A1.1: Wetterentropie

Eine Wetterstation fragt täglich verschiedene Regionen ab und bekommt als Antwort jeweils eine Meldung x zurück, nämlich

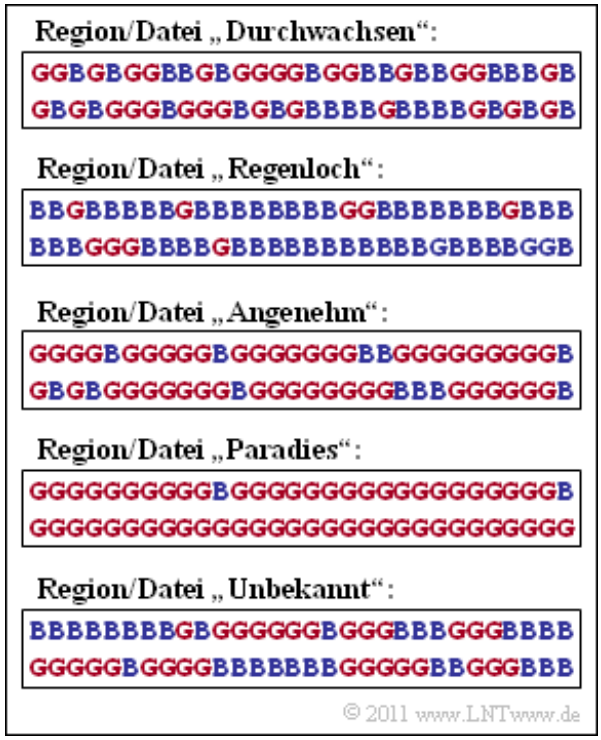
x = **B**: Das Wetter ist eher schlecht.
x = **G**: Das Wetter ist eher gut.

Die Daten wurden über viele Jahre für verschiedene Gebiete in Dateien abgelegt, so dass die Entropien der **B**/**G**–Folgen ermittelt werden können:

$$H = p_{\rm B} \cdot \log_2 \frac{1}{p_{\rm B}} + p_{\rm G} \cdot \log_2 \frac{1}{p_{\rm G}}$$

mit dem *Logarithmus dualis*

$$\log_2 \, p = \frac{\lg \, p}{\lg \, 2} \quad (= {\rm ld} \, p) \; .$$



„lg” kennzeichnet hierbei den Logarithmus zur Basis 10. Zu erwähnen ist ferner, dass jeweils noch die Pseudoeinheit „bit/Anfrage” anzufügen ist.

Die Grafik zeigt diese binären Folgen jeweils für 60 Tage und folgende Regionen:

- Region „Durchwachsen”: $p_{\rm B} = p_{\rm G} = 0.5$,
- Region „Regenloch”: $p_{\rm B} = 0.8$, $p_{\rm G} = 0.2$,
- Region „Angenehm”: $p_{\rm B} = 0.2$, $p_{\rm G} = 0.8$,
- Region „Paradies”: $p_{\rm B} = 1/30$, $p_{\rm G} = 29/30$.

Schließlich ist auch noch die Datei „Unbekannt” angegeben, deren statistische Eigenschaften zu schätzen sind.

Hinweis: Die Aufgabe bezieht sich auf das **Kapitel 1.1.** Für die vier ersten Dateien wird vorausgesetzt, dass die Ereignisse „B” und „G” statistisch unabhängig seien, eine für die Wetterpraxis allerdings eher unrealistische Annahme.

Fragebogen zu "A1.1: Wetterentropie"

a) Welche Entropie H_D weist die Datei „Durchwachsen" auf?

$\boldsymbol{H_D}$ = bit/Anfrage

b) Welche Entropie H_R weist die Datei „Regenloch" auf?

$\boldsymbol{H_R}$ = bit/Anfrage

c) Welche Entropie H_A weist die Datei „Angenehm" auf?

$\boldsymbol{H_A}$ = bit/Anfrage

d) Wie groß sind die Informationsgehalte der Ereignisse „B" und „G" bezogen auf die Datei „Paradies"?

$\boldsymbol{I_B}$ = bit/Anfrage

$\boldsymbol{I_G}$ = bit/Anfrage

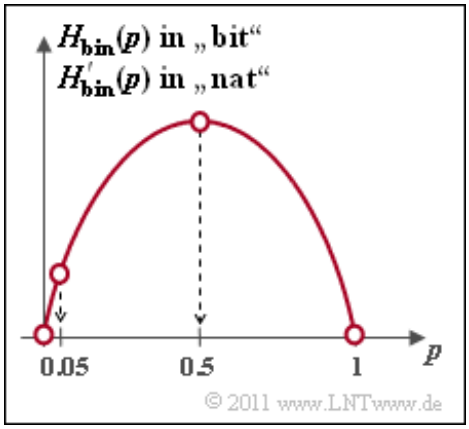
e) Wie groß ist die Entropie (das heißt: der mittlere Informationsgehalt) H_P der Datei „Paradies"? Interpretieren Sie das Ergebnis?

