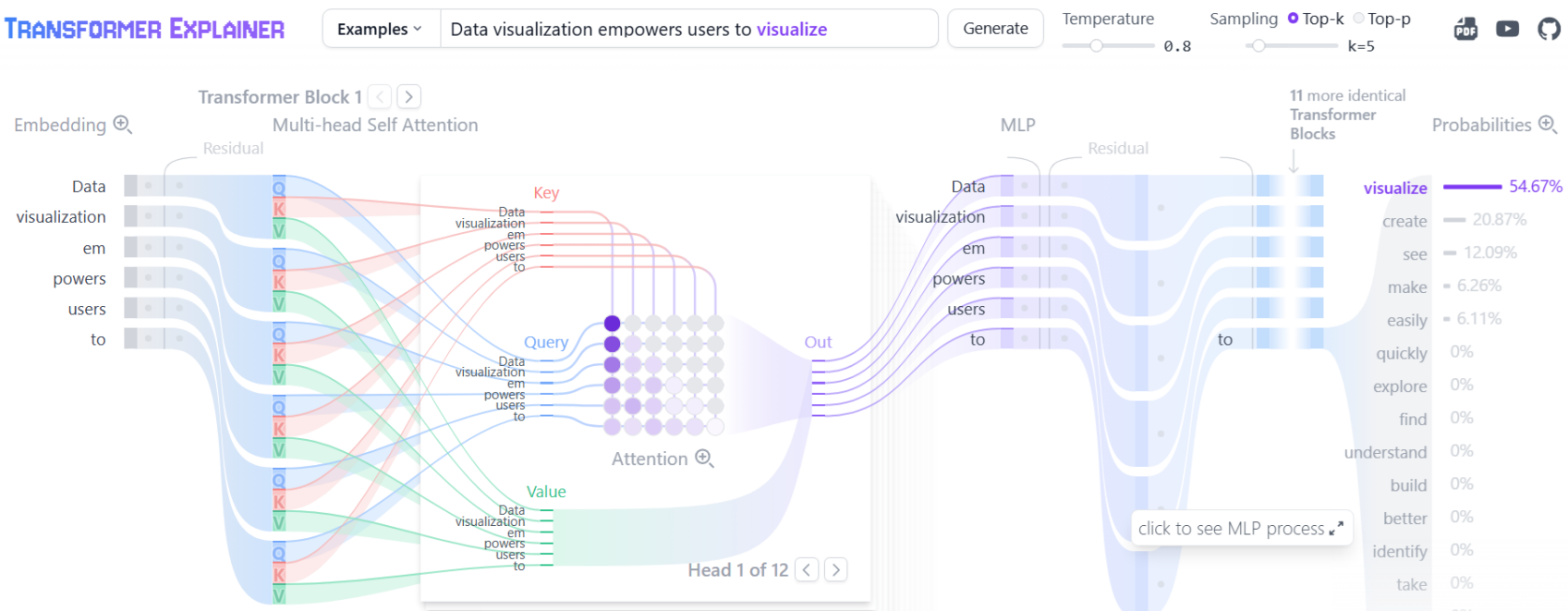


GPT (OpenAI, 2018) = Generative Pre-trained Transformer

Aspect	BERT	GPT
Type of Task	Mainly focused on understanding language (e.g., classification, NER)	Primarily focused on generating language (e.g., text generation, dialogue)
Training Objective	Masked Language Modeling (MLM)	Autoregressive Language Modeling (next word prediction)
Architecture	Encoder (bidirectional attention)	Decoder (unidirectional attention)
Applications	Text classification, question answering, NER	Text generation, summarization, chatbots
Context Understanding	Looks at both left and right context	Looks at left context only (sequential)

A **large language model (LLM)** is a language model trained with self-supervised machine learning on a vast amount of text, designed for natural language processing tasks, especially language generation.

