O statystycznym modelowaniu języka z elementami teorii informacji

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Instytut Podstaw Informatyki
Polskiej Akademii Nauk

Seminarium „Teoria informacji”
Wydział Psychologii UW, Warszawa, 17.01.2018

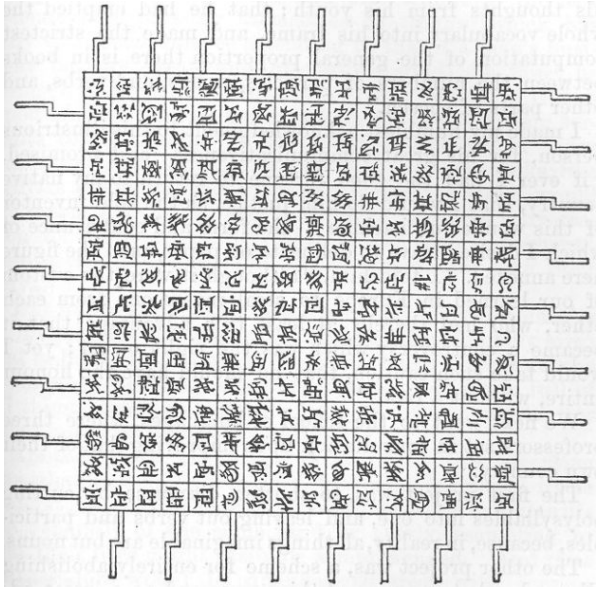
Moje zainteresowania

1. **Statystyczne modelowanie języka:**
   - Problem teoretyczny:
     Jakie przypisać prawdopodobieństwo dowolnym wypowiedziom w danym języku naturalnym (angielskim, polskim, ...)?
   - Zastosowania praktyczne:
     — automatyczne rozpoznawanie mowy, klawiatury telefonów komórkowych, maszynowe tłumaczenie, sztuczna inteligencja.
2. **Teoria informacji:**
   - Problem teoretyczny:
     Jak określić *ilość* informacji w zmiennej losowej bądź w ustalonym napisie? $\longrightarrow$ entropia, informacja wzajemna, złożoność Kołmogorowa...
   - Zastosowania praktyczne:
     — przesyłanie danych przez zaszumione kanały, kompresja danych, automatyczna korekcja błędów.

Jonathan Swift (1667–1745)

Podróże Guliwera (1726):

Zapewnił mnie, że ten wynalazek był owocem wszystkich jego myśli od wczesnej młodości, że użył całego dykcjonarza do tych ram i obliczył ściśle proporcje, jakie są w księgach między rodzajnikami, imionami, czasownikami i innymi rodzajami mowy.



Andriej Andriejewicz Markow (1856–1922)

Matematyk rosyjski. Autor pojęcia łańcucha Markowa. W wykładzie wygłoszonym w 1913 w Petersburgu przedstawił zastosowanie pojęcia łańcucha Markowa do analizy poematu *Eugeniusz Oniegin* Aleksandra Puszkina. Szacował w nim prawdopodobieństwo warunkowe występowania po sobie spółgłosek i samogłosek w analizowanym tekście.

Procesy Markowa

- Proces stochastyczny $(X_i)_{i=1}^{\infty}$ na przestrzeni $(\Omega, \mathcal{J}, P)$.
- Bloki zmiennych losowych $X_j^k := (X_j, X_{j+1}, ..., X_k)$.
  (notacja z teorii informacji)
- P-stwo warunkowe zależy tylko od ostatniej zmiennej:

$$P(X_i|X_1^{i-1}) = P(X_i|X_{i-1})$$

- Estymacja największej wiarogodności:

$$P_{MLE}(X_i|X_{i-1}) := \frac{N(X_{i-1}^i|x_1^n)}{N(X_{i-1}|x_1^{n-1})},$$

gdzie
— $N(w|z)$ to liczba wystąpień podsłowa w w słowie z,
— ciąg $x_1^N = (x_1, x_2, ..., x_N)$ to próba ucząca.

G. Udny Yule (1871–1951), Herbert A. Simon (1916–2001)

Statystyk brytyjski i polimat amerykański. Twórcy rozkładu Yule'a-Simona i procesu Yule'a (1925,1955), współcześnie znanych jako *preferential attachment* czy też efekt św. Mateusza, zaproponowanych w kontekście modelowania danych ekologicznych i lingwistycznych. Rozkład Yule'a jest przykładem procesu o potęgowym ogonie. G. U. Yule jest także autorem książki *The Statistical Study of Literary Vocabulary* (1944), w której wprowadził stałą $\boldsymbol{K}$ jako narzędzie w atrybucji autorstwa tekstów.

George Kingsley Zipf (1902–1950)

Lingwista amerykański. Autor książki *The Psycho-Biology of Language: An Introduction to Dynamic Philology* (1935). Przedstawił w niej empiryczne prawo zwane później prawem Zipfa. Prawo to głosi, że częstość dowolnego słowa w tekście jest z grubsza odwrotnie proporcjonalna do rangi tego słowa.

Gwoli ścisłości odkrywcą prawa Zipfa był Jean-Baptiste Estoup (1868–1950), stenograf francuski, autor książki *Gammes sténographiques* (1912).