$\boldsymbol{H_P}$ = bit/Anfrage

f) Welche Aussagen könnten für die Datei „Unbekannt" gelten?

- ☐ Die Ereignisse „B" und „G" sind etwa gleichwahrscheinlich.
- ☐ Die Folgenelemente sind statistisch voneinander unabhängig.
- ☐ Die Entropie dieser Datei ist $H_U \approx 0.7$ bit/Anfrage.
- ☐ Die Entropie dieser Datei ist $H_U = 1.5$ bit/Abfrage.

Z1.1: Binäre Entropiefunktion



Wir betrachten eine Folge von binären Zufallsgrößen mit dem Symbolvorrat {**A**, **B**} $\Rightarrow M = 2$. Die Auftrittswahrscheinlichkeiten der beiden Symbole seien $p_A = p$ und $p_B = 1 - p$.

Die einzelnen Folgenelemente sind statistisch unabhängig. Für die Entropie dieser Nachrichtenquelle gilt gleichermaßen:

$$H_{\rm bin}(p) = p \cdot {\rm ld}\,\frac{1}{p} + (1-p) \cdot {\rm ld}\,\frac{1}{1-p}\ {\rm in\ [bit]},$$

$$H'_{\rm bin}(p) = p \cdot {\rm ln}\,\frac{1}{p} + (1-p) \cdot {\rm ln}\,\frac{1}{1-p}\ {\rm in\ [nat]}.$$

In diesen Gleichungen werden als Kurzbezeichnungen verwendet:

- der *natürliche* Logarithmus $\ln p = \log_e p$,
- der Logarithmus *dualis* $\text{ld}\ p = \log_2 p$.

Die Grafik zeigt diese binäre Entropiefunktion in Abhängigkeit des Parameters p, wobei $0 \le p \le 1$ vorausgesetzt wird.

In den Teilaufgaben (e) und (f) soll der relative Fehler ermittelt werden, wenn die Symbolwahrscheinlichkeit p per Simulation (also als relative Häufigkeit h) ermittelt wurde und sich dabei fälschlicherweise $h = 0.9\ p$ ergeben hat. Der relative Fehler ist dann wie folgt gegeben:

$$\varepsilon_H = \frac{H_{\rm bin}(h) - H_{\rm bin}(p)}{H_{\rm bin}(p)}.$$

Hinweis: Die Aufgabe gehört zum **Kapitel 1.1.**

Fragebogen zu "Z1.1: Binäre Entropiefunktion"

a) Wie hängen $H_{\rm bin}(p)$ in bit und $H'_{\rm bin}(p)$ in nat zusammen?

- ☐ $H_{\rm bin}(p)$ und $H'_{\rm bin}(p)$ unterscheiden sich um einen Faktor.
- ☐ Es gilt $H'_{\rm bin}(p) = H_{\rm bin}(\ln\, p)$.
- ☐ Es gilt $H'_{\rm bin}(p) = 1 + H_{\rm bin}(2\, p)$.

b) Zeigen Sie, dass sich das Maximum der binären Entropiefunktion für $p = 0.5$ ergibt. Wie groß ist $H_{\rm bin}(p = 0.5)$?

$$\boldsymbol{H_{\rm bin}(p = 0.5)} = \quad \text{bit}$$

c) Berechnen Sie den binären Entropiewert für $p = 0.05$.

$$\boldsymbol{H_{\rm bin}(p = 0.05)} = \quad \text{bit}$$

d) Geben Sie den größeren der beiden p–Werte ein, die sich aus der Gleichung $H_{\rm bin}(p) = 0.5$ bit ergeben.

$$\boldsymbol{p} =$$

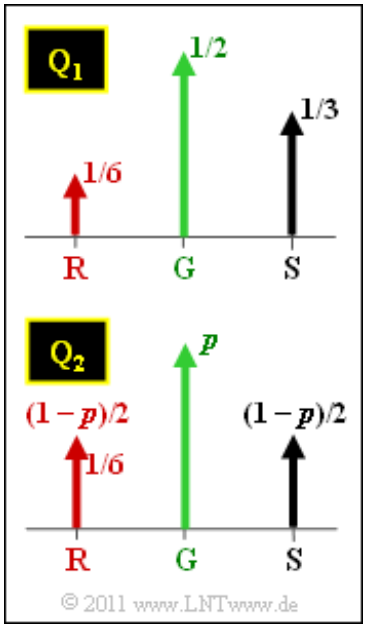
e) Durch unzureichende Simulation wurde $p = 0.5$ um 10% zu niedrig ermittelt. Wie groß ist der prozentuale Fehler hinsichtlich der Entropie?

$$\boldsymbol{p = 0.45 \text{ statt } p = 0.5:} \quad \varepsilon_H = \quad \%$$

f) Durch unzureichende Simulation wurde $p = 0.05$ um 10% zu niedrig ermittelt. Wie groß ist der prozentuale Fehler hinsichtlich der Entropie?

