Statistiques d'occurrences de mots

Régnier and Szpankowski

On pattern frequency occurrences in a markovian sequence?

Algorithmica, 1998

Régnier

A unified approach to word occurrences probabilities.

Discrete Applied Mathematics, 2000

Pierre Nicodème, Nov. 2001

Séries Génératrices

$f_0, f_1, f_2, \ldots, f_n, \ldots \quad \rightarrow \quad F(z) = \sum f_n z^n$

$1, 1, \ldots, 1, \ldots$

$\rightarrow F(z) = 1 + z + z^2 + \cdots + z^n + \cdots = \frac{1}{1-z}$

$1, 2, 4, \ldots, 2^n, \ldots$

$\rightarrow F(z) = 1 + 2z + 4z^2 + \cdots + 2^n z^n + \cdots = \frac{1}{1-2z}$

$F(z) \rightarrow f_n = [z^n]F(z)$, nth coefficient de Taylor

par analyse de singularité ou calcul formel

$[z^n]\frac{1}{1-2z} = 2^n$

Série génératrice d'un langage

$\mathcal{L} = \subset \Sigma^*\}$ ensemble de mots

$\rightarrow \quad F_{\mathcal{L}}(l_1, l_2, \ldots, l_p) = \sum_{w \in \mathcal{L}} \text{commute}(w)$

$(aabaa)^\star = \epsilon + aabaa + (aabaa)^2 + (aabaa)^3 + \cdots$

$\mathcal{L} = (aabaa)^\star \cup ccc \quad \rightarrow F_{\mathcal{L}}(a, b, c) = \frac{1}{1-a^4b} + c^3$

Si $A.B$ non-ambigu,

$F_{A \cdot B}(l_1, \ldots, l_p) = F_A(l_1, \ldots, l_p) \cdot F_B(l_1, \ldots, l_p)$

Si A et B disjoints

$F_{A \cup B}(l_1, \ldots, l_p) = F_A(l_1, \ldots, l_p) + F_B(l_1, \ldots, l_p)$

Langages et Séries génératrices

$\mathcal{L} \subset \Sigma^*$ $\quad (\Sigma = \{l_1, l_1, \ldots, l_q\}$ alphabet, $\quad \omega_i = \mathbf{P}(l_i))$

Série génératrice du languages

$M(l_1, l_2, \ldots, l_q) = \sum_{\alpha \in \mathcal{L}} \text{commute}(\alpha)$

Série génératrice de probabilité

$$\begin{aligned} F(z) &= M(\omega_1 z, \omega_2 z, \ldots, \omega_q z) \\ &= \sum_{\alpha \in \mathcal{L}} p_\alpha z^{|\alpha|} = \sum \pi_n z^n \end{aligned}$$

p_α proba. du mot α

π_n proba. qu'un mot taille n appartienne à $\mathcal{L}$

Exemple

$\Sigma = \{a, b\} \quad \omega_a = 1/3, \; \omega_b = 2/3 \quad (\epsilon$ mot vide$)$

$\mathcal{L} = \{\epsilon, aa, ab, ba, aaab\}$

$$\Rightarrow \begin{cases} M(a, b) = 1 + a^2 + 2ab + a^3 b \\ F(z) = M\left(\frac{z}{3}, \frac{2z}{3}\right) = 1 + 0z + \frac{5}{9} z^2 + 0z^3 + \frac{2}{81} z^4 \end{cases}$$

1 mot - Bernoulli

h mot considéré (longueur k, proba $p_h = p$)

Polynome d'autocorrélation

$h = ababa$

$$
\begin{array}{l}
ababa \\
\quad ababa \\
\qquad ababa
\end{array}
$$

$\mathcal{A} = \epsilon + ba + baba$

$A(z) = 1 + \omega_a \omega_b z^2 + \omega_a^2 \omega_b^2 z^4$

$\mathcal{A} = \{\ w, \quad h.w = u.h \quad \text{et} \quad |w| < |h|\ \}$

Langages utilisés

First $\mathcal{F} = \{\ w = u.h \quad \text{et} \quad \nexists r, s,\ w = r.h.s\ \}$

$aaaaaababa \subset \mathcal{F}, \quad bbbbbabababa \not\subset \mathcal{F}$

Minimal $\mathcal{M}$

$= \{\ w, \quad h.w = u.h \ \ \text{et} \ \ \nexists r, s,\ h.w = r.h.s\ \}$

$ababa$ $ababa$

$aaaaababa \ \subset \mathcal{M}$ $babbbbbbbbababa \ \not\subset \mathcal{M}$

$ababa$

$ba \subset \mathcal{M}$

Tail $\mathcal{T} = \{w, \quad \nexists r, s,\ h.w = r.h.s\}$

$ababa$ $ababa$

$aabbbabbbbbb \subset \mathcal{T}$ $babbbbbbbbb \not\subset \mathcal{T}$

Not $\mathcal{N} = \Sigma^\star - \Sigma^\star.h.\Sigma^\star = \{w,\ \nexists r, s,\ w = r.h.s\}$