Przykład listy rangowej

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*

ranga $r(w)$	częstość $f(w)$	słowo w	$r(w) \cdot f(w)$
1	14767	w	14767
2	12473	i	24946
3	11093	się	33279
...	...	...	...
210	214	ciągu	44940
211	213	jeśli	44943
212	212	czas	44944
213	210	ludzie	44730
...	...	...	...
38420	2	Aaa	76840
38421	1	życznej	38421
...	...	...	...
92963	1	aa	92963

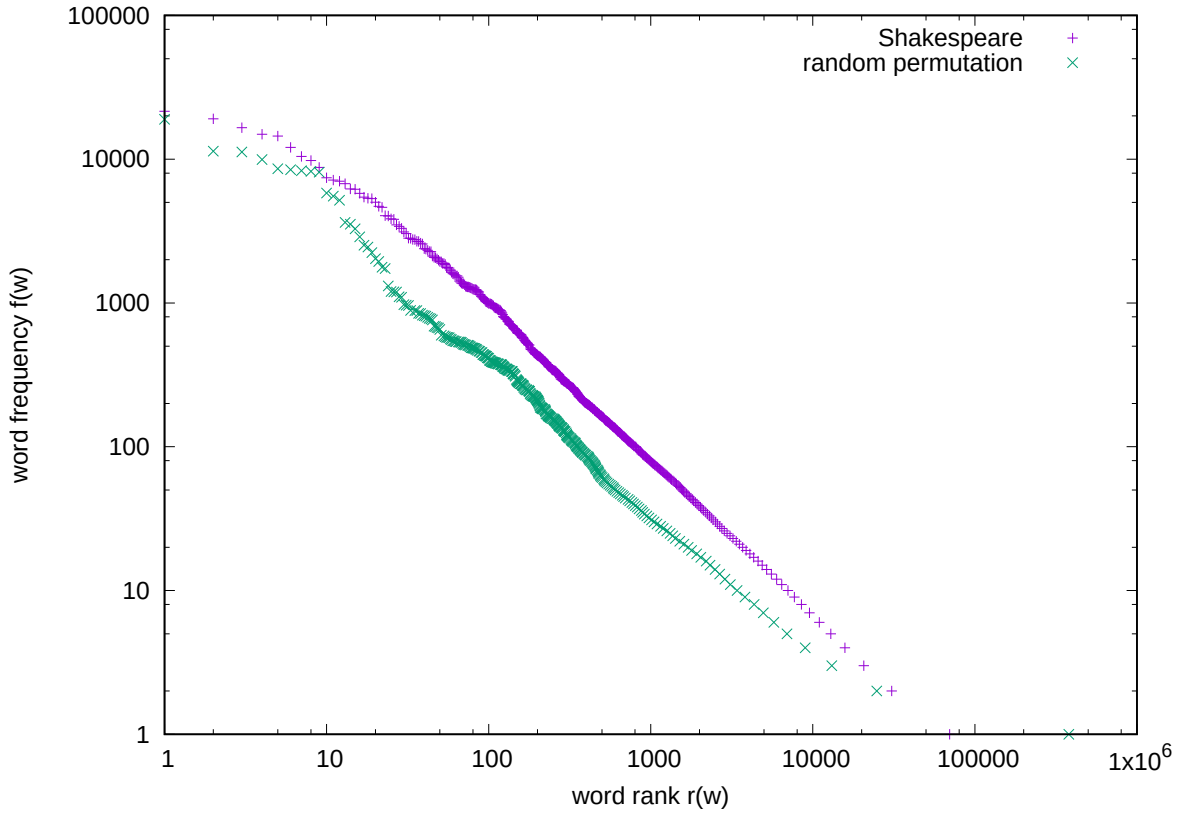
Benoît B. Mandelbrot (1924–2010)

Matematyk polsko-żydowskiego pochodzenia. Twórca geometrii fraktalnej i autor słowa „fraktal". Próbował zastosować pojęcie fraktali do modelowania języka naturalnego i zaobserwował w roku 1953, że prawo Zipfa spełnione jest przez teksty otrzymane przez niezależne losowanie kolejnych liter i odstępów w tekście.

George Armitage Miller (1920-2012)

Psycholog amerykański. Przyczynił się do narodzin psycholingwistyki i kognitywistyki. Niezależnie od B. B. Mandelbrota także zaobserwował w roku 1957, że prawo Zipfa spełnione jest przez teksty otrzymane przez niezależne losowanie kolejnych liter i odstępów w tekście.

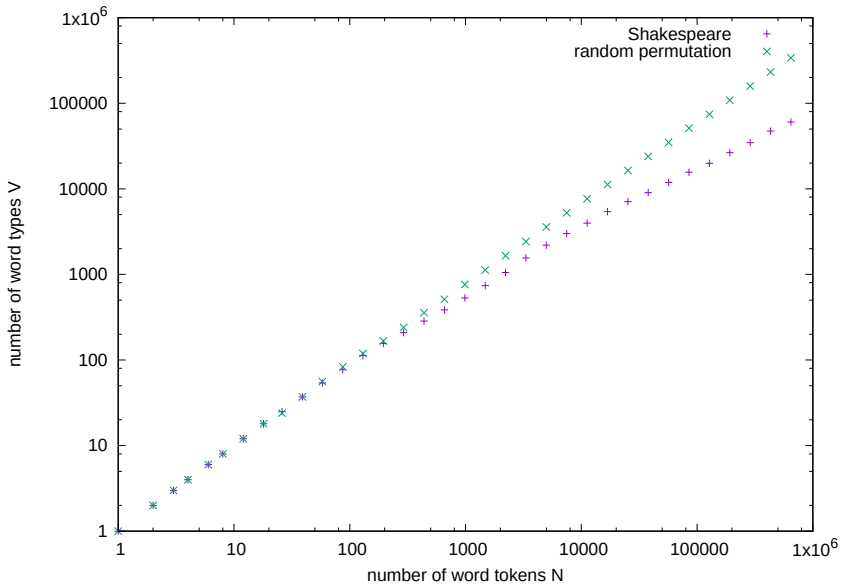
Prawo Zipfa (ranga-częstość)