$$\boldsymbol{p = 0.045 \text{ statt } p = 0.05:} \quad \varepsilon_H = \quad \%$$

A1.2: Entropie von Ternärquellen



Die Entropie einer wertdiskreten gedächtnislosen Nachrichtenquelle mit M möglichen Symbolen lautet:

$$H = \sum_{\mu=1}^{M} p_\mu \cdot \log_2 \frac{1}{p_\mu}, \quad \text{Pseudoeinheit: bit}.$$

Hierbei bezeichnen die p_μ die Auftrittswahrscheinlichkeiten der einzelnen Symbole bzw. Ereignisse. Im vorliegenden Beispiel werden die Ereignisse mit **R**(ot), **G**(rün) und **S**(chwarz) bezeichnet.

Bei einer binären Quelle mit den Auftrittswahrscheinlichkeiten p und $1 – p$ kann hierfür geschrieben werden:

$$H = H_{\rm bin}(p) = p \cdot \log_2 \frac{1}{p} + (1-p) \cdot \log_2 \frac{1}{1-p}.$$

Die Entropie einer mehrstufigen Quelle lässt sich häufig mit dieser „binären Entropiefunktion” $H_{\rm bin}(p)$ – ebenfalls mit der Pseudoeinheit „bit” – ausdrücken.

Betrachtet werden in dieser Aufgabe zwei Ternärquellen mit den Symbolwahrscheinlichkeiten gemäß der obigen Grafik:

- die Quelle Q_1 mit $p_G = 1/2$, $p_S = 1/3$, $p_R = 1/6$,
- die Quelle Q_2 mit $p_G = p$, $p_R = p_S = (1 – p)/2$.

Die Ternärquelle Q_2 lässt sich auch auf Roulette anwenden, wenn ein Spieler nur auf die Felder **R**ot, **S**chwarz und **G**rün (die „Null”) setzt. Dieser Spieltyp wird im Fragebogen mit „Roulette 1” bezeichnet.

Dagegen weist „Roulette 2” darauf hin, dass der Spieler auf einzelne Zahlen (**0**, ... , **36**) setzt.

Hinweis: Die Aufgabe bezieht sich auf das **Kapitel 1.1.**

Fragebogen zu "A1.2: Entropie von Ternärquellen"

a) Welche Entropie H besitzt die Quelle Q_1?

Q_1: H = bit

b) Welche der folgenden Aussagen sind zutreffend, wenn man **R**, **G** und **S** durch die Zahlenwerte –1, 0 und +1 darstellt?

- ☐ Es ergibt sich eine kleinere Entropie.
- ☐ Die Entropie bleibt gleich.
- ☐ Es ergibt sich eine größere Entropie.

c) Bestimmen Sie die Entropie der Quelle Q_2 unter Verwendung der binären Entropiefunktion $H_{bin}(p)$. Welcher Wert ergibt sich für $p = 0.5$?

Q_2; $p = 0.5$: H = bit

d) Für welchen p–Wert ergibt sich die maximale Entropie?

Q_2, $H \rightarrow H_{max}$: p =

e) Welche Entropie hat die Nachrichtenquelle „Roulette" hinsichtlich der Ereignisse **R**ot, **S**chwarz und **G**rün (die „Null")?

Roulette 1: H = bit

f) Welche Entropie weist „Roulette" hinsichtlich der Zahlen **0**, ... , **36** auf?

Roulette 2: H = bit

A1.3: $H_0, H_1, H_2, \ldots, H$

Die Grafik zeigt vier Symbolfolgen $\langle q_\nu \rangle$ mit jeweiliger Länge $N = 60$. Die Quellensymbole sind jeweils **A** und **B**. Daraus folgt direkt, dass für den Entscheidungsgehalt aller betrachteten Quellen $H_0 = 1$ bit/Symbol gilt. Die Symbole **A** und **B** treten mit den Wahrscheinlichkeiten p_A und p_B auf.

Die folgende Tabelle zeigt neben H_0 die Entropienäherungen

- H_1, basierend auf p_A und p_B (Spalte 2),
- H_2, basierend auf Zweiertupel (Spalte 3),
- H_3, basierend auf Dreiertupel (Spalte 4),
- H_4, basierend auf Vierertupel (Spalte 5),
- die tatsächliche Quellenentropie H, die sich aus H_k durch den Grenzübergang für $k \to \infty$ ergibt (letzte Spalte).

Zwischen diesen Entropien bestehen folgende Größenrelationen:

$$H \le \ldots \le H_3 \le H_2 \le H_1 \le H_0 \,.$$

Nicht bekannt ist die Zuordnung zwischen den Quellen Q1, Q2, Q3, Q4 und den in der Grafik gezeigten gezeigten Symbolfolgen (Schwarz, Blau, Rot, Grün). Es ist lediglich bekannt, dass die Quelle Q4 einen Wiederholungscode beinhaltet. Zu bestimmen sind für diese Nachrichtenquelle schließlich noch die Entropienäherungen H_2 und H_3.