$\Sigma^\star = \mathcal{N} + \mathcal{F}.(\mathcal{M})^\star.\mathcal{T} \ \Rightarrow\ \mathcal{L}_u = \mathcal{N} + \mathcal{F}u.(\mathcal{M}u)^\star.\mathcal{T}$

Equations sur les langages

(I) $\Sigma^\star = \mathcal{T} + \mathcal{M}\Sigma^\star$

$$h.\Sigma^\star = h.\mathcal{T} + h.\mathcal{M}\Sigma^\star$$

mot commencant par h

- 1 seule occurrence de $h \quad \Rightarrow \quad h.\mathcal{T}$
- plusieurs occurrences de $h \quad \Rightarrow \quad h.\mathcal{M}.\Sigma^\star$

(II) $\Sigma^\star h = \mathcal{F}.\mathcal{A} + \mathcal{F}.\Sigma^\star.h$

(remarque $\epsilon \subset \mathcal{A}$)

mot finissant par h

- première occurrence h chevauche dernière (ou seule occurrence) $\quad \Rightarrow \mathcal{F}.\mathcal{A}$
- première occurrence ne chevauche pas dernière $\quad \Rightarrow \mathcal{F}.\Sigma^\star.h$

Equations sur les langages (suite)

(III) $\mathcal{M}^+ = \Sigma^\star.h + \mathcal{A} - \epsilon$

$h.\mathcal{M}^+ = h.\Sigma^\star.h + h.(\mathcal{A} - \epsilon)$

- $h.\mathcal{M}^+$, $\Rightarrow$ mots commençant et finissant par h
- $h.\Sigma^\star.h$, $\Rightarrow$ 1ere et dernière occurrence non recouvrantes
- $h.(\mathcal{A} - \epsilon)$ $\Rightarrow$ 1ere et dernière occurrence recouvrantes

(IV) $\mathcal{N}.\sigma = \mathcal{F} + \mathcal{N} - \epsilon$

ajouter une lettre a mot de N $\Rightarrow$

- crée un match $\Rightarrow \mathcal{F}$
- ne crée pas de match $\Rightarrow \mathcal{N}$

(mot vide ϵ interdit)

Langages $\Rightarrow$ séries génératrices

$$\begin{array}{llcl}
\text{(I)} & \Sigma^\star = \mathcal{T} + \mathcal{M}\Sigma^\star & \Rightarrow & \frac{1}{1-z} = T(z) + \frac{M(z)}{1-z} \\
\text{(II)} & \Sigma^\star h = \mathcal{F}.\mathcal{A} + \mathcal{F}.\Sigma^\star.h & \Rightarrow & \frac{p_h z^k}{1-z} = F(z)\left(A(z) + \frac{p_h z^k}{1-z}\right) \\
\text{(III)} & \mathcal{M}^+ = \Sigma^\star.h + \mathcal{A} - \epsilon & \Rightarrow & \frac{M(z)}{1-M(z)} = \frac{p_h z^k}{1-z} + A(z) - 1 \\
\text{(IV)} & \mathcal{N}.\Sigma = \mathcal{F} + \mathcal{N} - \epsilon & \Rightarrow & zN(z) = F(z) + N(z) - 1
\end{array}$$

Ajouter la marque u après chaque match $\Rightarrow$ langage marqué $\mathcal{L}$

$$\mathcal{L} = \mathcal{N} + \mathcal{F}.u(\mathcal{M}.u)^\star\mathcal{T}$$

Traduction en série génératrice

$$L(z,u) = N(z) + \frac{uF(z)T(z)}{1 - uM(z)}$$

Solving $\Rightarrow$

$$L(z,u) = \frac{1}{1 - z + pz^k \frac{1-u}{u+(1-u)A(z)}}$$

1 mot - Markov stationnaire

mot $h = h_1 h_2 \cdots h_k$

$$p = P(h) = \pi_{h_1} \omega_{h_1,h_2} \ldots \omega_{h_{k-1},h_k}$$

$$\text{Matrice Markov } \mathbb{P} = \begin{pmatrix} \omega_{a,a} & \omega_{a,b} \\ \omega_{b,a} & \omega_{b,b} \end{pmatrix}$$

$$\text{Matrice stationnaire } \Pi = \lim_{n \to \infty} \mathbb{P}^n = \begin{pmatrix} \pi_a & \pi_b \\ \pi_a & \pi_b \end{pmatrix}$$

$$\text{distribution stationnaire } \pi = (\pi_a, \pi_b)$$

$$\pi \mathbb{P} = \pi \quad \Rightarrow \quad \Pi \times \mathbb{P} = \Pi$$

$P(w_1|w_2)$ probabilité d'apparition de w_1, conditionnellement à w_1 précédé par w_2