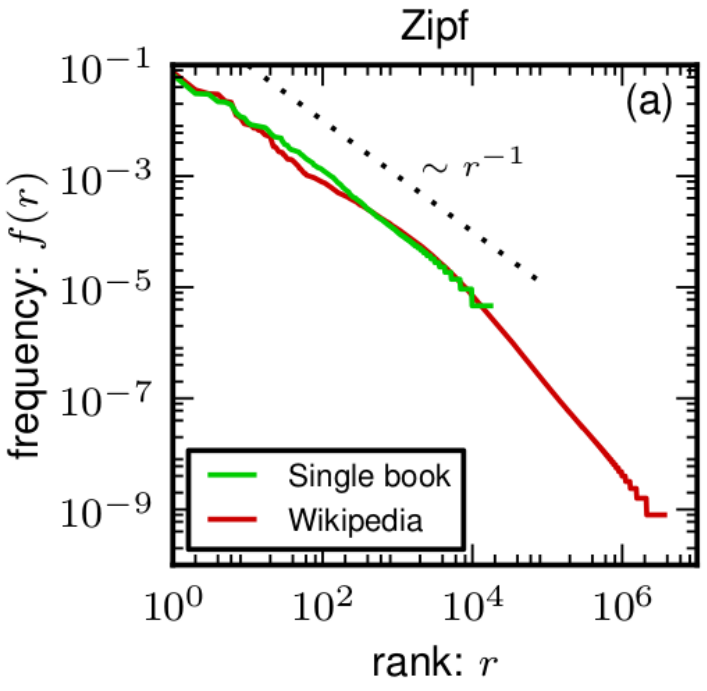
Jak działa GPT?



https://poloclub.github.io/transformer-explainer/

Prawo Zipfa

- historyczne sformułowanie:
  Jeśli słowa (typy) otrzymają rangę (ranking) zgodnie z częstotliwością wystepowania $r = 1, 2, ..., V$, to częstość $f(r)$ r-tego słowa skaluje się z rangą jak
  $$f(r) = \frac{f(1)}{r},$$
  gdzie $f(1)$ to częstość słowa najczęstszego.
- obecna postać:
  $$f(r) = \frac{\beta}{r^{\alpha}},$$
  gdzie β jest stałą normującą, a $\alpha \geq 1$,



Pojedyncza książka, tu Moby Dick Melville'a, do pobrania ze strony Projektu Gutenberg `http://www.gutenberg.org/`. Czerwony – angielska Wikipedia, pobrana z `http://dumps.wikimedia.org`

Przyswajanie nowych słów może być rozumiane jako proces przyswajania innowacji, a to z kolei jest często zapisywane jako tzw. model Bassa dany równaniem

$$\frac{d\rho(t)}{dt} = (a + b\rho(t))(1 - \rho(t)) \tag{1}$$

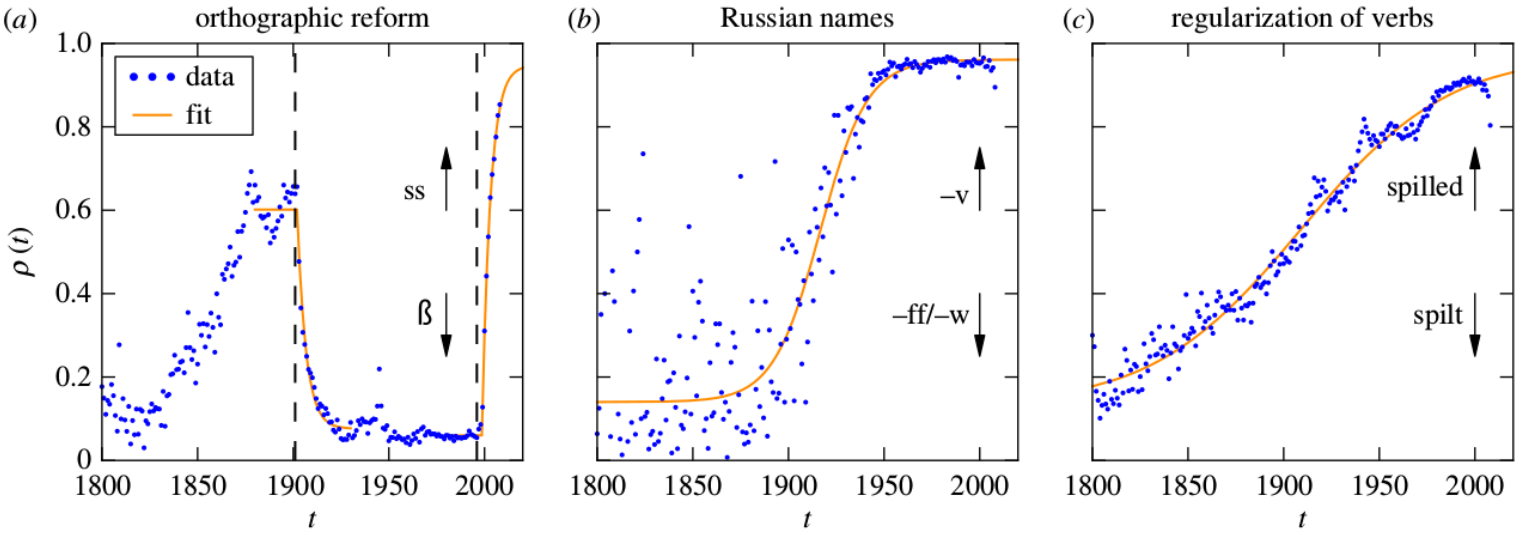
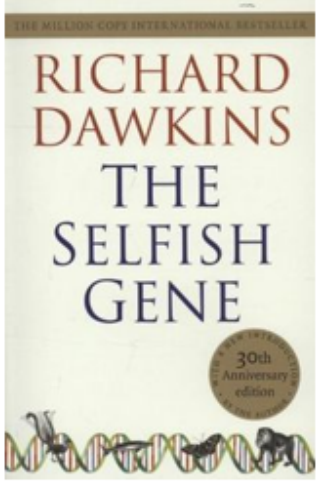


