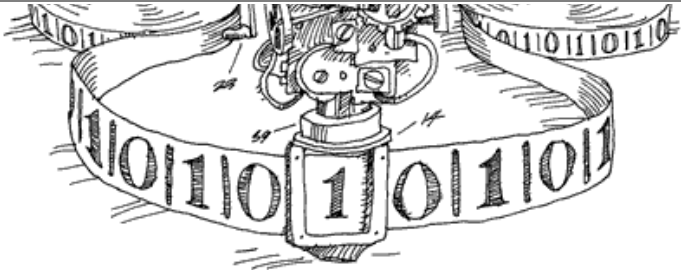


# Theoretische Grundlagen der Informatik

## Tutorium 14

Institut für Theoretische Informatik



- Jede Information bzw. Nachricht besitzt eine Quelle
  - Oft randomisiert a.k.a. Zufallsquellen
  - Wenn alle gesendete Nachrichten unabhängig voneinander sind, ist die Quelle gedächtnislos
- Es gibt immer einen Empfänger, der die Nachrichten beobachtet
- Je unvorhersehbarer die Nachricht, desto mehr Informationsgehalt
  - Wird deshalb auch manchmal Überraschungswert genannt
- Entropie ist ein Begriff für die Dichte der Informationen

- Informationsgehalt soll nicht negativ sein
- Ein sicheres Ergebnis ( $p = 1$ ) enthält keine Information
- Informationen von unabhängigen Nachrichten sollen sich addieren
- Kleine Änderungen der Wahrscheinlichkeit  $\Rightarrow$  kleine Änderung des Informationsgehalts
- $I(x) = -\log_b(p(x)) = \log_b(\frac{1}{p(x)})$  erfüllt diese Bedingungen
  - Meist wird als Basis  $b = 2$  verwendet

- Entropie ist entsprechend definiert

$$H(X) = \sum_{x \in X} (p(x) \cdot \log_2(\frac{1}{p(x)})) = \sum_{x \in X} (p(x) \cdot I(x))$$

- Zufallsquelle 1:  $p(A) = \frac{1}{2}, p(B) = \frac{1}{2}$
- $I(A) = \log_2\left(\frac{1}{0.5}\right) = \log_2(2) = 1 = I(B)$
- $H(X) = (p(A) \cdot I(A)) + (p(B) \cdot I(B)) = 0.5 + 0.5 = 1$

- Zufallsquelle 2:  $p(A) = \frac{1}{16}, p(B) = \frac{15}{16}$
- $I(A) = \log_2\left(\frac{1}{0.0625}\right) = \log_2(16) = 4$
- $I(B) = \log_2\left(\frac{1}{0.9375}\right) = \log_2\left(\frac{16}{15}\right) = 0.0931 \dots$
- $H(X) = (p(A) \cdot I(A)) + (p(B) \cdot I(B))$   
 $= \left(\frac{1}{16} \cdot 4\right) + \left(\frac{15}{16} \cdot 0.0931\right) = \frac{1}{4} + 0.873 = 0.337$

1. Wie groß sind der Informationsgehalt und die Entropie, wenn eine Quelle mit dem Alphabet  $\{0, 1\}$  nur aus dem Zeichen 0 bestehende Folgen sendet?
2. An einer Quelle mit  $n$  Zeichen tritt jedes Zeichen gleichverteilt auf. Wie groß sind der Informationsgehalt und die Entropie eines einzelnen Zeichens?
3. Berechnen Sie die Entropie des Wurfes eines idealen Würfels mit 8 Seiten, dessen Wahrscheinlichkeit für jede Seite  $p = \frac{1}{8}$  ist!
4. Was ist der Unterschied zwischen den beiden Folgen, die aus verschiedenen gedächtnislosen Quellen mit der gleichen Wahrscheinlichkeit für 0 und 1 gesendet werden, wenn man sie unter dem Aspekt Entropie und Ordnung betrachtet?
  - 4.1 ...101010101010101010...
  - 4.2 ...01101100110111000010...

