Uniwersalność kodu PPM i informacja wzajemna

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Instytut Podstaw Informatyki
Polskiej Akademii Nauk

Seminarium ZAMS, 08.01.2018

1. Kod PPM
2. Słowa PPM i informacja wzajemna
3. Język naturalny
4. Konkluzje

1 Kod PPM

2 Słowa PPM i informacja wzajemna

3 Język naturalny

4 Konkluzje

Model Markowa k-tego rzędu

- 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 k ostatnich zmiennych:

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

- Estymacja największej wiarogodności

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

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.

Kod PPM (Prediction by Partial Matching)

Cleary & Witten, 1984; Ryabko, 1984:

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

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

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

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: $h := \lim_{n\to\infty} \frac{H(X_1^n)}{n}$

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.

P-stwo **PPM** a p-stwo $\mathbf{PPM}_k$

Mamy

$$-\log \mathrm{PPM}_G(x_1^n) \le \log \mathrm{PPM}(x_1^n) \le 1/2 + \log(k+2) - \log \mathrm{PPM}_k(x_1^n),$$

gdzie rząd PPM G to najmniejsza liczba taka, że

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

Pomocnicze funkcje

$$\log^* n := \begin{cases} 0, & n = 0 \\ \log n! - n \log n + n, & n \geq 1 \end{cases}$$

$$\mathfrak{H}(n_1, \ldots, n_l) := \begin{cases} \sum_{i=1:n_i>0}^{l} n_i \log \left(\frac{\sum_{j=1}^{l} n_j}{n_i} \right), & \text{jeśli } n_j > 0 \text{ istnieje} \\ 0, & \text{inaczej} \end{cases}$$

$$\mathfrak{K}(n_1, \ldots, n_l) := \sum_{i=1}^{l} \log^* n_i - \log^* \left(\sum_{i=1}^{l} n_i \right)$$

Dwa człony kodu $\mathbf{PPM}_k$

$$-\log \mathrm{PPM}_k(x_1^n) = k \log D - \log \mathrm{PPM}_k^0(x_1^n) - \log \mathrm{PPM}_k^1(x_1^n)$$

gdzie

$$-\log \mathrm{PPM}_k^0(x_1^n) := \sum_{u \in \mathbb{X}^k} \mathfrak{H}\left(N(u1|x_1^n), \ldots, N(uD|x_1^n)\right)$$

(warunkowa entropia rozkładu empirycznego)

$$-\log \mathrm{PPM}_k^1(x_1^n) := \sum_{u \in \mathbb{X}^k} \mathfrak{H}\left(N(u|x_1^{n-1}), D-1\right) - \sum_{u \in \mathbb{X}^k} \mathfrak{K}\left(N(u1|x_1^n), \ldots, N(uD|x_1^n), D-1\right)$$

(poprawka na uniwersalność)

Człon $\mathbf{PPM}_k^0$ a entropia warunkowa

Entropia warunkowa: $H(X_1|X_{-k+1}^0) = \mathbb{E}\left[-\log P(X_1|X_{-k+1}^0)\right]$

Twierdzenie

Dla ergodycznego procesu stacjonarnego $(X_i)_{i=1}^{\infty}$ o skończ. alf.,

$$\lim_{n\to\infty} \frac{1}{n}\left[-\log \mathrm{PPM}_k^0(X_1^n)\right] = H(X_{k+1}|X_1^k) \text{ pnp..}$$

Dowód z twierdzenia ergodycznego.

Człon $\mathbf{PPM}_k^1$ a subword complexity

Oznaczmy zbiór wszystkich podsłów długości m w napisie x_1^n jako

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

Ogólnie zachodzi nierówność

$$\operatorname{card} V(k|x_1^n) \leq \min\left\{ D^k, n-k+1 \right\}.$$

(subword complexity)

Twierdzenie

$$0 \leq -\log \mathrm{PPM}_k^1(x_1^n) < D \operatorname{card} V(k|x_1^{n-1}) \, (2 + \log n)$$

Uniwersalność raz jeszcze

Mamy

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

Z poprzednich rozważań wynika następujący wynik:

Twierdzenie

Dla ergodycznego procesu stacjonarnego $(X_i)_{i=1}^{\infty}$ o skończ. alf.,

$$\limsup_{n\to\infty} \frac{1}{n}\left[-\log \mathrm{PPM}(X_1^n)\right] \leq h \text{ pnp.},$$

$$-\log \mathrm{PPM}(X_1^n) \leq 1/2 + n \log D.$$

Wniosek z rozkładu ergodycznego i twierdzenia SMB

Twierdzenie

Niech Q będzie kodem, tzn. $Q(x_1^n) \geq 0$ i $\sum_{x_1^n} Q(x_1^n) \leq 1$.