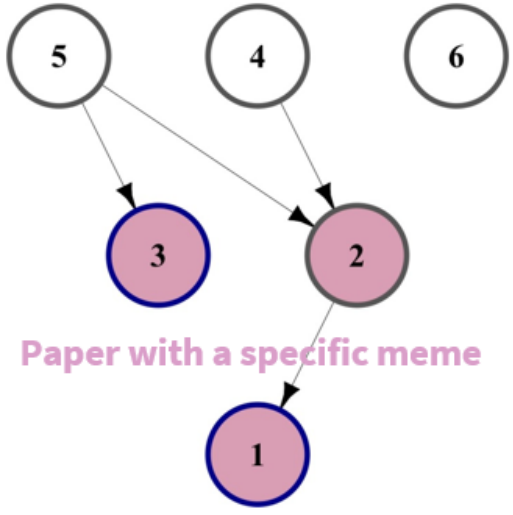
Figure 1. Examples of linguistic changes showing different adoption curves. We estimate the fraction of adopters $\rho(t)$ by the relative frequency as $\rho(t) = \sum_w n_1^w / \sum_w \sum_q n_q^w$, where n_q^w is the total number of occurrences (tokens) of variant q for the word w at year t. (*a*) The orthography of German words that changed to 'ss' ($q = 1$) from 'ß' ($q = 2$) in the orthographic reform of 1996 (many words changed from 'ss' to 'ß' in the 1901 reform). (*b*) The transliteration of Russian names ending with the letter 'в' when written in English (Latin alphabet), changed to an ending in 'v' ($q = 1$) from endings in 'ff' ($q = 2$) or 'w' ($q = 3$) (e.g. $w =$ 'Саратов' is nowadays almost unanimously written as 'Saratov', but it used to be written also as 'Saratoff' or 'Saratow'). (*c*) The past form of the verb spill changed to its regular form 'spilled' ($q = 1$) from the irregular form 'spilt' ($q = 2$). The light curve shows the fit of equation (1.2). The estimated parameters a and b are (*a*) $\hat{a} = 0.218$, $\hat{b} = 0.000$ in 1901, and $\hat{a} = 0.229$, $\hat{b} = 0.000$ in 1996; (*b*) $\hat{a} = 0.000$, $\hat{b} = 0.099$; and (*c*) $\hat{a} = 0.001$, $\hat{b} = 0.030$. The corpus is the Google-ngram [3,4] plotted in the minimum (yearly) resolution, see the electronic supplementary material, section I, for details on the data and section IIIB for details on the fit. (Online version in colour.)