Für die Quelle Q4 sind nur $H_1 = 1$ bit/Symbol ($\Rightarrow$ **A** und **B** gleichwahrscheinlich), die Näherung $H_4 \approx 0.789$ bit/Symbol und der Entropie–Endwert $H = 0.5$ bit/Symbol angegeben. Letzterer aufgrund der Tatsache, dass bei der entsprechenden Symbolfolge jedes zweite Symbol keinerlei Information lierfert.

Quelle	H_0	H_1	H_2	H_3	H_4	...	H
Q1	1.000	0.500	0.500	0.500	0.500	...	0.500
Q2	1.000	1.000	0.750	0.667	0.625	...	0.500
Q3	1.000	1.000	1.000	1.000	1.000	...	1.000
Q4	1.000	1.000	???	???	**0.789**	...	0.500

Hinweis: Die Aufgabe gehört zum Themengebiet von **Kapitel 1.2.** Für die k–te Entropienäherung gilt bei Binärquellen ($M = 2$) mit der Verbundwahrscheinlichkeit $p_i^{(k)}$ eines k–Tupels:

$$H_k = \frac{1}{k} \cdot \sum_{i=1}^{2^k} p_i^{(k)} \cdot \log_2 \frac{1}{p_i^{(k)}} \quad \text{(Einheit: bit/Symbol)}.$$

Fragebogen zu "A1.3: H_0, H_1, H_2, ... , H"

a) Von welcher Quelle stammt die schwarze Symbolfolge?

- ☐ Q1,
- ☐ Q2,
- ☐ Q3,
- ☐ Q4.

b) Von welcher Quelle stammt die blaue Symbolfolge?

- ☐ Q1,
- ☐ Q2,
- ☐ Q3,
- ☐ Q4.

c) Von welcher Quelle stammt die rote Symbolfolge?

- ☐ Q1,
- ☐ Q2,
- ☐ Q3,
- ☐ Q4.

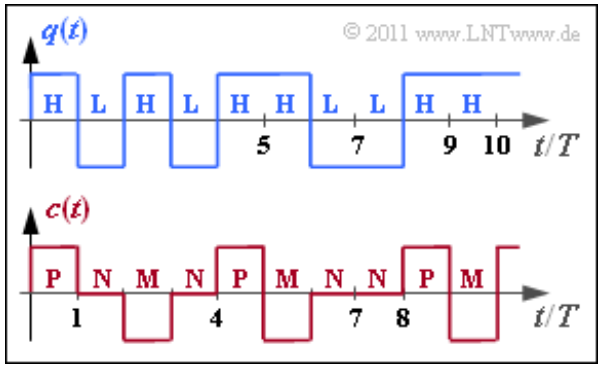
d) Berechnen Sie die Entropienäherung H_2 des Wiederholungscodes.

$\boldsymbol{H_2}$ = bit/Symbol

e) Berechnen Sie die Entropienäherung H_3 des Wiederholungscodes.

$\boldsymbol{H_3}$ = bit/Symbol

A1.4: Entropienäherungen H_k



Die Grafik zeigt oben das binäre Quellensignal $q(t)$, das man ebenfalls durch die Symbolfolge $\langle q_\nu \rangle$ mit $q_\nu \in \{\mathbf{L}, \mathbf{H}\}$ beschreiben kann. In der gesamten Aufgabe gelte $p_\mathrm{L} = p_\mathrm{H} = 0.5$.

Das codierte Signal $c(t)$ und die dazugehörige Symbolfolge $\langle c_\nu \rangle \in \{\mathbf{P}, \mathbf{N}, \mathbf{M}\}$ ergibt sich aus der AMI–Codierung (*Alternate Mark Inversion*) nach folgender Vorschrift:

- Das Binärsymbol **L** ⇒ *Low* wird stets durch das Ternärsymbol **N** ⇒ *Null* dargestellt.
- Das Binärsymbol **H** ⇒ *High* wird ebenfalls deterministisch, aber alternierend (daher der Name „AMI") durch die Symbole **P** ⇒ *Plus* und **M** ⇒ *Minus* codiert.

In dieser Aufgabe sollen die Entropienäherungen für das AMI–codierte Signal berechnet werden:

- Die Näherung H_1 bezieht sich nur auf die Symbolwahrscheinlichkeiten p_P, p_N und p_M.
- Die k–te Entropienäherung ($k = 2, 3, \ldots$) kann nach folgender Gleichung ermittelt werden:

$$H_k = \frac{1}{k} \cdot \sum_{i=1}^{3^k} p_i^{(k)} \cdot \log_2 \frac{1}{p_i^{(k)}} \quad \text{(Einheit: bit/Symbol)}.$$

  Hierbei bezeichnet $p_i^{(k)}$ die i–te Verbundwahrscheinlichkeit eines k–Tupels.

