# Twierdzenie o faktach i słowach dla procesów stacjonarnych

Łukasz Dębowski
ldebowsk@ipipan.waw.pl


Instytut Podstaw Informatyki
Polskiej Akademii Nauk

Seminarium TISO, WMIM UW
15 grudnia 2021

# Punkt wyjścia

**Teoria informacji:**

- **Podstawowy problem:**
  Jak zakodować dany napis (ciąg znaków) przy pomocy jak najmniejszej liczby cyfr binarnych (zer lub jedynek)?
- **Zastosowania:**
  — kompresja i przesyłanie danych, teoretyczne podstawy informatyki, statystyki, a nawet matematyki jako takiej.

**Statystyczne modelowanie języka naturalnego:**

- **Podstawowy problem:**
  Jakie przypisać prawdopodobieństwo dowolnym wypowiedziom w danym języku naturalnym (angielskim, polskim, chińskim)?
- **Zastosowania:**
  — automatyczne rozpoznawanie mowy, klawiatury telefonów komórkowych, maszynowe tłumaczenie i wiele innych.

# Teoria informacji a semantyka

*The concept of information [...] at first seems disappointing and bizarre—disappointing because it has nothing to do with meaning, and bizarre because [...] information and uncertainty find themselves to be partners.*

*[...] information and meaning may prove to be something like a pair of canonically conjugate variables in quantum theory, they being subject to some joint restriction that condemns a person to the sacrifice of the one as he insists on having much of the other.*

*[...] entropy not only speaks the language of arithmetic; it also speaks the language of language.*

— *Warren Weaver (1949)*

# „Ożywione” procesy stochastyczne

*If a Martian scientist sitting before his radio in Mars accidentally received from Earth the broadcast of an extensive speech [...], what criteria would he have to determine whether the reception represented the effect of animate process [...]? It seems that [...] the only clue to the animate origin would be this: the arrangement of the occurrences would be neither of rigidly fixed regularity such as frequently found in wave emissions of purely physical origin nor yet a completely random scattering of the same.*

— *George Kingsley Zipf (1965:187)*

# Temat wystąpienia

**Twierdzenie o faktach i słowach (sformułowanie nieformalne)**

Liczba niezależnych faktów opisywanych przez skończony tekst jest z grubsza mniejsza niż liczba różnych słów w tymże tekście.

W analogii do języka naturalnego, pojęcia faktu i słowa można zdefiniować dla dość dowolnego procesu stochastycznego.

# Ku sformułowaniu formalnemu

**Notacja:** Napis $x_j^k = x_j x_{j+1} \ldots x_k$, gdzie $x_i$ to symbole.

Na dalszych slajdach zdefiniujemy:

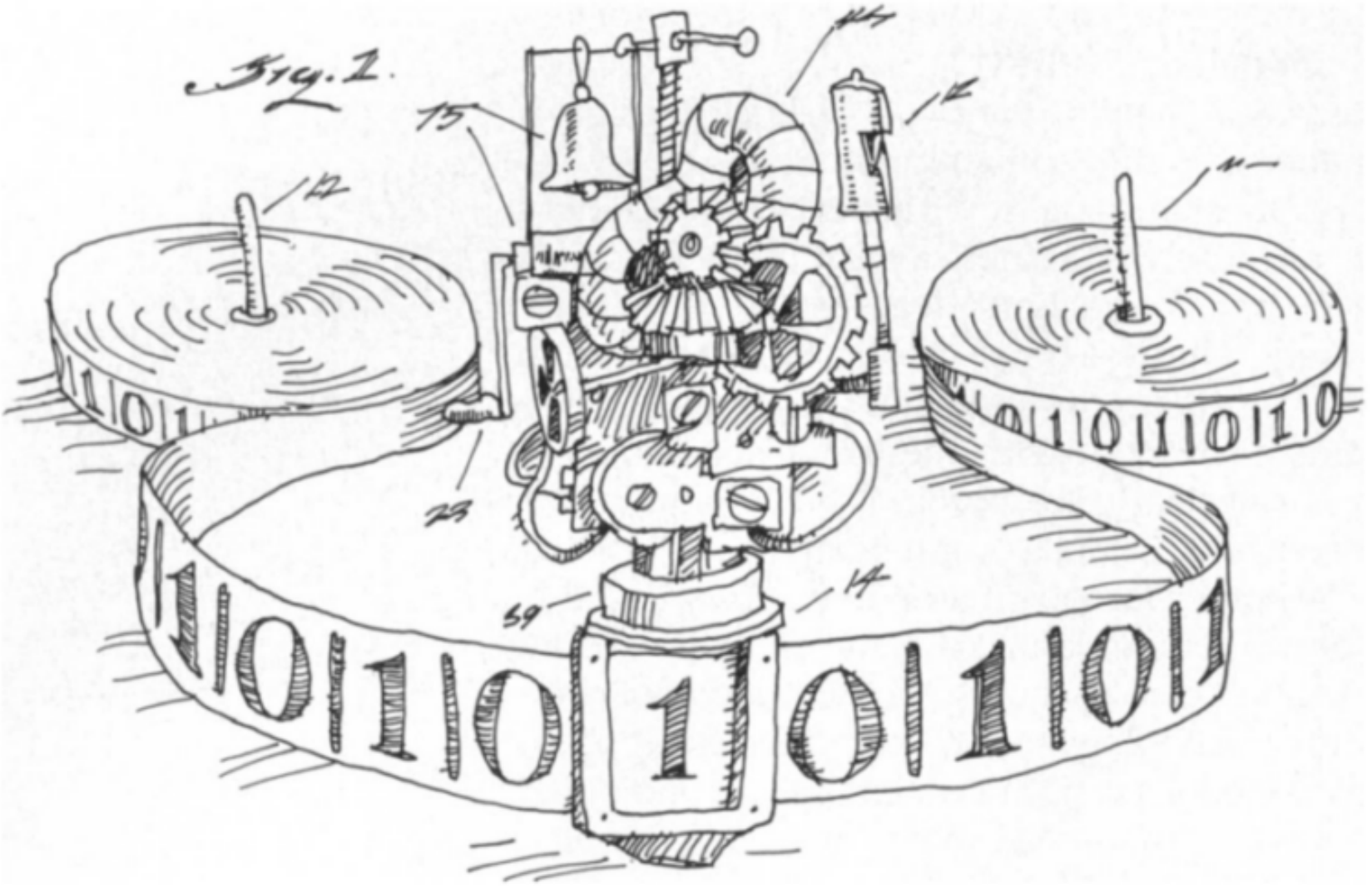
- $U(x_1^n)$ — liczba niezależnych faktów opisywanych przez $x_1^n$,
- $V(x_1^n)$ — liczba różnych słów użytych w $x_1^n$,
- $J(x_1^n; x_{n+1}^m)$ — informacja algorytmiczna między $x_1^n$ a $x_{n+1}^m$.
- wykładnik Hilberga: $\operatorname{hilb}_{n\to\infty} n^\beta = \beta$.

## Twierdzenie o faktach i słowach

Dla procesu stacjonarnego $(X_i)_{i=1}^\infty$ o skończonym alfabecie i dodatniej intensywności entropii Rényiego zachodzi nierówność

$$\operatorname{hilb}_{n\to\infty} \mathbb{E}\, U(X_1^n) \leq \operatorname{hilb}_{n\to\infty} \mathbb{E}\, J(X_1^n; X_{n+1}^{2n}) \leq \operatorname{hilb}_{n\to\infty} \mathbb{E}\, V(X_1^n).$$

# Maszyna Turinga z dzwonkiem — Tom Dunne, 2002



# Algorytmiczna teoria informacji

- Mamy ustaloną bezprefiksową maszynę Turinga.
- Złożoność Kołmogorowa $K(x_1^n)$ to długość najkrótszego programu generującego napis $x_1^n$:

$$K(x_1^n) := min\left\{|p| : \mathcal{U}(p) = x_1^n\right\},$$

gdzie $\mathcal{U}(p)$ to wynik programu $p$.

(Funkcja $x_1^n \mapsto K(x_1^n)$ nie jest obliczalna.)

- Nieskończony ciąg binarny $(x_i)_{i=1}^{\infty} = (x_1, x_2, x_3, ...)$ jest nazywany algorytmicznie losowym, gdy istnieje $c < \infty$ takie, że

$$K(x_1^n) \geq n - c.$$

(Zachodzi to, gdy najkrótszy program ma postać **print** $x_1^n$.)

- Ciąg rzutów uczciwą monetą jest algorytmicznie losowy z prawdopodobieństwem $1$.