Séries génératrices d'un langage $\mathcal{B}$

non-conditionné	conditionné par h
$B(z) = \sum_{w \in \mathcal{B}} P(w) z^{\lvert w \rvert}$	$B_{[h]}(z) = \sum_{w \in \mathcal{B}} P(w\lvert h) z^{\lvert w \rvert}$

Equations de langage

(I) $\Sigma^\star(z) = \Sigma^\star_{[h]}(z) = \dfrac{1}{1-z}$

$$\Sigma^\star_{[h]}(z) = \sum_{n=0}^{\infty} \sum_{|w|=n} P(w|h) z^n = 1 + \sum_{n=1}^{\infty} \sum_{\alpha\in\Sigma} \mathbb{P}^n_{h_k,\alpha} = 1 + \sum_{n=1}^{\infty} z^n$$

(II) $\mathcal{C} = \Sigma^\star h \quad \Rightarrow \quad C_{[h]}(z) = \left(\dfrac{1}{1-z} + \mathbb{K}(z)_{h_k,h_1}\right) \times p z^k$

$$\text{où} \quad \mathbb{K}(z) = \frac{1}{\pi_{h_1}} \times \frac{\mathbb{P}-\Pi}{\mathbb{I} - z(\mathbb{P}-\Pi)}$$

ceci vient de

$$C_{[h]}(z) = \sum_{n\geq 1} \mathbb{P}^n_{h_k,h_1} z^n \times \frac{p}{\pi_{h_1}} z^{k-1}$$

et

$$(1) \quad \mathbb{P}^n = \Pi^n + (\mathbb{P}-\Pi)^n$$

$$(2) \quad \Pi^n_{h_k,h_1} = \Pi_{h_k,h_1} = \pi_{h_1}$$

(III) $\mathcal{F}_{|h|}(z) = \mathcal{F}(z)$ (IV) $\mathcal{M}_{|h|}(z) = \mathcal{M}(z)$

(V) $\mathcal{T}_{|h|}(z) = \mathcal{T}(z)$

1 mot - Markov 1

Solution du système

$$L(z,u) = \frac{1 + pz^k \frac{(1-u)}{u+(1-u)A(z)} \kappa(z)}{1 - z + pz^k \frac{1-u}{u+(1-u)A(z)} \left(1 + (1-z)\kappa(z)\right)}$$

avec $\kappa(z) = \mathbb{K}(z)_{h_m,h_1}$

$$= \frac{1}{\pi_{h_1}} \left((\mathbb{P} - \Pi)(\mathbb{I} - z(\mathbb{P} - \Pi))^{-1}\right)_{h_m,h_1}$$

"Multi"- mots

$H = \{h_1, h_2, \ldots, h_m\}$

$h_i = h_{i,1} h_{i,2} \ldots h_{i,l_i}$

$\mathcal{A}_{i,j} = \{w,\ h_i.w = u.h_j, \text{ et } \nexists r, s, t \quad h_i.w = s.h_r.t\}$

$\Rightarrow \mathbb{A}(z) = (A_{i,j}(z))$

Exemple $\quad H = \{aab, abaa\}$

$$\mathbb{A}(z) = \begin{pmatrix} 1 & 0 \\ \pi_b z & 1 + \pi_a^2 \pi_b z^3 \end{pmatrix}$$

$\mathcal{F}_i, \mathcal{M}_{i,j}, \mathcal{T}_i \;\Rightarrow\; F_i(z), M_{i,j}(z), T_i(z)$

$\Rightarrow \mathbf{F}(z, u_1, \ldots, u_m), \mathbb{M}(z, u_1, u_2, \ldots, u_m), \mathbf{T}(z)$

$H, \mathbb{P}, \Pi \quad \Rightarrow \quad \mathbb{K}(z)$

Séries génératrices multivariées

$$L(z, u_1, \ldots, u_k)$$

$$= N(z) + \mathbf{F}^t(z, u_1, \ldots, u_k) \frac{\mathbb{M}(z, u_1, \ldots, u_k)}{\mathbb{I} - \mathbb{M}(z, u_1, \ldots, u_k)} \mathbf{T}(z)$$

$$\mathbf{Cov}(N_{h_i}(n), N_{h_j}(n))$$

$$= [z^n] \left(\frac{\partial^2 L}{\partial u_i \partial u_j} - \frac{\partial L}{\partial u_i} \frac{\partial L}{\partial u_j} \right) (z, 1, \ldots, 1)$$