Die Huffman-Codierung ist ein Algorithmus zur verlustfreien Datenkompression.

## Problemdefinition

### ■ Gegeben

- Ein Alphabet  $A = \{a_0, a_1, \dots, a_n\}$  der Größe  $n$
- Gewichte  $W = \{w_0, w_1, \dots, w_n\}$  für alle  $a \in A$ .  
Meist die Wahrscheinlichkeit, dass ein Zeichen auftritt.

### ■ Gesucht

- Eine binäre Codierung für alle Zeichen aus  $A$ , sodass die erwartete Code-Wortlänge in Bezug auf die Gewichte minimal ist.

Die Huffman-Codierung ist ein Algorithmus zur verlustfreien Datenkompression.

## Problemdefinition

### ■ Gegeben

- Ein Alphabet  $A = \{a_0, a_1, \dots, a_n\}$  der Größe  $n$
- Gewichte  $W = \{w_0, w_1, \dots, w_n\}$  für alle  $a \in A$ .  
Meist die Wahrscheinlichkeit, dass ein Zeichen auftritt.

### ■ Gesucht

- Eine binäre Codierung für alle Zeichen aus  $A$ , sodass die erwartete Code-Wortlänge in Bezug auf die Gewichte minimal ist.

Lässt sich sowohl auf konkrete Wörter anwenden als auch auf Quellen, von denen man weiß, wie wahrscheinlich sie welches Zeichen sendet.

# Huffman-Codierung Beispiel

Gegeben sei das Wort **abacabadabacaba**. Wie lautet eine Huffman-Codierung?



# Huffman-Codierung Beispiel

Gegeben sei das Wort **abacabadabacaba**. Wie lautet eine Huffman-Codierung?

- **#a = 8**
- **#b = 4**
- **#c = 2**
- **#d = 1**

# Huffman-Codierung Beispiel

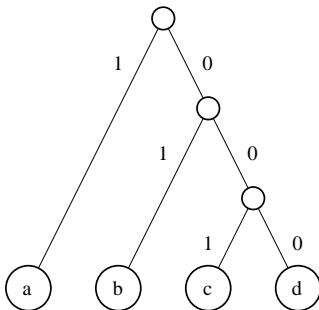
Gegeben sei das Wort **a****b****a****c****a****b****a****d****a****b****a****c****a****b****a**. Wie lautet eine Huffman-Codierung?

■ #a = 8

■ #b = 4

■ #c = 2

■ #d = 1



Gegeben sei eine Quelle mit Alphabet  $\{A, B, C, D\}$  und mit den folgenden Wahrscheinlichkeiten:

$$P(A) = \frac{1}{2}, P(B) = \frac{1}{4}, P(C) = \frac{1}{8}, P(D) = \frac{1}{8}$$

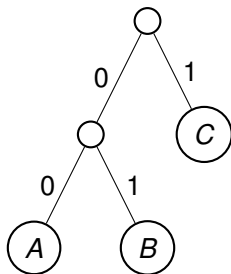
- Berechnen Sie die Entropie der Quelle!
- Erstellen Sie eine entsprechende Huffman-Codierung!
- Was ist die mittlere Codewortlänge? Gibt es einen Zusammenhang zur Entropie?

## Aufgabe B11 A3

Gegeben sei eine Quelle mit Alphabet  $\{A, B, C, D\}$  und mit den folgenden Wahrscheinlichkeiten:

$$P(A) = \frac{1}{2}, P(B) = \frac{1}{4}, P(C) = \frac{1}{8}, P(D) = \frac{1}{8}$$

■ Gegeben sei der folgende Huffman-Baum:



Dekodieren Sie 011011101100101011! Ist der Huffman-Code geeignet?