# Prawdopodobieństwo stopu

- Prawdopodobieństwo stopu $\Omega$ to liczba

$$\Omega = \sum_{p:\mathcal{U}(p)\downarrow} 2^{-|p|} \in (0,1).$$

- Zdefiniujmy rozwinięcie binarne $(\Omega_k)_{k=1}^{\infty} = (\Omega_1, \Omega_2, \Omega_3, \ldots)$, gdzie $\Omega_k \in \{0,1\}$ oraz $\sum_{k=1}^{\infty} 2^{-k}\Omega_k = \Omega$.
- Ciąg $(\Omega_k)_{k=1}^{\infty}$ jest ciągiem algorytmicznie losowym.
- Ponadto, gdybyśmy znali napis $\Omega_1^n$, moglibyśmy odpowiedzieć na pytanie, które stwierdzenia matematyczne długości mniejszej od $n$ są prawdziwe, a które nie.

Prawdopodobieństwo stopu $\Omega$ jest „kamieniem filozoficznym”.
Cyfry $\Omega_k$ są niezależnymi faktami matematycznymi.

# Proces Santa Fe

1. Niech $(K_i)_{i=1}^{\infty}$ będzie ciągiem niezależnych zmiennych losowych o wartościach w liczbach naturalnych i o rozkładzie

$$P(K_i = k) = \frac{k^{-\alpha}}{\zeta(\alpha)}, \quad \alpha > 1,$$

gdzie $\zeta(\alpha) := \sum_{k=1}^{\infty} k^{-\alpha}$.

2. Proces Santa Fe to ciąg zmiennych $(X_i)_{i=1}^{\infty}$ złożonych z par

$$X_i = (K_i, \Omega_{K_i}),$$

gdzie $\Omega_k$ to cyfry prawdopodobieństwa stopu.

# Od procesu Santa Fe do prawdopodobieństwa stopu

- Zdefiniujmy obliczalną funkcję:

$$g(k, x_1^n) = \begin{cases} 0 & \text{jeśli } \exists_{1 \le i \le n} x_i = (k, 0) \text{ i } \neg\exists_{1 \le i \le n} x_i = (k, 1), \\ 1 & \text{jeśli } \exists_{1 \le i \le n} x_i = (k, 1) \text{ i } \neg\exists_{1 \le i \le n} x_i = (k, 0), \\ 2 & \text{inaczej.} \end{cases}$$

- Określmy liczbę niezależnych faktów opisywanych przez $x_1^n$:

$$U(x_1^n) := \min\left\{k \in \mathbb{N} : g(k, x_1^n) \neq \Omega_k\right\} - 1.$$

- Dla procesu Santa Fe mamy wzrost potęgowy

$$\lim_{n \to \infty} \frac{\mathbb{E}\, U(X_1^n)}{n^{1/\alpha}} \in (0, \infty).$$

# Liczba opisywanych faktów — w ogólności

- Niech $(z_k)_{k=1}^{\infty} = (z_1, z_2, z_3, \ldots)$ będzie ustalonym binarnym ciągiem algorytmicznie losowym, niekoniecznie równym rozwinięciu binarnemu prawdopodobieństwa stopu.
- Symbole $z_k$ nazywać będziemy niezależnymi faktami.
- Niech $g(k, x_1^n)$ będzie ustaloną funkcją obliczalną.
- Niech $x_1^n$ będzie dowolnym napisem.

Liczba niezależnych faktów opisywanych przez napis $x_1^n$:

$$U(x_1^n) := \min\left\{k \in \mathbb{N} : g(k, x_1^n) \neq z_k\right\} - 1.$$

# Jak zdefiniować słowo w dowolnym tekście?

- W przypadku tekstów w wielu językach naturalnych, słowo to ciąg liter od spacji do spacji:

  Все люди рождаются свободными и равными в своем достоинстве и правах. Они наделены разумом и совестью и должны поступать в отношении друг друга в духе братства.

- Co począć z japońskim?

  すべての人間は、生まれながらにして自由であり、かつ、尊厳と権利とについて平等である。人間は、理性と良心とを授けられており、互いに同 胞の精神をもって行動しなければならない。

- Albo z ciągiem zer i jedynek?

  $\Omega = 0.000000100000010000011000100001101000111\ldots$

# Podejście toporne, acz skuteczne

- Subword complexity, czyli liczba różnych podsłów długości $k$ w napisie $x_1^n$ to

$$V(k|x_1^n) := \left|\left\{x_{t+1}^{t+k} : 0 \leq t \leq n-k\right\}\right|.$$

- Jak wybrać optymalne $k$?