Jeżeli dla ergodycznych procesów stacjonarnych $(X_i)_{i=1}^{\infty}$ o sk. alf.,

$$\limsup_{n\to\infty} \frac{1}{n}\left[-\log Q(X_1^n)\right] \leq h \text{ pnp.},$$

$$-\log Q(X_1^n) \leq cn,$$

to dla dowolnego procesu stacjonarnego $(X_i)_{i=1}^{\infty}$ o sk. alf.,

$$\lim_{n\to\infty} \frac{1}{n}\left[-\log Q(X_1^n)\right] = h_F \text{ pnp.},$$

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

1 Kod PPM

2 Słowa PPM i informacja wzajemna

3 Język naturalny

4 Konkluzje

Rząd PPM i słownik PPM

- Rząd PPM G to najmniejsza liczba taka, że

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

- Zbiór różnych słów PPM w napisie x_1^n to zbiór wszystkich podsłów długości G,

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

Informacja wzajemna PPM i słowa PPM

$$\mathbb{I}_Q(u; v) := -\log Q(u) - \log Q(v) + \log Q(uv)$$

Twierdzenie

$$\mathbb{I}_{\mathrm{PPM}_k^0}(x_1^n; x_{n+1}^{n+m}) \leq 0$$

$$\begin{aligned}\mathbb{I}_{\mathrm{PPM}}(x_1^n; x_{n+1}^{n+m}) \leq 1 &+ 4 \log \left[G_{\mathrm{PPM}}(x_1^{n+m}) + 2\right] \\ &+ \left[G_{\mathrm{PPM}}(x_1^{n+m}) + 1\right] \log D \\ &+ 2D \operatorname{card} V_{\mathrm{PPM}}(x_1^{n+m}) \left[2 + \log(n+m)\right]\end{aligned}$$

Wykładnik Hilberga

Wykładnik Hilberga definiujemy jako

$$\operatorname*{hilb}_{n\to\infty} s(n) := \limsup_{n\to\infty} \frac{\log^+ s(2^n)}{\log 2^n}, \quad \log^+ x = \begin{cases} \log(x+1), & x \geq 0, \\ 0, & x < 0. \end{cases}$$

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

Twierdzenie

Niech $\mathfrak{J}(n) := 2\mathfrak{G}(n) - \mathfrak{G}(2n)$. Jeżeli $\lim_{n\to\infty} \mathfrak{G}(n)/n = \mathfrak{g}$, to

$$\operatorname*{hilb}_{n\to\infty} (\mathfrak{G}(n) - n\mathfrak{g}) \leq \operatorname*{hilb}_{n\to\infty} \mathfrak{J}(n),$$

gdzie równość zachodzi, gdy $\mathfrak{J}(n) \geq 0$.

Twierdzenie o słowach i informacji wzajemnej

Informacja wzajemna Shannona dla procesu stacjonarnego:

$$\begin{aligned} I(X_1^n; X_{n+1}^{2n}) &:= H(X_1^n) + H(X_{n+1}^{2n}) - H(X_1^{2n}) \\ &= 2H(X_1^n) - H(X_1^{2n}) \end{aligned}$$

Twierdzenie (informacja wzajemna i słowa PPM)

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

$$\begin{aligned} \operatorname*{hilb}_{n\to\infty} I(X_1^n; X_{n+1}^{2n}) &\leq \operatorname*{hilb}_{n\to\infty} \mathbb{E}\, \mathbb{I}_{\mathrm{PPM}}(X_1^n; X_{n+1}^{2n}) \\ &\leq \operatorname*{hilb}_{n\to\infty} \mathbb{E}\left[G_{\mathrm{PPM}}(X_1^n) + \operatorname{card} V_{\mathrm{PPM}}(X_1^n)\right]. \end{aligned}$$