Das Hamilton-Kreis Problem ist NP-vollständig

In der Klasse NP liegen nicht-entscheidbare Probleme

Das Vertex-Cover Problem ist NP-vollständig

Semi-entscheidbare Sprachen sind unter Komplementbildung abgeschlossen



Nichtdeterministische endliche Automaten sind echt mächtiger als deterministische

Zu jeder CH-2-Sprache gibt es eine CH-1-Grammatik

Um zu zeigen, dass ein Problem  $\Pi$  NP-vollständig ist, genügt es, ein NP-schweres Problem auf  $\Pi$  zu reduzieren.


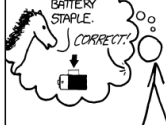
Das Travelman Salesmen Problem ist NP-vollständig

N-SAT ist immer NP-vollständig

## Deterministische Kellerautomaten erkennen Chomsky-2

$$NP \neq co - NP \implies P \neq NP$$

# Bis zum nächsten Mal!

<p>□□□□□□□□□□□□□□</p> <p>UNCOMMON (NON-GIBBERISH) BASE WORD</p> <p>ORDER UNKNOWN</p> <p>Trøub4dor &amp; 3</p> <p>CAPS? □</p> <p>COMMON SUBSTITUTIONS □□□</p> <p>NUMERAIL □□□</p> <p>PUNCTUATION □□□□</p> <p>(YOU CAN ADD A FEW MORE BITS TO ACCOUNT FOR THE FACT THAT THIS IS ONLY ONE OF A FEW COMMON FORMATS.)</p>	<p>~28 BITS OF ENTROPY</p> <p>□□□□□□□□</p> <p>□□□□□□□□</p> <p>□□□□</p> <p>□□□□</p> <p>□□□□</p> <p><math>2^{28} = 3 \text{ DAYS AT } 1000 \text{ GUESSES/SEC}</math></p> <p>(PLAUSIBLE ATTACK ON A WEAK REMOTE WEB SERVICE: YES, CRACKING A STOKEN HANGUP IS FINEST, BUT IT'S NOT WHAT THE AVERAGE USER SHOULD WORRY ABOUT.)</p> <p>DIFFICULTY TO GUESS: <b>EASY</b></p>	<p>WAS IT TROMBONE? NO, TROUBADOR. AND ONE OF THE 0s WAS A ZERO?</p> <p>AND THERE WAS SOME SYMBOL...</p>  <p>DIFFICULTY TO REMEMBER: <b>HARD</b></p>
<p>correct horse battery staple</p> <p>□□□□□□ □□□□□□ □□□□□□ □□□□□□</p> <p>FOUR RANDOM COMMON WORDS</p>	<p>~44 BITS OF ENTROPY</p> <p>□□□□□□□□□□</p> <p>□□□□□□□□□□</p> <p>□□□□□□□□□□</p> <p>□□□□□□□□□□</p> <p><math>2^{44} = 580 \text{ YEARS AT } 1000 \text{ GUESSES/SEC}</math></p> <p>DIFFICULTY TO GUESS: <b>HARD</b></p>	<p>THAT'S A BATTERY STAPLE.</p> <p>CORRECT!</p>  <p>DIFFICULTY TO REMEMBER: YOU'VE ALREADY MEMORIZED IT</p>

THROUGH 20 YEARS OF EFFORT, WE'VE SUCCESSFULLY TRAINED EVERYONE TO USE PASSWORDS THAT ARE HARD FOR HUMANS TO REMEMBER, BUT EASY FOR COMPUTERS TO GUESS.





Dieses Werk ist unter einem "Creative Commons Namensnennung-Weitergabe unter gleichen Bedingungen 3.0 Deutschland"-Lizenzvertrag lizenziert. Um eine Kopie der Lizenz zu erhalten, gehen Sie bitte zu <http://creativecommons.org/licenses/by-sa/3.0/de/> oder schreiben Sie an Creative Commons, 171 Second Street, Suite 300, San Francisco, California 94105, USA.

Davon ausgenommen sind das Titelbild, welches aus der März-April 2002 Ausgabe von American Scientist erschienen ist und ohne Erlaubnis verwendet wird, sowie das KIT Beamer Theme. Hierfür gelten die Bestimmungen der jeweiligen Urheber.