Estoup 1916; Zipf 1935; Mandelbrot 1953, Miller 1957:

$$f(w) \propto \left[\frac{1}{B + r(w)}\right]^{\alpha}, \quad \alpha > 1$$

Prawo Heapsa (liczba różnych słów)



Kuraszkiewicz i Łukaszewicz 1951; Herdan 1964; Heaps 1978:

$$V \propto N^{\beta}, \quad \beta < 1, \quad \beta \approx 1/\alpha$$

V — liczba różnych słów w tekście (typów/types)
N — liczba wszystkich słów tekście (okazów/tokens)

Claude Elwood Shannon (1902–2001)

Inżynier amerykański. Twórca teorii informacji, autor pojęcia entropii zmiennej losowej i modelu n-gramowego (1948). Motywujące założenie teorii informacji stanowi, że teksty w języku naturalnym można modelować jako proces stochastyczny. Modele n-gramowe, czyli modele Markowa n-tego rzędu, są pewną próbą estymacji rozkładu p-stwa tego procesu.

Entropia i intensywność entropii

- Entropia zmiennej losowej:

$$H(X) = -\sum_x P(X = x) \log P(X = x)$$

- Entropia warunkowa:

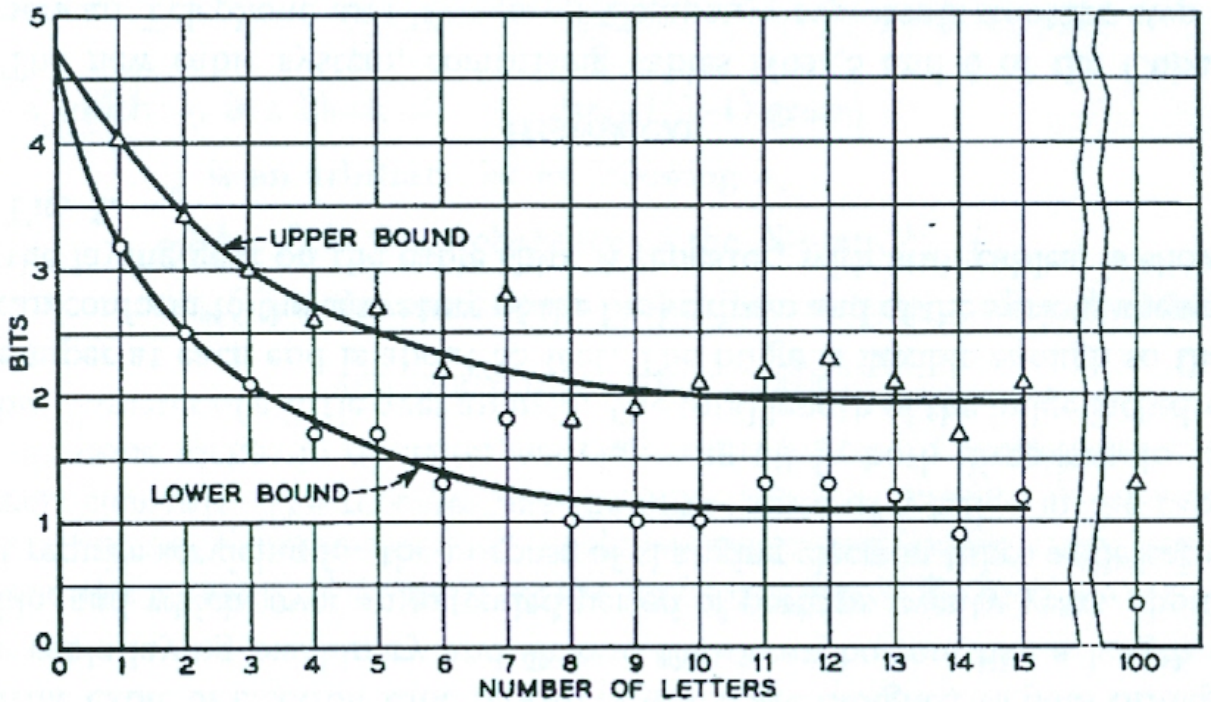
$$H(X|Y) = -\sum_x P(X = x, Y = y) \log P(X = x|Y = y)$$

- Intensywność entropii procesu stacjonarnego $(X_i)_{i=1}^{\infty}$:

$$h = \lim_{n\to\infty} \frac{H(X_1^n)}{n} = \lim_{n\to\infty} H(X_n|X_1^{n-1})$$