Hinweis: Die Aufgabe gehört zu **Kapitel 1.2.** In der **Aufgabe Z1.4** wird die tatsächliche Entropie der Codesymbolfolge $\langle c_\nu \rangle$ zu $H = 1$ bit/Symbol berechnet. Zu erwarten sind die folgenden Größenrelationen:

$$H \le \ldots \le H_3 \le H_2 \le H_1 \le H_0 .$$

Fragebogen zu "A1.4: Entropienäherungen H_k"

a) Wie groß ist der Entscheidungsgehalt des AMI–Codes?

H_0 = bit/Symbol

b) Berechnen Sie die erste Entropienäherung.

H_1 = bit/Symbol

c) Wie groß ist die Entropienäherung H_2, basierend auf Zweiertupel?

H_2 = bit/Symbol

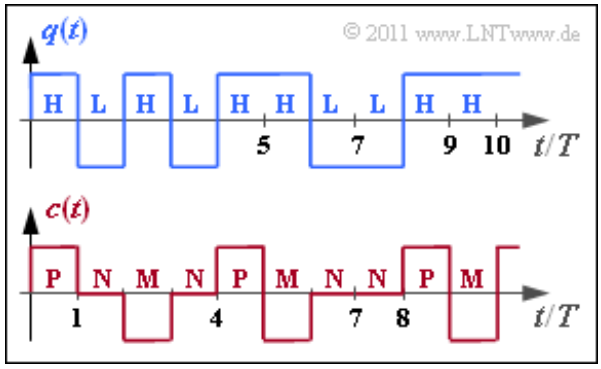
d) Welchen Wert liefert die Entropienäherung H_3, basierend auf Dreiertuptel?

H_3 = bit/Symbol

e) Welche Aussagen gelten für die Entropienäherung H_4?

- ☐ Es muss über $3^4 = 81$ Summanden gemittelt werden.
- ☐ Es gilt 1 bit/Symbol $< H_4 < H_3$.
- ☐ Nach langer Rechnung erhält man $H_4 = 1.333$ bit/Symbol.

Z1.4: Entropie der AMI–Codierung



Wir gehen von ähnlichen Voraussetzungen wie in der **Aufgabe A1.4** aus: Eine Binärquelle liefert die Quellensybolfolge $\langle q_\nu \rangle$ mit $q_\nu \in \{\mathbf{L}, \mathbf{H}\}$, wobei es keine statistischen Bindungen zwischen den einzelnen Folgenelementen gibt.

Für die Symbolwahrscheinlichkeiten gelte:

- $p_\mathrm{L} = p_\mathrm{H} = 1/2$ (Teilaufgaben a und b),
- $p_\mathrm{L} = 1/4$, $p_\mathrm{H} = 3/4$ (Teilaufgaben c, d und e),
- $p_\mathrm{L} = 3/4$, $p_\mathrm{H} = 1/4$ (Teilaufgabe f).

Das dargestellte Codesignal $c(t)$ und die zugehörige Symbolfolge $\langle c_\nu \rangle$ mit $c_\nu \in \{\mathbf{P}, \mathbf{N}, \mathbf{M}\}$ ergibt sich aus der AMI–Codierung (*Alternate Mark Inversion*) nach folgender Vorschrift:

- Das Binärsymbol **L** ⇒ *Low* wird stets durch das Ternärsymbol **N** ⇒ *Null* dargestellt.
- Das Binärsymbol **H** ⇒ *High* wird ebenfalls deterministisch, aber alternierend (daher der Name „AMI") durch die Symbole **P** ⇒ *Plus* und **M** ⇒ *Minus* codiert.

In dieser Aufgabe sollen für die drei oben genannten Parametersätze der Entscheidungsgehalt H_0 sowie die resultierende Entropie H_C der Codesymbolfolge $\langle c_\nu \rangle$ bestimmt werden. Die relative Redundanz der Codefolge ergibt sich daraus entsprechend der Gleichung

$$r_\mathrm{C} = \frac{H_0 - H_\mathrm{C}}{H_\mathrm{C}}.$$

Hinweis: Die Aufgabe gehört zu **Kapitel 1.2.** Allgemein bestehen folgende Relationen zwischen dem Entscheidungsgehalt H_0, der Entropie H (hier gleich H_C) und den Entropienäherungen:

$$H \le \ldots \le H_3 \le H_2 \le H_1 \le H_0 .$$

In **Aufgabe A1.4** wurden für gleichwahrscheinliche Symbole **L** und **H** die Entropie–Näherungen wie folgt berechnet (jeweils in bit/Symbol):

$$H_1 = 1.500 , \quad H_2 = 1.375 , \quad H_3 = 1.292 .$$

Fragebogen zu "Z1.4: Entropie der AMI–Codierung"

a) Die Quellensymbole seien gleichwahrscheinlich. Wie groß ist die Entropie H_C der Codesymbolfolge $\langle c_\nu \rangle$?

