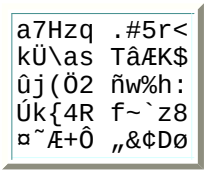
Kryptologie

Buchstabenhäufigkeiten stochastischer Sprachen



Einleitung

Es ist eine empirische Beobachtung, dass der Koinzidenzindex $\kappa(a,b)$ für »natürliche« Sprachen fast eine Konstante ist.

Z. B. ist $\kappa(a,b) \approx 0.0762$ für gleich lange deutsche Texte, wenn deren Länge r genügend groß ist (etwa > 100).

Der entsprechende Wert für Englisch ist 0.0661.

Welche statistischen Eigenschaften »natürlicher« Sprachen liegen derartigen Phänomenen zu Grunde?

Bei der statistischen Kryptoanalyse der monoalphabetischen Substitution wurde die (empirisch genügend gesicherte) Annahme verwendet, dass die durchschnittliche Häufigkeit des Buchstabens $s \in \Sigma$ in genügend langen Texten einer solchen Sprache nahe bei einem Wert p_s liegt. Dies gilt auch, wenn man nur eine beliebige feste Stelle j betrachtet, zumindest für die meisten j - die Anfangsbuchstaben von Texten haben durchaus abweichende Häufigkeiten.

Sei also $M \subseteq \Sigma^*$ eine Sprache, $M_r = M \cap \Sigma^r$ für $r \in \mathbf{N}$.

Die durchschnittliche Häufigkeit von $s \in \Sigma$ an der Stelle $j \in [0 \ldots r\text{-}1]$ für Texte in M_r ist

$$\mu_{sj}^{(r)} := \frac{1}{\#M_r} \cdot \sum_{a \in M_r} \delta_{sa_j}$$

(Die Summe zählt die $a \in M_r$, an deren j-ter Stelle s steht.)

Beispiel

Sei $M = \Sigma^*$. Dann ist

$$\mu_{sj}^{(r)} = \frac{1}{n^r} \cdot \sum_{a \in \Sigma^r} \delta_{sa_j} = \frac{1}{n} \quad \text{für alle } s \in \Sigma, j = 1, \ldots, r-1$$

(Es gibt genau n^{r-1} mögliche Texte, wenn $a_j = s$ festgehalten wird.)

Definition

Die Sprache $M \subseteq \Sigma^*$ heißt **stochastisch**, wenn es eine endliche Ausnahmemenge $J \subseteq \mathbf{N}$ gibt, so dass für alle $j \in \mathbf{N}\text{-}J$ und alle $s \in \Sigma$

$$p_s := \lim_{r \to \infty} \mu_{sj}^{(r)}$$

gleichmäßig in j existiert und unabhängig von j ist.

Die p_S heißen die **Buchstabenhäufigkeiten** von M.

Bemerkungen und Beispiele. 1.) Die Ausnahmemenge J besteht bei »natürlichen« Sprachen meist nur aus der Stelle 0; d. h. an der Anfangsposition von Texten darf die Buchstabenverteilung abweichen. Z. B. ist im Deutschen das »e« nicht der häufigste Anfangsbuchstabe eines Textes.

2.) Die Sprache $M = \Sigma^*$ ist stochastisch.

3.) Da stets $\Sigma_{s \in \Sigma} \mu_{sj}^{(r)} = 1$, folgt $\Sigma_{s \in \Sigma} p_s = 1$.

Achtung. Diese Definition ist in der Literatur nicht üblich. Dort werden meist viel stärkere Einschränkungen gemacht.

Autor: Klaus Pommerening, 5. März 2000; letzte Änderung: 23. November 2004.
E-Mail an `Pommerening »AT« imbei.uni-mainz.de`.