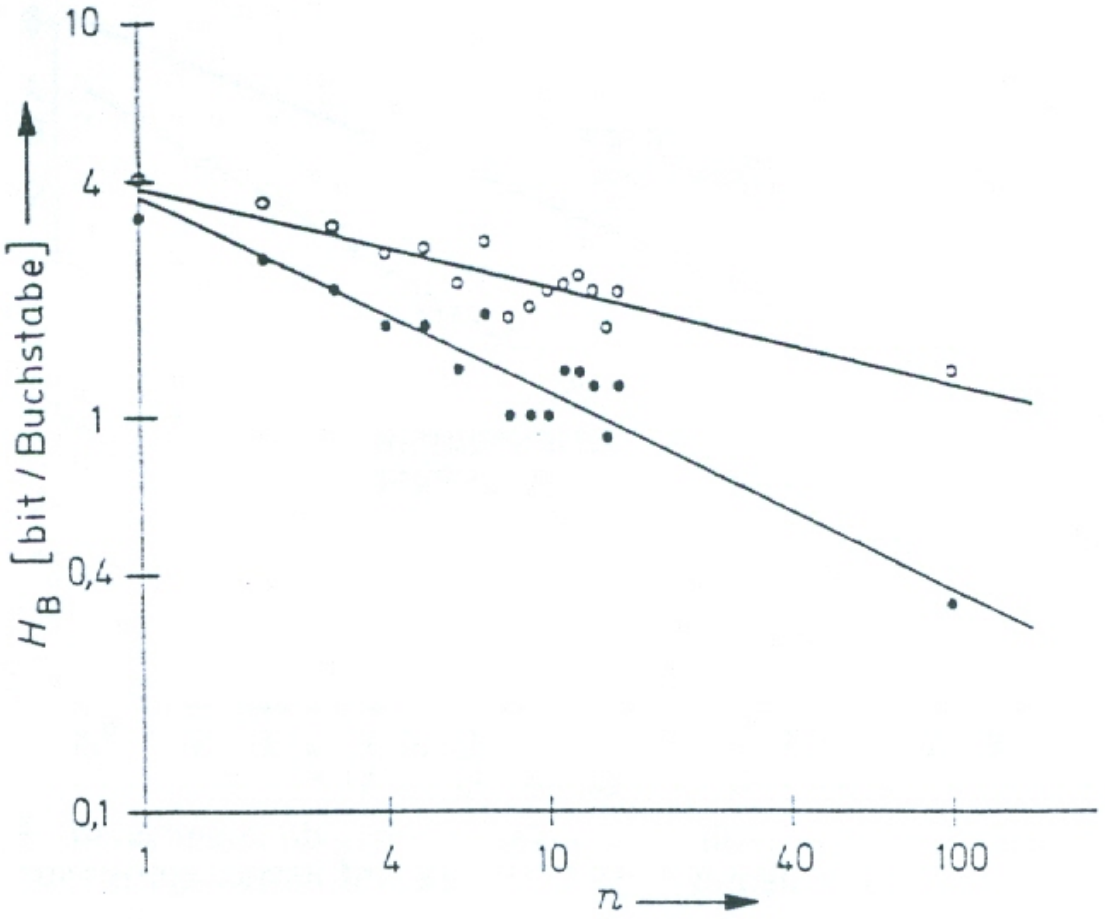
Intensywność entropii dla języka angielskiego

Shannon (1951), *Prediction and entropy of printed English*.



Intensywność entropii języka naturalnego jest rzędu 1 bita na literę.

Hipoteza Hilberga (1990)



$$H(X_n|X_1^{n-1}) \approx Bn^{\beta-1} + h, \quad \beta \approx 1/2, \quad n \leq 100$$

Model n-gramowy

- Proces stochastyczny $(X_i)_{i=1}^{\infty}$ na przestrzeni $(\Omega, \mathcal{J}, P)$.
- Bloki zmiennych losowych $X_j^k := (X_j, X_{j+1}, ..., X_k)$.
- P-stwo warunkowe zależy tylko od $n-1$ ostatnich zmiennych:

$$P(X_i|X_1^{i-1}) = P(X_i|X_{i-n+1}^{i-1})$$

- Estymacja największej wiarogodności

$$P_{MLE}(X_i|X_{i-n+1}^{i-1}) = \frac{N(X_{i-n+1}^{i}|x_1^N)}{N(X_{i-n+1}^{i-1}|x_1^{N-1})},$$

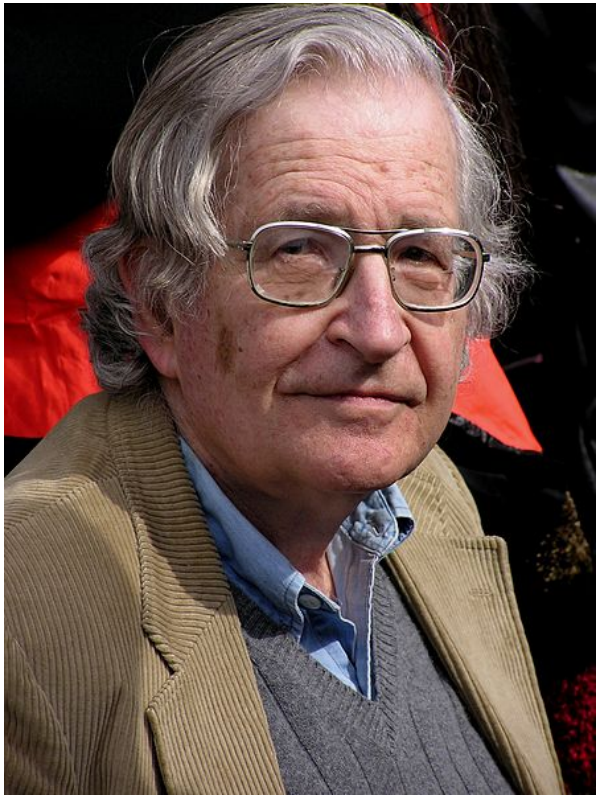
gdzie
— $N(w|z)$ to liczba wystąpień podsłowa w w słowie z,
— ciąg $x_1^N = (x_1, x_2, ..., x_N)$ to próba ucząca.