$p_L = p_H$: H_C = bit/Ternärsymbol

b) Wie groß ist die relative Redundanz der Codesymbolfolge?

$p_L = p_H$: r_C = %

c) Für die Binärquelle gelte nun $p_L = 1/4$ und $p_H = 3/4$. Welcher Wert ergibt sich nun für die Entropie der Codesymbolfolge?

$p_L = 1/4$: H_C = bit/Ternärsymbol

d) Wie groß ist nun die relative Redundanz der Codesymbolfolge?

$p_L = 1/4$: r_C = %

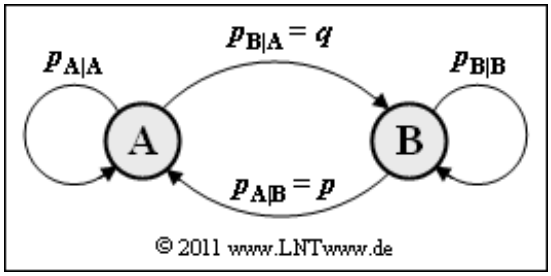
e) Berechnen Sie die Näherung H_1 der Coderentropie für $p_L = 1/4$, $p_H = 3/4$.

$p_L = 1/4$: H_1 = bit/Ternärsymbol

f) Berechnen Sie die Näherung H_1 der Coderentropie für $p_L = 3/4$, $p_H = 1/4$.

$p_L = 3/4$: H_1 = bit/Ternärsymbol

A1.5: Binäre Markovquelle



Die **Aufgabe A1.4** hat gezeigt, dass die Berechnung der Entropie bei einer gedächtnisbehafteten Quelle sehr aufwändig sein kann. Man muss dann zunächst (sehr viele) Entropienäherungen H_k für k–Tupel berechnen und kann erst dann mit dem Grenzübergang $k \to \infty$ die Quellenentropie ermitteln:

$$H = \lim_{k \to \infty} H_k \,.$$

Oft tendiert dabei H_k nur sehr langsam gegen den Grenzwert H.

Der Rechengang wird drastisch reduziert, wenn die Nachrichtenquelle **Markoveigenschaften** besitzt. Die Grafik zeigt das Übergangsdiagramm für eine binäre Markovquelle mit den zwei Zuständen (Symbolen) **A** und **B**. Dieses ist durch die beiden bedingten Wahrscheinlichkeiten $p_{A|B} = p$ und $p_{B|A} = q$ eindeutig bestimmt. Die bedingten Wahrscheinlichkeiten $p_{A|A}$ und $p_{B|B}$ sowie die Symbolwahrscheinlichkeiten p_A und p_B lassen sich daraus ermitteln.

Die Entropie der binären Markovkette (mit der Einheit „bit/Symbol") lautet dann:

$$H = p_{AA} \cdot \log_2 \frac{1}{p_{A|A}} + p_{AB} \cdot \log_2 \frac{1}{p_{B|A}} + p_{BA} \cdot \log_2 \frac{1}{p_{A|B}} + p_{BB} \cdot \log_2 \frac{1}{p_{B|B}} \,.$$

Bei dieser Gleichung ist zu beachten, dass im Argument des *Logarithmus dualis* jeweils die *bedingten Wahrscheinlichkeiten* $p_{A|A}$, $p_{B|A}$, ... einzusetzen sind, während für die Gewichtung die *Verbundwahrscheinlichkeiten* p_{AA}, p_{AB}, ... zu verwenden sind.

Mit der Entropienäherung erster Ordnung,

$$H_1 = p_A \cdot \log_2 \frac{1}{p_A} + p_B \cdot \log_2 \frac{1}{p_B} \quad \text{(Einheit: bit/Symbol)} \,,$$

sowie der oben angegebenen (tatsächlichen) Entropie H lassen sich bei einer Markovquelle auch alle weiteren Entropienäherungen (k = 2, 3, ...) direkt berechnen:

$$H_k = \frac{1}{k} \cdot [H_1 + (k-1) \cdot H_M] \,.$$

Hinweis: Diese Aufgabe gehört zum Themengebiet von **Kapitel 1.2.** Mit Ausnahme der Teilaufgabe (f) sei $p = 1/4$ und $q = 1/2$.

Für die (ergodischen) Symbolwahrscheinlichkeiten einer Markovkette erster Ordnung gilt:

$$p_A = \frac{p_{A|B}}{p_{A|B} + p_{B|A}} \,, \quad p_B = \frac{p_{B|A}}{p_{A|B} + p_{B|A}} \,.$$

Fragebogen zu "A1.5: Binäre Markovquelle"

a) Geben Sie die Übergangswahrscheinlichkeiten für $p = 1/4$ und $q = 1/2$ an.

