

Aufgabe 1) In einem Korpus mit 500 000 Adjektiv-Nomen-Paaren taucht das Adjektiv **bare** 18-mal, das Nomen **Münze** 43-mal und das Paar **bare Münze** 14-mal auf.

Wie berechnen Sie Schätzwerte für die Wahrscheinlichkeiten $p(\text{bare})$, $p(\text{Münze})$, $p(\text{bare, Münze})$, $p_1(\text{bare}|\text{Münze})$, $p_2(\text{Münze}|\text{bare})$, wobei $p_1(x|y)$ die Wahrscheinlichkeit von x **vor** y und $p_2(x|y)$ die Wahrscheinlichkeit von x **nach** y ist. (2 Punkte)

Aufgabe 2) Wie können Sie zeigen, dass das Wortpaar **bare Münze** in Aufgabe 1 signifikant häufiger ist, als bei statistischer Unabhängigkeit der Wörter zu erwarten wäre? Was müssen Sie hierfür konkret berechnen? (2 Punkte)

Aufgabe 4) Wie lautet die Formel für die Berechnung der Entropie $H(X)$ einer Zufallsvariablen X ?

Wie hoch ist die Entropie einer Zufallsvariablen mit einer uniformen Verteilung über 4 Werte (also wenn alle Werte die gleiche Wahrscheinlichkeit haben)? Wie hoch ist die Entropie, wenn einer der Werte die Wahrscheinlichkeit 1 besitzt? (2 Punkte)

Aufgabe 6) Angenommen Sie haben das Wort w_1 10-mal in einem Korpus der Größe 10 000 gesehen. In welchem Bereich liegt die erwartete Häufigkeit desselben Wortes in einem neuen Korpus derselben Größe aus derselben Quelle? Zur Auswahl stehen 0-9 Mal, 9-10 Mal, 10-11 Mal, mehr als 11 Mal. (1 Punkt)