Modele *n*-gramowe — przykłady generowanych tekstów

- model **2**-gramowy:
  `oć sświatw wsz sta paku wo pojz ktćda bi iańcychy`
- model **3**-gramowy:
  `kłobez odzie na w swarza z le czenie niasną drały`
- model **4**-gramowy:
  `rykomadzie jużbie, w rancza Rzeciwsze z nie`
- model **5**-gramowy:
  `yk Siedziesiąt tysię, na pan Muszając; przysia`
- model **6**-gramowy:
  `k Sieniu oka mgnieniu, męstwę i rzuciła ślady`
- model **10**-gramowy:
  `enkiewicz, Pan Wołodyjowski wyjechać na objazd.`
- model **20**-gramowy:
  `Pan Wołodyjowski zaniepokoił się tym bardzo`

Noam Chomsky (1928–)

Lingwista amerykański. Twórca hierarchii Chomsky'ego, czyli hierarchii języków formalnych, i formalnych teorii składni języka naturalnego (1957). Znany jest z bardzo wpływowego sceptycznego stosunku do statystycznego modelowania języka naturalnego.



Wpływowa w lingwistyce była też praca E. Marka Golda (1967) *Language identification in the limit* pokazująca, że w pewnym ujęciu niestatystycznym niemożliwe jest nauczenie się z wyłącznie danych pozytywnych nieskończonych języków formalnych.

Andriej Nikołajewicz Kołmogorow (1903–1987)

Matematyk rosyjski. Twórca współczesnej teorii prawdopodobieństwa. Miał także wątpliwości, czy prawdopodobieństwo zdań i dłuższych tekstów w języku naturalnym ma sensowną interpretację częstościową, ale z tego punktu widzenia zaproponował algorytmiczne podejście do definicji ilości informacji zawartej w dowolnym napisie (1965). Współcześnie wielkość ta nazywana jest złożonością Kołmogorowa.

Algorytmiczna teoria informacji

- Złożoność Kołmogorowa:

$$K(w) = min\{|p| : U(p) = w\}$$

gdzie $U(p)$ to wynik programu p.

(Złożoność Kołmogorowa nie jest efektywnie obliczalna.)

- Napis w jest nazywany algorytmicznie losowym, gdy:

$$K(w) \approx |w|$$

Zachodzi to, gdy najkrótszy program ma postać **print** w**;**

- Dla efektywnie obliczalnego rozkładu p-stwa:

$$0 \leq \mathbb{E}\, K(X_1^n) - H(X_1^n) \leq K(P) + C$$

Zachodzi też podobna relacja prawie na pewno.

Frederick Jelinek (1932–2010)

Amerykański inżynier czeskiego pochodzenia. Twórca systemów automatycznego rozpoznawania mowy opartych na statystycznym modelowaniu języka naturalnego, ukrytych modelach Markowa i modelach n-gramowych. Często cytowane jest jego powiedzenie: *Every time I fire a linguist, the performance of the speech recognizer goes up.*

Automatyczne rozpoznawanie mowy

- Reguła Bayesa:

$$P(\text{tekst}|\text{mowa}) = \frac{P(\text{mowa}|\text{tekst})P(\text{tekst})}{P(\text{mowa})}$$

  Wybieramy tekst o najwyższym p-stwie a posteriori.
- Model języka $P(\text{tekst})$ szacuje się jako model n-gramowy, najczęściej używając $n = 3$ dla słów (trigramy).

Problem rzadkości danych

- Przeciętne czynne słownictwo człowieka $\approx \mathbf{10^4}$ słów.
- Liczba różnych trigramów $\approx \mathbf{10^{12}}$.
- Współczesne korpusy tekstów $\approx \mathbf{10^9}$ słów.
- Nie jesteśmy w stanie sensownie wyestymować p-stw w oparciu o estymację największej wiarogodności

$$P_{MLE}(X_i|X_{i-n+1}^{i-1}) = \frac{N(X_{i-n+1}^{i}|x_1^N)}{N(X_{i-n+1}^{i-1}|x_1^{N-1})},$$

gdzie
— $N(w|z)$ to liczba wystąpień podsłowa w w słowie z,
— ciąg $x_1^N = (x_1, x_2, ..., x_N)$ to próba ucząca.

Jak uniknąć zerowych i nieokreślonych p-stw warunkowych?

- Przykładowe wygładzanie prawdopodobieństw:

$$P_n(X_i|X_{i-n+1}^{i-1}) = \frac{N(X_{i-n+1}^{i}|x_1^N) + \lambda_n P_{n-1}(X_i|X_{i-n+2}^{i-1})}{N(X_{i-n+1}^{i-1}|x_1^{N-1}) + \lambda_n},$$

  gdzie λ_n to wolne parametry.
- Parametry λ_n dobiera się minimalizując entropię krzyżową

$$-\sum_{i=3}^{M} \log P_3(X_i = y_i|X_{i-2}^{i-1} = y_{i-2}^{i-1})$$

  na danych walidacyjnych $y_1^M = (y_1, y_2, ..., y_M)$.
- Zaproponowano wiele innych technik wygładzania (np. estymator Gooda-Turinga).
- Entropia krzyżowa takich modeli jest rzędu 1,5 bita na literę.

Maksymalne powtórzenie

Maksymalne powtórzenie (maximal repetition) $L(x_1^n)$ w tekście $x_1^n = (x_1, x_2, \ldots, x_n)$ to maksymalna długość powtarzającego się podsłowa.

Formalnie,

$$L(x_1^n) := \max\left\{k : x_{i+1}^{i+k} = x_{j+1}^{j+k} \text{ dla pewnych } 0 \leq i < j \leq n-k\right\}.$$

Przykład:

$$x_1^n = \text{“O szyby deszcz dzwoni, deszcz dzwoni jesienny.”}$$
$$L(x_1^n) = |\text{“ deszcz dzwoni”}| = 14.$$

Maksymalne powtórzenie $L(x_1^n)$ można policzyć w czasie $O(n)$ sortując drzewo sufiksów (Kolpakov & Kucherov, 1999).