$\boldsymbol{p = 1/4, q = 1/2}$: $p_{A|A}$ =

$p_{B|B}$ =

b) Wie groß sind die Symbolwahrscheinlichkeiten?

$\boldsymbol{p = 1/4, q = 1/2}$: p_A =

p_B =

c) Geben Sie die Entropienäherung erster Ordnung an.

$\boldsymbol{p = 1/4, q = 1/2}$: H_1 = bit/Symbol

d) Welche Entropie besitzt diese Markovquelle?

$\boldsymbol{p = 1/4, q = 1/2}$: H = bit/Symbol

e) Welche Näherungen H_k ergeben sich aufgrund der Markoveigenschaften?

$\boldsymbol{p = 1/4, q = 1/2}$: H_2 = bit/Symbol

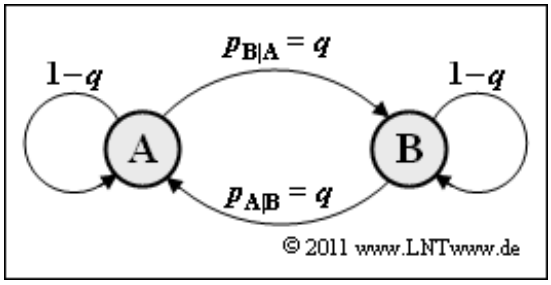
H_3 = bit/Symbol

H_4 = bit/Symbol

f) Welche Entropie besitzt die Markovquelle mit $p = 1/4$ und $q = 3/4$?

$\boldsymbol{p = 1/4, q = 3/4}$: H = bit/Symbol

Z1.5: Symmetrische Markovquelle



In der **Aufgabe A1.5** wurde eine binäre Markovquelle behandelt, bei der die Übergangswahrscheinlichkeiten von **A** nach **B** sowie von **B** nach **A** unterschiedlich waren. In dieser Aufgabe soll nun gelten:

$$p_{A|B} = p_{B|A} = q \qquad (0 \le q \le 1)\,.$$

Alle in der Aufgabe A1.5 angegebenen Gleichungen gelten auch hier:

- **Entropie:**

$$H = p_{AA} \cdot \text{ld}\,\frac{1}{p_{A|A}} + p_{AB} \cdot \text{ld}\,\frac{1}{p_{B|A}} + p_{BA} \cdot \text{ld}\,\frac{1}{p_{A|B}} + p_{BB} \cdot \text{ld}\,\frac{1}{p_{B|B}},$$

- **Erste Entropienäherung**:

$$H_1 = p_A \cdot \text{ld}\,\frac{1}{p_A} + p_B \cdot \text{ld}\,\frac{1}{p_B},$$

- ***k*–te Entropienäherung** (k = 2, 3, ...):

$$H_k = \frac{1}{k} \cdot [H_1 + (k-1) \cdot H]\,, \qquad H = \lim_{k \to \infty} H_k\,.$$

Hinweis: Die Aufgabe bezieht sich auf **Kapitel 1.2, Seite 5c.** Bei allen Entropien ist die Pseudoeinheit „bit/Symbol” hinzuzufügen.

Fragebogen zu "Z1.5: Symmetrische Markovquelle"

a) Berechnen Sie die Symbolwahrscheinlichkeiten.

$\boldsymbol{q = 1/4}$: p_A =

p_B =

b) Berechnen Sie die Quellenentropie H. Welches Ergebnis liefert $q = 1/4$?

$\boldsymbol{q = 1/4}$: H = bit/Symbol

c) Welche Entropienäherungen erhält man für $q = 1/4$?

$\boldsymbol{q = 1/4}$: H_1 = bit/Symbol

H_2 = bit/Symbol

H_3 = bit/Symbol

d) Bestimmen Sie q derart, dass H maximal wird. Interpretation.

$H \rightarrow$ **Maximum:** q =

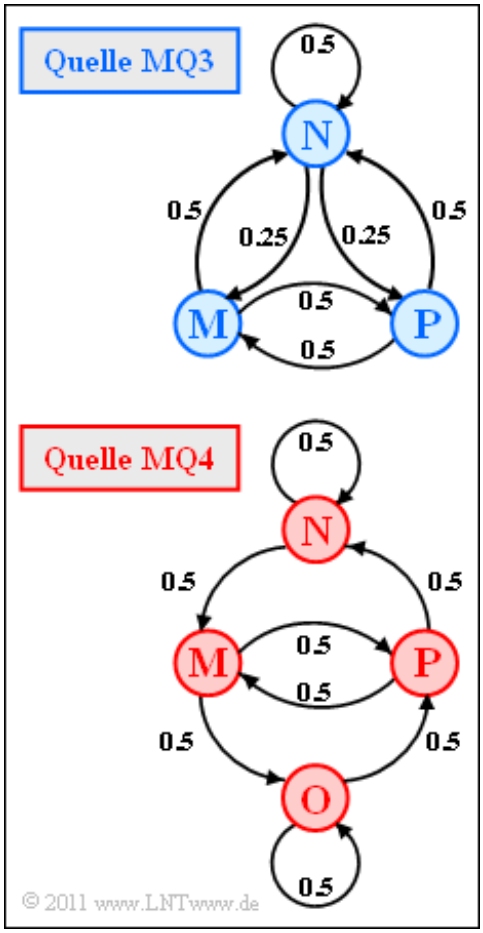
e) Welche Symbolfolgen sind mit $q = 0$ möglich?