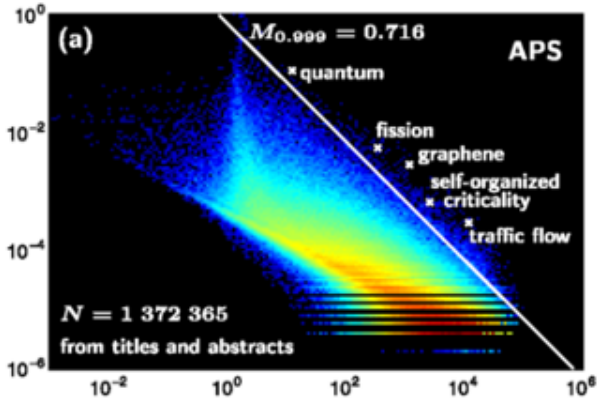
Analiza memetyczna

Selfish gene, Dawkins 1989

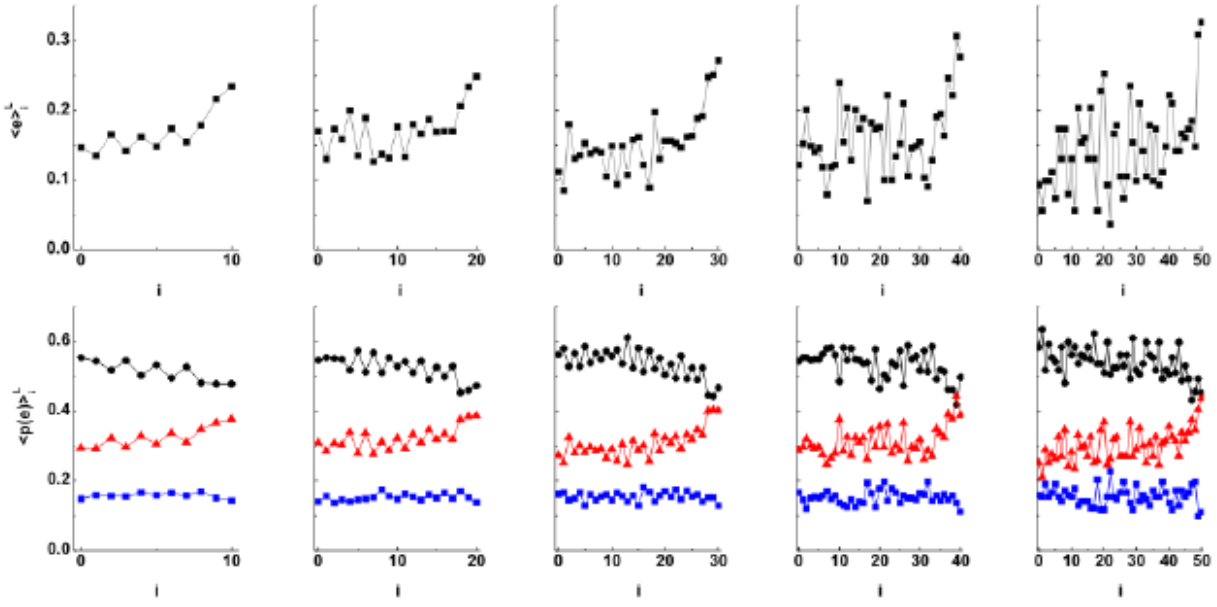
Cultural entities such as words, melodies, recipes, and ideas evolve similarly as genes, involving replication and mutation but using human culture instead of the gene pool as their medium of propagation.



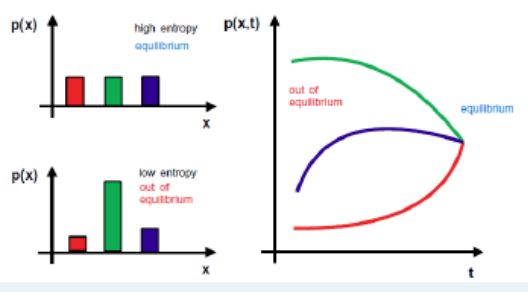




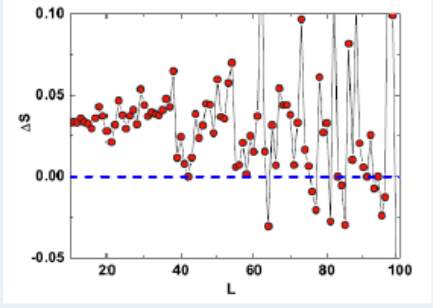
Badanie i modelowanie zmian sentymentu w dialogach on-line.