Z punktu widzenia probabilistów... (Erdős & Rényi, 1970)

Niech $(X_i)_{i=1}^{\infty}$ będzie procesem IID, tzn. nieskończonym ciągiem niezależnych zmiennych losowych o identycznym rozkładzie,

$$P(X_1^n = x_1^n) = \prod_{i=1}^{n} p(x_i).$$

Można wówczas udowodnić, że istnieje taka stała $A > 0$, że

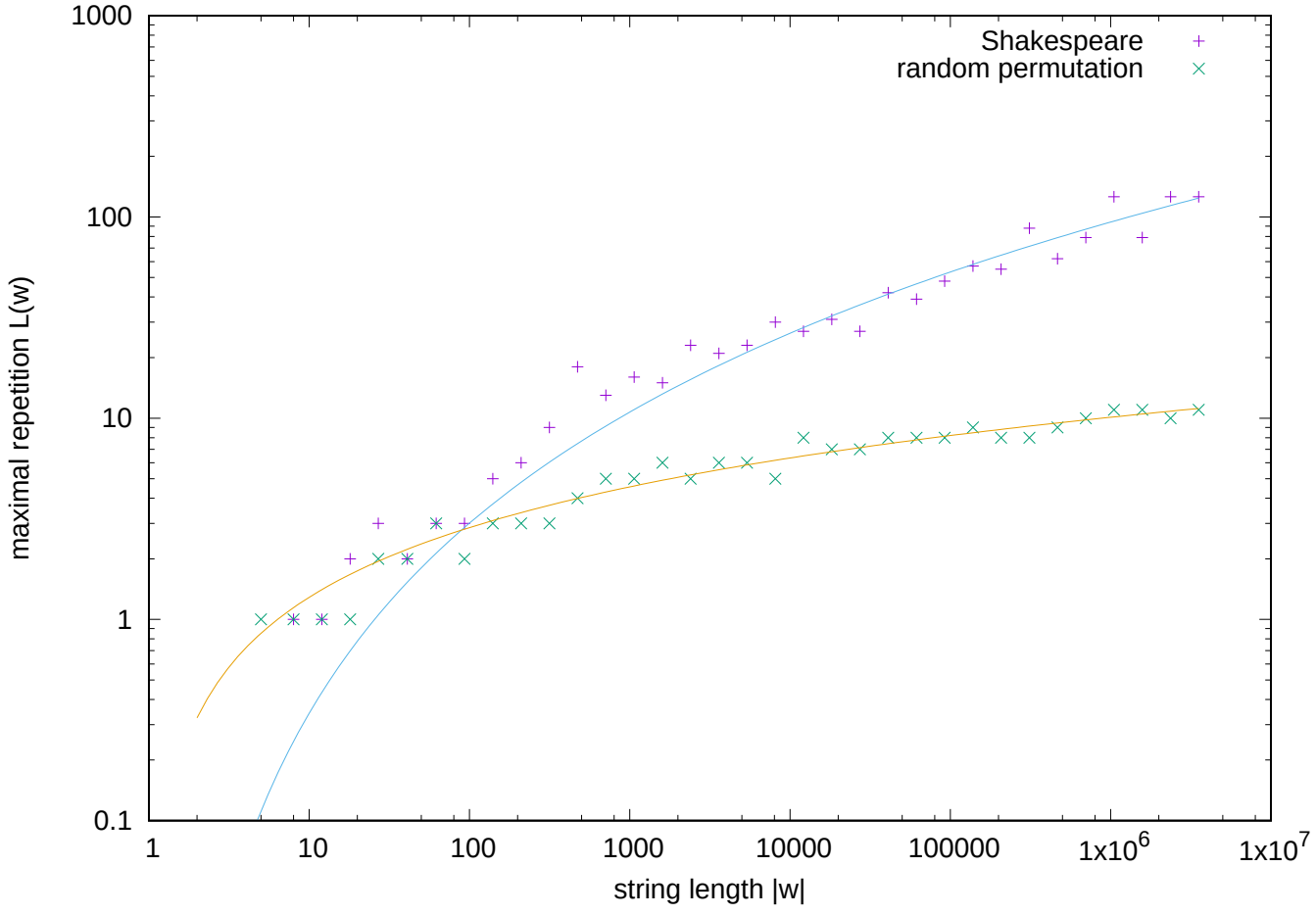
$$L(X_1^n) \leq A \log n$$

dla dostatecznie dużych n z prawdopodobieństwem 1.

Inaczej pisząc,

$$P\left(\limsup_{n\to\infty} \frac{L(X_1^n)}{\log n} \leq A\right) = 1.$$

A w odniesieniu do języka... (Dębowski, 2015)



$L(x_1^n) \approx 0.02498\,(\log n)^{3.136}$ dla tekstu w języku angielskim.

$L(x_1^n) \approx 0.4936\,(\log n)^{1.150}$ dla losowej permutacji znaków.

Kod PPM (Prediction by Partial Matching)

Definiujemy

$$\mathbf{PPM}_k(x_i|x_1^{i-1}) := \begin{cases} \dfrac{1}{D}, & i \leq k, \\ \dfrac{N(x_{i-k}^{i}|x_1^{i-1}) + 1}{N(x_{i-k}^{i-1}|x_1^{i-2}) + D}, & i > k, \end{cases}$$

$$\mathbf{PPM}_k(x_1^n) := \prod_{i=1}^{n} \mathbf{PPM}_k(x_i|x_1^{i-1}),$$

$$\mathbf{PPM}(x_1^n) := \frac{6}{\pi^2} \sum_{k=-1}^{\infty} \frac{\mathbf{PPM}_k(x_1^n)}{(k+2)^2}.$$

Wielkość $\mathbf{PPM}(x_1^n)$ nazywa się p-stwem PPM napisu x_1^n.