1 mot - Lois asymptotiques

$$\mathcal{L}^{(r)} = \mathcal{F}.\mathcal{M}^{r-1}.\mathcal{T}$$

$$L^{(r)}(z) =$$

$$\frac{pz^k\left(\big(1-z\big)A(z) + pz^k\Big(1+\big(1-z\big)\kappa(z)\Big)+z-1\right)^{r-1}}{\left(\big(1-z\big)A(z) + pz^k\Big(1+\big(1-z\big)\kappa(z)\Big)\right)^{r+1}}$$

$$\phi_\kappa(z) = \big(1-z\big)A(z) + pz^k\Big(1+\big(1-z\big)\kappa(z)\Big)$$

$$\phi(z) = \big(1-z\big)A(z) + pz^k$$

Analyse de singularité

$$\phi_\kappa(z) = (1-z)A(z)+pz^k\Big(1+(1-z)\kappa(z)\Big) = \frac{1-z}{1-M(z)}$$

$$\alpha(z) = \frac{1}{1-M(z)} \quad \text{série génératrice de} \quad \mathcal{M}^\star$$

$\Rightarrow \quad \alpha(z)$ converge pour $z < 1$

$$\left(\text{“dominée” par } \frac{1}{1-z}\right)$$

$\phi_\kappa(1) = p \neq 0 \Rightarrow z = 1$ pôle simple de $\dfrac{1}{1-M(z)}$

$\alpha(z)$ coeff. positifs $\Rightarrow$ pas autres poles module 1

$\phi(z)$ polynome $\Rightarrow \exists \rho > 1,\ \phi(\rho) = 0$ (racine simple)

Rouché $\Rightarrow \exists \rho_\kappa > 1, \quad \phi_\kappa(\rho_\kappa) = 0$

Poisson pour occurrences rares

analyse de singularité

$$\Rightarrow \mathbf{P}\big(N_n(h) = r\big) = \frac{1}{\rho^{n+1}} \frac{1}{r!} \left(\frac{n}{\rho}\right)^r \left(1 + O\left(\frac{1}{n}\right)\right)$$

Loi centrale limite

$$\mathbf{E}\left(e^{tX_n}\right) \to e^{t^2/2} \quad \Rightarrow \quad X_n \to \mathcal{N}(0,1)$$

convergence via la transformée de Laplace

$$L(z,u) = \sum_{n\geq 0} L_n(u) z^n = N(z) + \frac{upz^k}{\phi^2(z)\big(1-uM(z)\big)}$$

$$\begin{aligned} L_n(u) &= \frac{1}{2\pi i}\oint \frac{L(z,u)}{z^{n+1}} \\ &\asymp \frac{1}{2\pi i}\oint \frac{up}{\phi^2(z)\big(1-uM(z)\big)z^{n+1-k}} \end{aligned}$$

pôle, racine de $1-uM(z)=0$

substitutions: $u=e^t$, $z=e^\xi$ analytique

$1-e^tM\left(e^\xi\right)=0 \Rightarrow$ solution $\xi(t)$

$$\Rightarrow L_n\left(e^t\right) = C(t)e^{-(n+1-k)\xi(t)} + O\left(R^{-n}\right)$$

$\xi(t) = t\xi'(0) + \frac{t^2}{2}\xi''(0) + O(t^3)$

$\mu_n \asymp -n\xi'(0), \quad \sigma_n^2 \asymp -n\xi''(0), \quad$ subst. $t = \dfrac{\tau}{\sigma_n} \to 0$

$$\Rightarrow \; e^{-\frac{\tau\mu_n}{\sigma_n}} L_n\left(e^{\frac{\tau}{\sigma_n}}\right) = e^{\frac{\tau^2}{2}}\left(1+O\left(\frac{1}{\sqrt{n}}\right)\right)$$

Large deviations

$r = (1+\delta)\mu_n$

$$\lim_{n\to\infty} \frac{\log L_n\left(e^t\right)}{n} = \xi(t)$$

Gärtner-Ellis $\Rightarrow \displaystyle\lim_{n\to\infty} \frac{\log\left(\mathbf{P}(N_n(h) > na)\right)}{n} = -I(a)$

$I(a) = a\beta_a + \xi(\beta_a)$

et β_a vérifie $\xi'(\beta_a) = -a$

Méthode par shift de moyenne

$$\Lambda(e^t) = \frac{L_n(e^{t+v})}{L_n(e^v)}$$

v calculé pour que $\mu_\Lambda = (1+\delta)\mu_n$

$$\mathbf{P}\left(N_n(h) = (1+\delta)\mu_n\right)$$
$$= \frac{1}{\sigma_a\sqrt{2\pi(n-k+1)}} e^{-(n-k+1)I(a)}\left(1 + O\left(\frac{1}{n}\right)\right)$$

avec $\sigma_a^2 = -\xi''(\beta_a)$