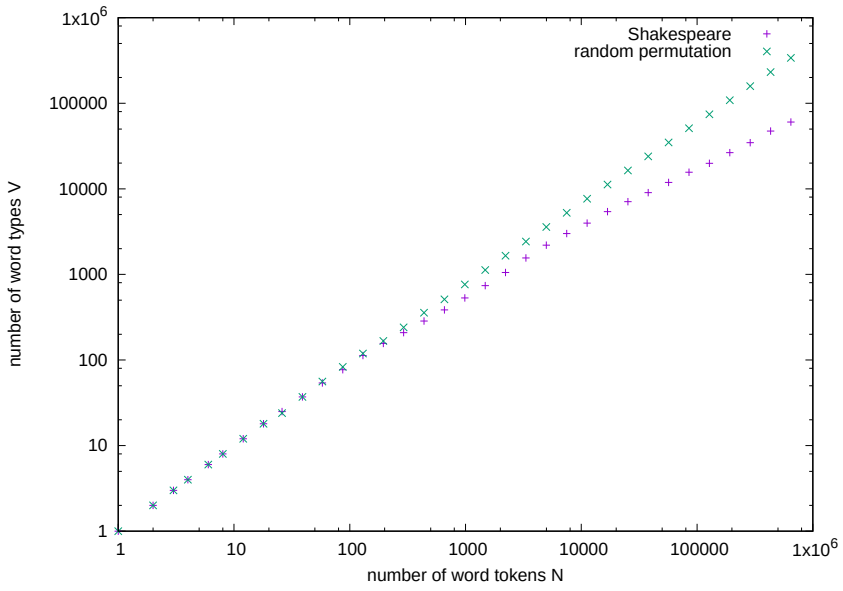
1 Kod PPM

2 Słowa PPM i informacja wzajemna

3 Język naturalny

4 Konkluzje

Prawo Heapsa (liczba różnych słów ortograficznych)



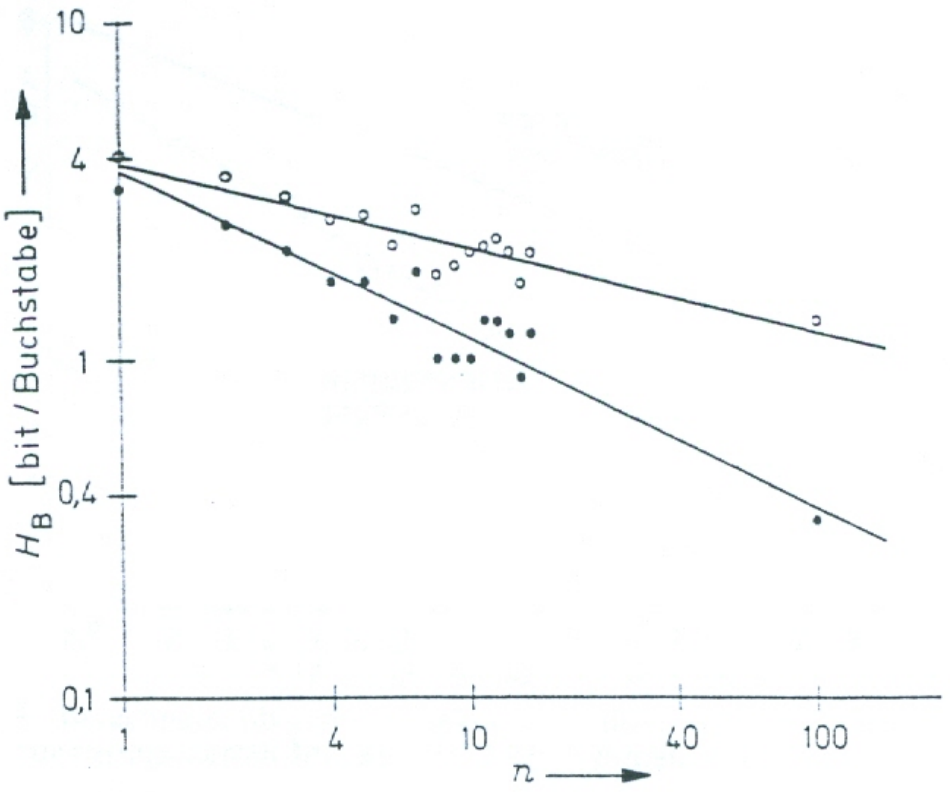
Kuraszkiewicz i Łukaszewicz 1951; Herdan 1964; Heaps 1978:

$$V \propto N^{\gamma}, \quad \gamma < 1$$

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

Hipoteza Hilberga dla entropii Shannona (1990)

$$H(X) := \mathbb{E}\left[-\log P(X)\right] \quad H(X|Y) := \mathbb{E}\left[-\log P(X|Y)\right]$$