Zauważmy, że $\mathbf{PPM}_k(x_1^n) = D^{-n}$ dla $k > L(x_1^n)$.

Uniwersalność p-stwa PPM

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

Intensywność entropii: $h = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[-\log P(X_1^n)\right]$

Twierdzenie

P-stwo PPM jest p-stwem uniwersalnym, tzn. zachodzi

$$\mathbb{E}\left[-\log \mathrm{PPM}(X_1^n)\right] \geq H(X_1^n)$$

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[-\log \mathrm{PPM}(X_1^n)\right] = h$$

dla dowolnego procesu stacjonarnego $(X_i)_{i=1}^{\infty}$ o skończ. alfabecie.

Rząd PPM i słownik PPM

- Rząd PPM $G_{\mathrm{PPM}}(x_1^n)$ to najmniejsza liczba G taka, że

  $-\log \mathrm{PPM}_G(x_1^n) \leq -\log \mathrm{PPM}_k(x_1^n)$ dla każdego $k \geq -1$.

- Zbiór wszystkich podsłów długości m w napisie x_1^n to

  $$V(m|x_1^n) := \left\{ y_1^m : x_{t+1}^{t+m} = y_1^m \text{ dla pewnego } 0 \leq t \leq n-m \right\}.$$

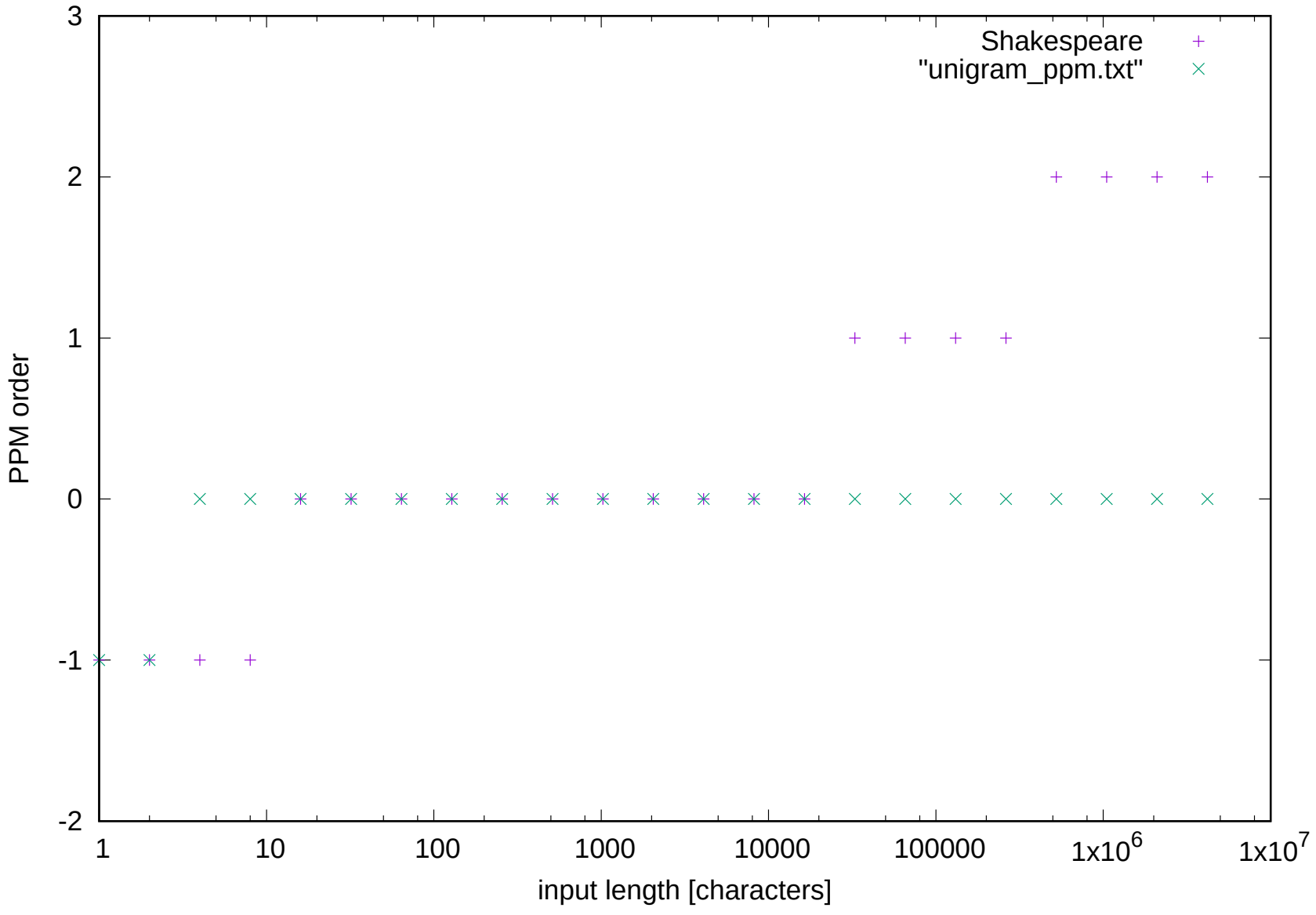
- Zbiór różnych słów PPM w napisie X_1^n to