# Estymator rzędu Markowa

- Częstość podsłowa $a_1^k$ w napisie $x_1^n$:

$$N(a_1^k|x_1^n) := \sum_{i=1}^{n-k+1} 1\left\{x_i^{i+k-1} = a_1^k\right\}$$

- Warunkowa entropia empiryczna $k$-tego rzędu:

$$\mathcal{H}(k|x_1^n) := \sum_{a_1^{k+1}} \frac{N(a_1^{k+1}|x_1^n)}{n-k} \log \frac{N(a_1^k|x_1^{n-1})}{N(a_1^{k+1}|x_1^n)}.$$

- Bezprefiksowa złożoność Kołmogorowa: $K(x_1^n)$.
- Estymator rzędu Markowa (zgodny!!!):

$$M(x_1^n) := \inf\left\{k \geq 0 : (n-k)\mathcal{H}(k|x_1^n) \leq K(x_1^n)\right\}.$$

Liczba różnych słów użytych w napisie $x_1^n$:

$$V(x_1^n) := V(M(x_1^n)|x_1^n).$$

# Wykładnik Hilberga

- Dla języka naturalnego mamy w przybliżeniu prawo Herdana

$$V(x_1^n) \propto n^{0.8},$$

jeżeli w estymatorze rzędu Markowa przybliżymy złożoność Kołmogorowa przez długość kodu uniwersalnego.

- Ponadto dla języka naturalnego hipoteza Hilberga głosi, że

$$J(x_1^n; x_{n+1}^{2n}) \propto n^{\beta}, \beta \in (0, 1).$$

- Wykładnik Hilberga definiujemy jako

$$\operatorname*{hilb}_{n\to\infty} s(n) := \limsup_{n\to\infty} \frac{\log s(n)}{\log n}.$$

Mamy $\operatorname*{hilb}_{n\to\infty} n^{\beta} = \beta$.

- Dla procesu Santa Fe mamy

$$\operatorname*{hilb}_{n\to\infty} \mathbb{E}\, U(X_1^n) = \frac{1}{\alpha} \in (0, 1).$$

# Intensywność entropii Rényiego

- Proces stacjonarny: ciąg zmiennych $(X_i)_{i=1}^{\infty}$, gdzie

$$P(X_{j+1}^{j+n} = x_1^n) = P(X_1^n = x_1^n).$$

- Entropia Rényiego:

$$H_2(X_1^n) := -\log \sum_{x_1^n} P(X_1^n = x_1^n)^2.$$

- Intensywność entropii Rényiego:

$$h_2 := \inf_{n \geq 1} \frac{H_2(X_1^n)}{n}.$$

# Informacja algorytmiczna

Równości i nierówności poniżej są z dokładnością do stałej.

- Bezprefiksowa złożoność Kołmogorowa: $K(u)$.
- Informacja algorytmiczna:

$$J(u;v) := K(u) + K(v) - K(u,v).$$

- Złożoność warunkowa:

$$\bar{K}(u|v) := K(u|v, K(v)) = K(u,v) - K(v) \leq K(u).$$

- Reguły łańcuchowe:

$$\bar{K}(u|v) + J(u;v) = K(u),$$
$$\bar{K}(u|v) + K(v) = K(u,v).$$

# Twierdzenie o faktach i słowach

Na poprzednich slajdach zdefiniowaliśmy:

- $U(x_1^n)$ — liczba niezależnych faktów opisywanych przez $x_1^n$,
- $V(x_1^n)$ — liczba różnych słów użytych w $x_1^n$,
- $J(x_1^n; x_{n+1}^m)$ — informacja algorytmiczna między $x_1^n$ a $x_{n+1}^m$.
- wykładnik Hilberga: $\operatorname{hilb}_{n\to\infty} n^\beta = \beta$.

## Twierdzenie o faktach i słowach

Dla procesu stacjonarnego $(X_i)_{i=1}^\infty$ o skończonym alfabecie i dodatniej intensywności entropii Rényiego zachodzi nierówność

$$\operatorname{hilb}_{n\to\infty} \mathbb{E}\, U(X_1^n) \le \operatorname{hilb}_{n\to\infty} \mathbb{E}\, J(X_1^n; X_{n+1}^{2n}) \le \operatorname{hilb}_{n\to\infty} \mathbb{E}\, V(X_1^n).$$

# Twierdzenie o wykładnikach Hilberga

## Twierdzenie

Niech $\nabla S(n) := 2S(n) - S(2n)$. Jeżeli $\lim_{n\to\infty} S(n)/n = s$, to

$$\operatorname{hilb}_{n\to\infty} [S(n) - ns] \leq \operatorname{hilb}_{n\to\infty} \nabla S(n),$$

gdzie równość zachodzi, gdy $S(n) - ns \geq 0$.