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

Prawa potęgowe dla entropii i informacji wzajemnej

Hipoteza Hilberga:

$$H(X_n|X_1^{n-1}) \approx Bn^{\beta-1} + h$$

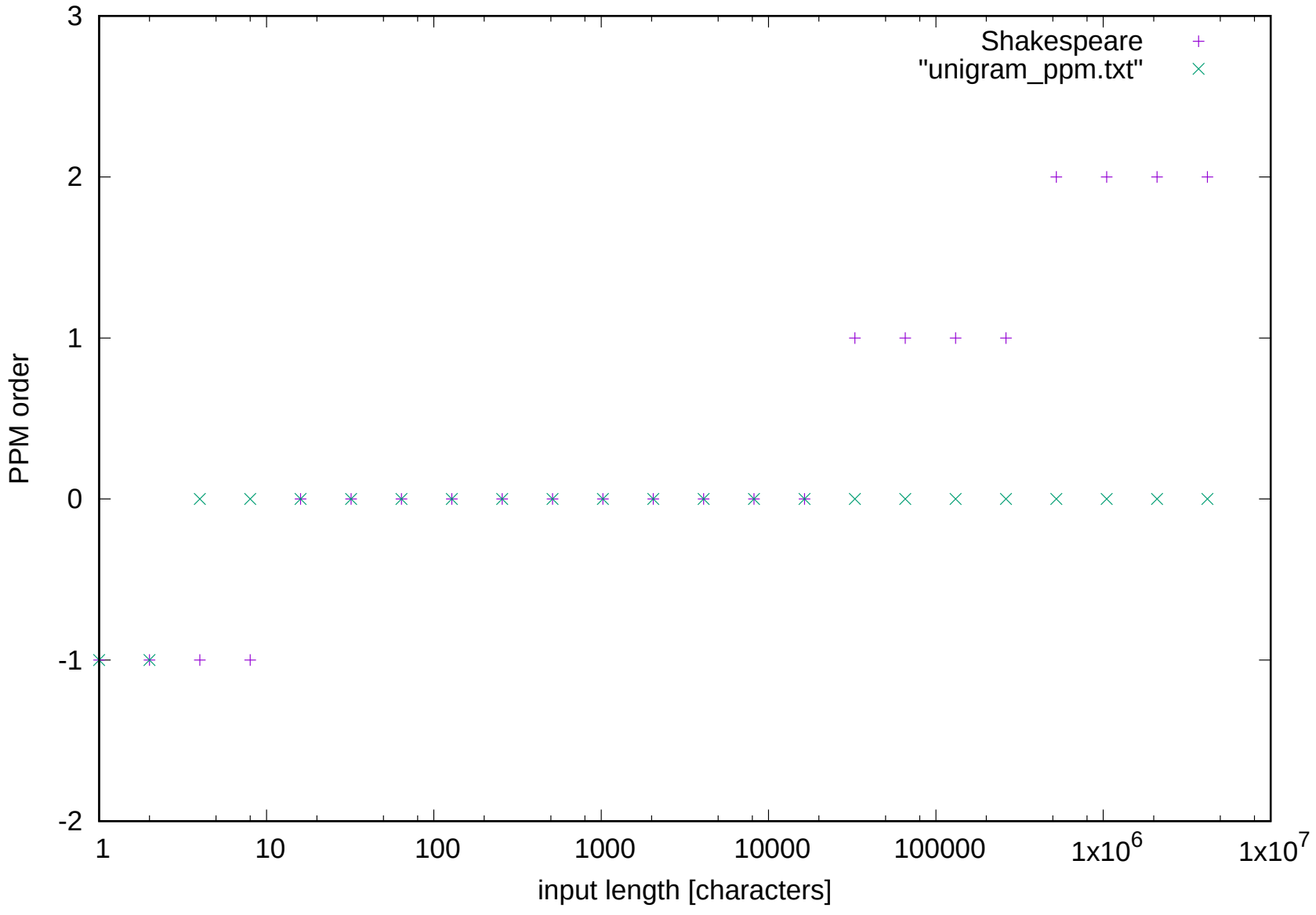
$$\Downarrow$$

$$H(X_1^n) = \sum_{i=1}^{n} H(X_i|X_1^{i-1}) \approx B'n^{\beta} + hn$$

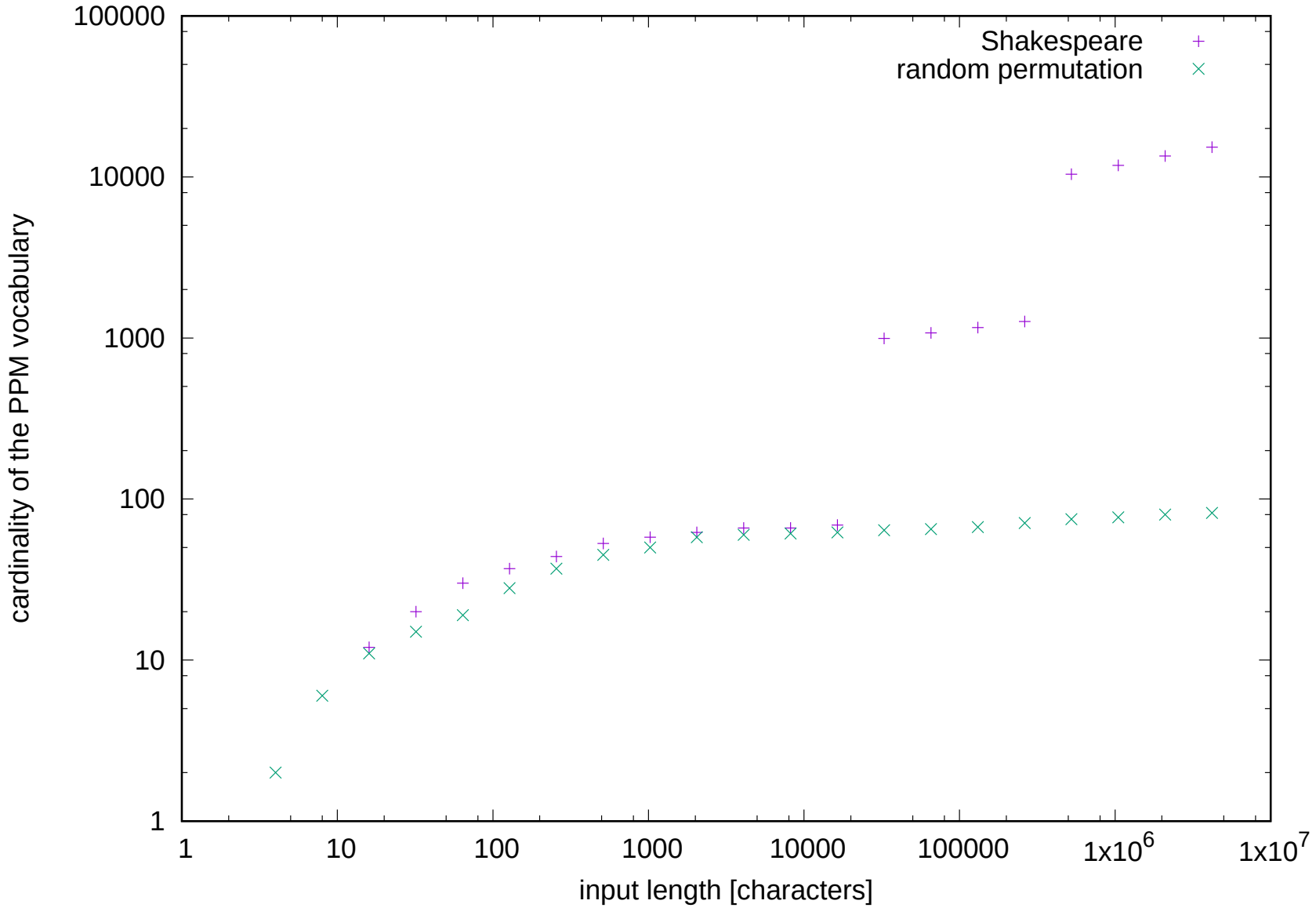
$$\Downarrow$$

$$I(X_1^n; X_{n+1}^{2n}) \approx B''n^{\beta}$$

Rząd PPM na wykresie



Moc słownika PPM na wykresie



Wykładnik Hilberga dla mocy słownika PPM

Dla procesów Markowa k-tego rzędu nad skończonym alfabetem

$$\lim_{n\to\infty} G_{\mathrm{PPM}}(X_1^n) \leq k, \text{ czyli } \operatorname{hilb}_{n\to\infty} \operatorname{card} V_{\mathrm{PPM}}(X_1^n) = 0.$$

Dla języka naturalnego mamy prawdopodobnie

$$\operatorname{hilb}_{n\to\infty} \operatorname{card} V_{\mathrm{PPM}}(X_1^n) = \beta, \quad \beta \in (0,1).$$

Czyli język naturalny nie jest procesem Markowa.

1 Kod PPM

2 Słowa PPM i informacja wzajemna

3 Język naturalny

4 Konkluzje

Konkluzje

- Kod PPM jest praktycznym algorytmem kompresji danych, ale ma też ciekawe własności teoretyczne:
  - Może służyć jako test niemarkowskości procesu stochastycznego.
  - Odwołując się do niego można udowodnić twierdzenie Shannona-McMillana-Breimana dla procesów nieergodycznych i asymptotycznie średnio stacjonarnych (AMS).
- Możliwe jest, że język naturalny spełnia hipotezę Hilberga.