  $$V_{\mathrm{PPM}}(x_1^n) := V(G_{\mathrm{PPM}}(x_1^n)|x_1^n).$$

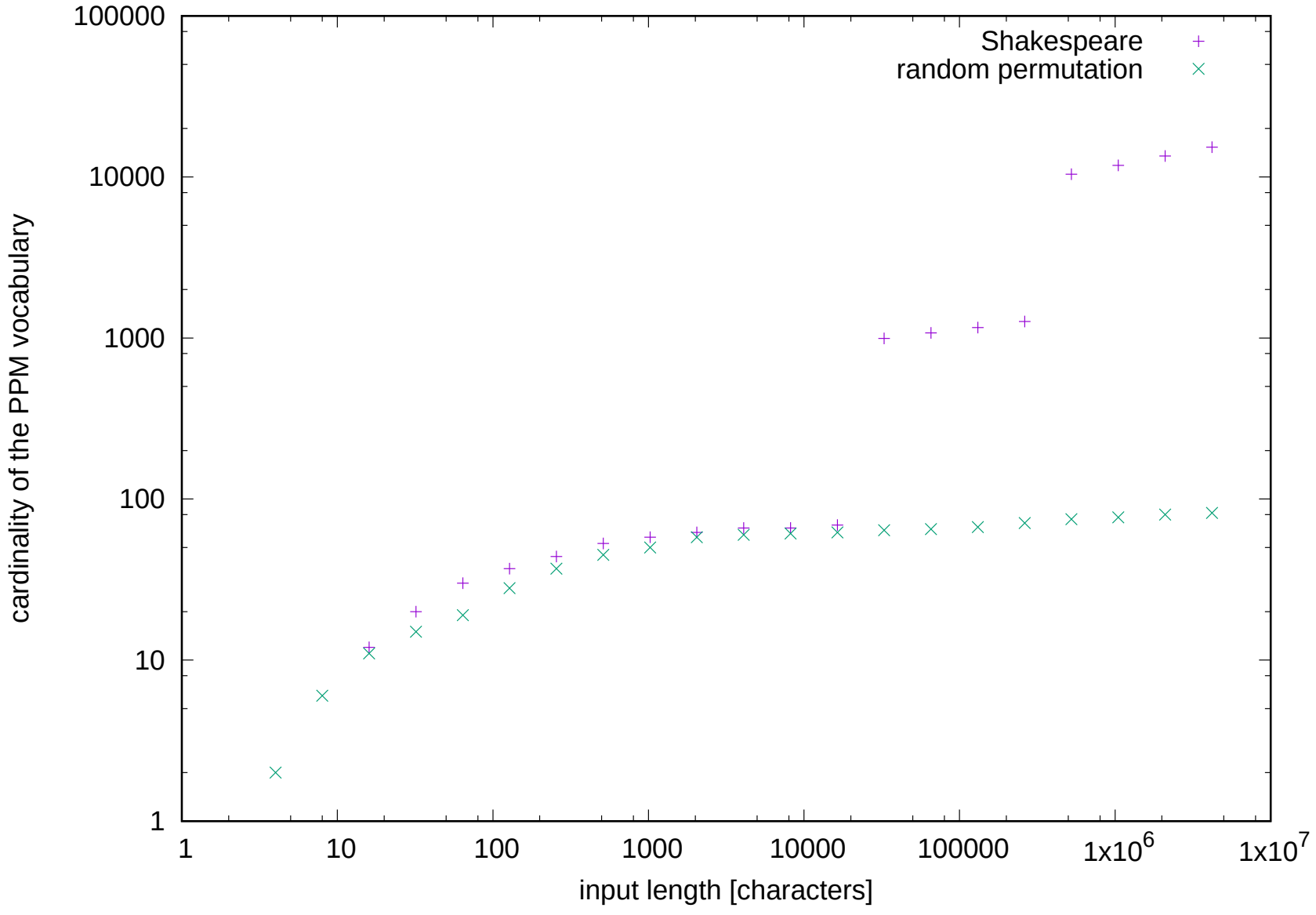
- Ogólnie zachodzi nierówność

  $$\operatorname{card} V_{\mathrm{PPM}}(x_1^n) \leq \min \left\{ D^{G_{\mathrm{PPM}}(x_1^n)}, n - G_{\mathrm{PPM}}(x_1^n) + 1 \right\}.$$

Rząd PPM na wykresie



Moc słownika PPM na wykresie



Statystyczne prawa językowe

Teksty w języku naturalnym spełniają przybliżone prawa ilościowe:

1. **Prawo Zipfa:** częstość słowa jest odwrotnie proporcjonalna do rangi słowa.
2. **Prawo Heapsa:** liczba różnych słów w tekście rośnie potęgowo z długością tekstu.
3. **Intensywność entropii Shannona:** jest rzędu 1 bita na literę.
4. **Hipoteza Hilberga:** entropia warunkowa litery maleje potęgowo z długością kontekstu.
5. **Prawo kodu PPM:** liczba różnych „słów" wykrywanych przez algorytm PPM w tekście rośnie potęgowo z długością tekstu.
6. **Prawo maksymalnego powtórzenia:** długość maksymalnego powtórzenia rośnie jak sześcian logarytmu długości tekstu.

Czy można coś wywnioskować o języku jako procesie stochastycznym na podstawie tych obserwacji/hipotez?

Pytania matematyka

1. Czy istnieje idealny probabilistyczny model języka?
2. Czy model ten może być modelem Markowa?
3. Czy model ten może być ukrytym modelem Markowa?
4. Czy model ten jest ergodyczny?
5. Czy model ten jest stacjonarny?
6. Czy model ten jest asymptotycznie średnio stacjonarny?
7. Czy model ten jest kodem uniwersalnym?
8. Czy model ten jest efektywnie obliczalny?