# Dowód

Mamy tożsamość teleskopową

$$\sum_{k=0}^{\infty} \frac{\nabla S(2^k n)}{2^{k+1}} = S(n) - ns.$$

Oznaczmy $\beta = \mathbf{hilb}_{n\to\infty} \nabla S(n) < 1$ b.z.o.r. Mamy $\nabla S(n) \leq n^{\beta+\epsilon}$ dla dużych $n$. Dla $\epsilon < 1 - \beta$, tożsamość teleskopowa daje

$$S(n) - ns \leq n^{\beta+\epsilon} \sum_{k=0}^{\infty} \frac{2^{k(\beta+\epsilon)}}{2^{k+1}} = \frac{n^{\beta+\epsilon}}{2(1 - 2^{\beta+\epsilon-1})}.$$

Biorąc dowolnie małe $\epsilon$, otrzymujemy $\mathbf{hilb}_{n\to\infty} (S(n) - ns) \leq \beta$.

Załóżmy teraz $S(n) \geq ns$ dla każdego $n$. Mamy

$$S(n) - ns = \frac{\nabla S(n)}{2} + \frac{S(2n) - 2ns}{2} \geq \frac{\nabla S(n)}{2}.$$

A zatem $\mathbf{hilb}_{n\to\infty} (S(n) - ns) \geq \beta$.

# Ważny przykład dla procesu stacjonarnego

- Entropia Shannona:

$$H(X_1^n) := \mathbb{E}\left[-\log P(X_1^n)\right].$$

- Intensywność entropii Shannona:

$$h := \lim_{n\to\infty} \frac{H(X_1^n)}{n} = \inf_{k\geq 0} H(X_{k+1}|X_1^k).$$

- Informacja wzajemna Shannona:

$$I(X_1^n; X_{n+1}^{2n}) := H(X_1^n) + H(X_{n+1}^{2n}) - H(X_1^{2n}) = \nabla H(X_1^n).$$

- Ponieważ $H(X_1^n) \geq hn$, to

$$\operatorname*{hilb}_{n\to\infty}\left[H(X_1^n) - nh\right] = \operatorname*{hilb}_{n\to\infty} I(X_1^n; X_{n+1}^{2n}).$$

# Twierdzenie o faktach

Rozpatrujemy napisy nad alfabetem skończonym $\{1, 2, ..., D\}$.

Z reguły łańcuchowej i nierówności $U(x_1^n) \leq K(x_1^n) \leq n \log D$,

$$K(x_1^n) = K(x_1^n, z_1^{U(x_1^n)}) = K(z_1^{U(x_1^n)}) + \bar{K}(x_1^n | z_1^{U(x_1^n)})$$
$$\geq U(x_1^n) + \bar{K}(x_1^n | z_1^{n \log D}) - O(\log n).$$

Z nierówności kodowania $\mathbb{E}\, \bar{K}(X_1^n | z_1^{n \log D}) \geq H(X_1^n) \geq hn$, więc

$$\mathbb{E}\, K(X_1^n) \geq \mathbb{E}\, U(X_1^n) + hn - O(\log n).$$

W rezultacie mamy twierdzenie o faktach:

$$\operatorname*{hilb}_{n\to\infty} \mathbb{E}\, U(X_1^n) \leq \operatorname*{hilb}_{n\to\infty} \left[\mathbb{E}\, K(X_1^n) - nh\right].$$

# Kodowanie uniwersalne (prediction by partial matching)

Estymatory Laplace'a dla modeli Markowa $k$-tego rzędu:

$$R_k(x_1^n) := \frac{1}{D^{k+1}} \prod_{i=k+1}^{n} \frac{N(x_{n+1-k}^{n+1}|x_1^n)+1}{N(x_{n+1-k}^{n}|x_1^{n-1})+D}, \quad n \geq k+1.$$

Przepiszmy je jako

$$\begin{aligned} R_k(x_1^n) &= \frac{1}{D^k} \prod_{a_1^k} \frac{\prod_{a_{k+1}} 1 \cdot 2 \ldots N(a_1^{k+1}|x_1^n)}{D \cdot (D+1) \ldots (N(a_1^k|x_1^{n-1})+D-1)} \\ &= \frac{1}{D^k} \prod_{a_1^k} \frac{(D-1)! \prod_{a_{k+1}} N(a_1^{k+1}|x_1^n)!}{(N(a_1^k|x_1^{n-1})+D-1)!}. \end{aligned}$$

Z przybliżenia Stirlinga $\frac{n^n}{e^n} \leq n! \leq \frac{(n+1)^{n+1}}{e^n}$ otrzymujemy

$$-\log R_k(x_1^n) \leq k \log D + (n-k)\mathcal{H}(k|x_1^n) + DV(k|x_1^n) \log n.$$

# Twierdzenie o informacji algorytmicznej

Oznaczmy $\mathcal{S}(k|u) := (|u| - k)\mathcal{H}(k|u)$.