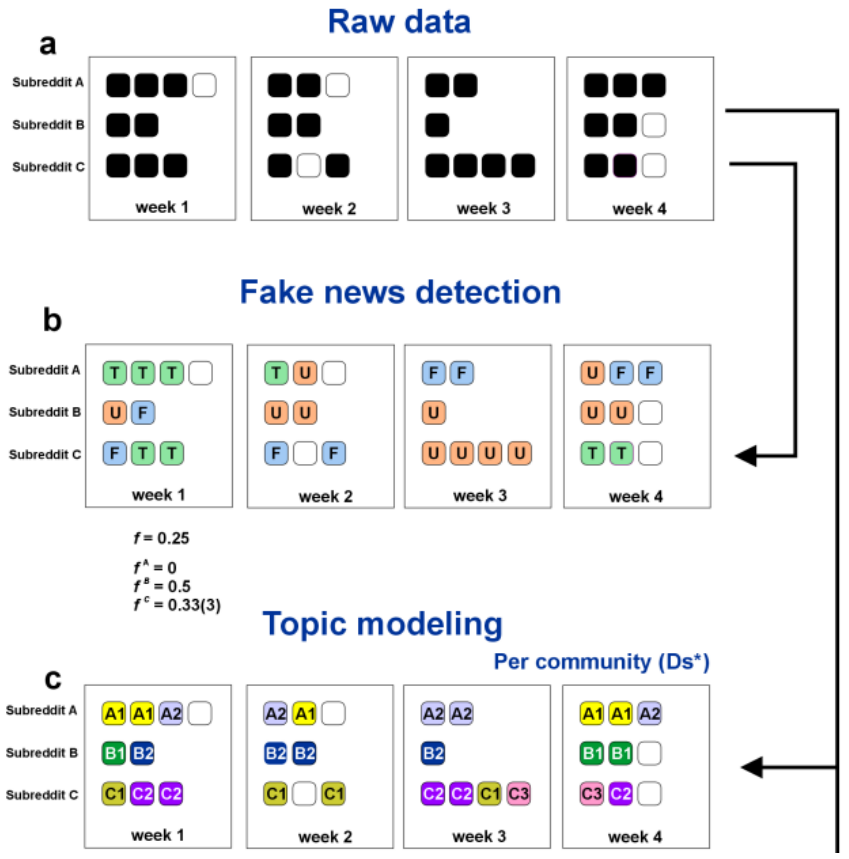
S jako wskaźnik...



... fazy dialogu.



Badanie zależnosci pomiędzy fałszywymi wiadomościami (fake news) a przeciążeniem informacyjnym w oparciu o model BERT.



Fizyka atmosfery arXiv:2211.02556

Atmosfera to pole na siatce przestrzennej. Token = kawałek atmosfery w danym miejscu i poziomie ciśnienia. Mechanizm uwagi uczy się, które regiony atmosferyczne są wzajemnie informatywne — np. że układ niżowy nad Atlantykiem wpływa na pogodę w Polsce za 3 dni

Fizyka statystyczna arXiv:2510.07401

Transformer trenowany metodą uczenia półnadzorowanego na konfiguracjach Monte Carlo modelu Isinga. Konfiguracje spinów na sieci -> tokeny. W fazie ferromagnetycznej spiny są skorelowane dalekozasięgowo – mechanizm uwagi uczy się łatwo struktury.

Fizyka kwantowa arXiv:2202.02541

Transformer z uwzględnieniem symetrii obrotowej, potencjały międzyatomowe. Atom = token, embedding = lokalne środowisko chemiczne. Mechanizm uwagi między atomami i j odzwierciedla, jak bardzo atom *j* wpływa na energię atomu *i*.

Fale grawitacyjne arXiv:2509.10912

Dwuwymiarowe obrazy zbudowane ze spektrogramu SFT (Short Fourier Transform. Vision Transformer dzieli ten obraz na prostokątne elementy i każdy z nich staje się tokenem. Mechanizm uwagi działa między fragmentami obrazu czasowo-częstotliwościowego.