Z kodowania Shannona-Fano otrzymujemy nierówność

$$K(u) \leq 3 \log |u| + k \log D + \mathcal{S}(k|u) + DV(k|u) \log |u| \,.$$

Z twierdzenia ergodycznego otrzymujemy

$$\lim_{n\to\infty} \mathcal{H}(k|X_1^n) = H(X_{k+1}|X_1^k) \implies \lim_{n\to\infty} \frac{K(X_1^n)}{n} = h.$$

W rezultacie mamy twierdzenie o informacji algorytmicznej:

$$\operatorname*{hilb}_{n\to\infty} \left[\mathbb{E}\, K(X_1^n) - nh\right] = \operatorname*{hilb}_{n\to\infty} \mathbb{E}\, J(X_1^n; X_{n+1}^{2n}).$$

# Twierdzenie o słowach

Entropia empiryczna jest superaddytywna (maximum likelihood):

$$\mathcal{S}(k|u) + \mathcal{S}(k|v) - \mathcal{S}(k|uv) \leq 0.$$

Jeżeli położymy $k = M(w)$, gdzie $w = uv$, to

$$\begin{aligned} J(u;v) &\leq K(u) + K(v) - K(uv) \leq 3\log|u| + 3\log|v| \\ &\quad + 2k\log D + DV(k|u)\log|u| + DV(k|v)\log|v| \\ &\leq 6\log|w| + 2M(w)\log D + 2DV(w)\log|w|\,. \end{aligned}$$

Estymator rzędu Markowa spełnia nierówność:

$$\frac{M(w)}{\log|w|} \leq \frac{|w|}{K(w)}, \text{ gdzie } \mathbb{E}\left(\frac{n}{K(X_1^n)}\right) \leq \frac{4}{h_2}.$$

W rezultacie dla $h_2 > 0$ mamy twierdzenie o słowach:

$$\operatorname*{hilb}_{n\to\infty} \mathbb{E}\, J(X_1^n; X_{n+1}^{2n}) \leq \operatorname*{hilb}_{n\to\infty} \mathbb{E}\, V(X_1^n).$$

# Konkluzje

1. Przedstawiliśmy twierdzenie o faktach i słowach, które orzeka, że liczba niezależnych faktów opisywanych przez skończony tekst jest mniejsza niż liczba różnych słów w tymże tekście.
2. Dla tekstów w języku naturalnym, liczba różnych słów zdaje się rosnąć potęgowo z długością tekstu.
3. Czy zatem liczba niezależnych faktów opisywanych przez teksty w języku naturalnym również rośnie potęgowo z długością tekstu?
4. Czy proces, dla którego liczba słów i liczba faktów są zbliżone, musi przypominać proces Santa Fe?
5. Jakie są inne przykłady procesów perygraficznych, czyli o potęgowo rosnącej liczbie faktów?

# Literatura


1. Ł. Dębowski, (2021). A Refutation of Finite-State Language Models Through Zipf's Law for Factual Knowledge. *Entropy*, vol. 23, pp. 1148.
2. Ł. Dębowski, (2020). Information Theory Meets Power Laws: Stochastic Processes and Language Models, Wiley.
3. Ł. Dębowski, (2018). Is Natural Language a Perigraphic Process? The Theorem about Facts and Words Revisited. *Entropy*, vol. 20(2), pp. 85.
4. R. Takahira, K. Tanaka-Ishii, Ł. Dębowski, (2016). Entropy Rate Estimates for Natural Language—A New Extrapolation of Compressed Large-Scale Corpora. *Entropy*, vol. 18(10), pp. 364.
5. Ł. Dębowski, (2011). On the Vocabulary of Grammar-Based Codes and the Logical Consistency of Texts. *IEEE Transactions on Information Theory*, vol. 57, pp. 4589–4599.
6. Ł. Dębowski, (2006). On Hilberg's law and its links with Guiraud's law. *Journal of Quantitative Linguistics*, vol. 13, pp. 81–109.

# Dziękuję za uwagę!