- ☐ **AAAAAA** ...
- ☐ **BBBBBB** ...
- ☐ **ABABAB** ...

f) Welche Symbolfolgen sind mit $q = 1$ möglich?

- ☐ **AAAAAA** ...
- ☐ **BBBBBB** ...
- ☐ **ABABAB** ...

A1.6: Nichtbinäre Markovquellen



Die Grafik zeigt zwei ergodische Markovquellen (MQ):

- Die Quelle MQ3 ist durch $M = 3$ Zustände (Symbole) **N**, **M**, **P** gekennzeichnet. Aufgrund der Stationarität haben die Wahrscheinlichkeiten folgende Werte:

$$p_{\rm N} = 1/2\,,\quad p_{\rm M} = p_{\rm P} = 1/4\,.$$

- Bei der Quelle MQ4 ist zusätzlich der Zustand **O** möglich ⇒ $M = 4$. Aufgrund der symmetrischen Übergänge sind die stationären Wahrscheinlichkeiten alle gleich:

$$p_{\rm N} = p_{\rm M} = p_{\rm O} = p_{\rm P} = 1/4\,.$$

Informationstheoretisch sind Markovquellen von besonderer Bedeutung, da bei diesen – und nur bei diesen – durch H_1 (Entropienäherung, nur auf den Symbolwahrscheinlichkeiten basierend) und H_2 (zweite Entropienäherung, berechenbar mit den Verbundwahrscheinlichkeiten für alle Zweiertupel) gleichzeitig auch

- die weiteren Entropienäherungen $H_k (k = 3, 4, \text{ ... })$ und
- die tatsächliche Quellenentropie H

bestimmt sind. Es gelten folgende Bestimmungsgleichungen:

$$H = 2 \cdot H_2 - H_1\,,\quad H_k = \frac{1}{k} \cdot \left[H_1 + (k-1) \cdot H\right].$$

Hinweis: Die Aufgabe gehört zu **Kapitel 1.2.** Hier finden Sie auch Hinweise zur Berechnung der ersten und zweiten Entropienäherung.

Fragebogen zu "A1.6: Nichtbinäre Markovquellen"

a) Berechnen Sie die Entropienäherung H_1 der Markovquelle MQ3.

MQ3: H_1 = bit/Symbol

b) Berechnen Sie die Entropienäherung H_2 der Markovquelle MQ3.

MQ3: H_2 = bit/Symbol

c) Wie groß sind die Näherungen H_3 und H_4 und die Quellenentropie H?

MQ3: H_3 = bit/Symbol

H_4 = bit/Symbol

H = bit/Symbol

d) Berechnen Sie die Entropienäherung H_1 der Markovquelle MQ4.

MQ4: H_1 = bit/Symbol

e) Berechnen Sie die Entropienäherung H_2 der Markovquelle MQ4.

MQ4: H_2 = bit/Symbol

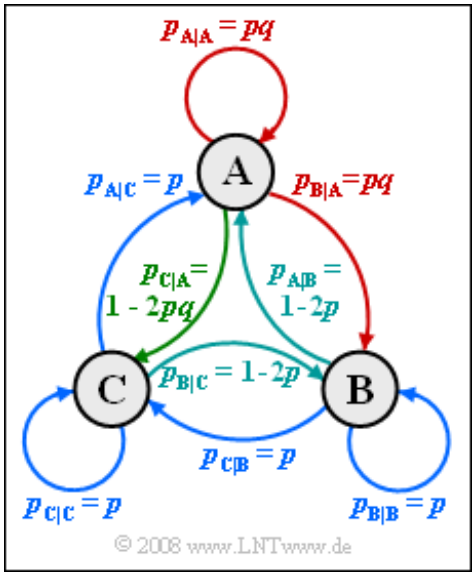
f) Wie groß sind hier die Näherungen H_3 und H_4 und die Quellenentropie H?

MQ4: H_3 = bit/Symbol

H_4 = bit/Symbol

H = bit/Symbol

Z1.6: Ternäre Markovquelle



Die Grafik zeigt eine Markovquelle mit $M = 3$ Zuständen **A**, **B** und **C**. Für die beiden Parameter dieses Markovprozesses soll gelten:

$$0 \le p \le 0.5, \quad 0 \le q \le 1 .$$

Aufgrund der Markoveigenschaft dieser Quelle kann die Entropie auf unterschiedliche Weise ermittelt werden:

- Man berechnet die beiden ersten **Entropienäherungen** H_1 und H_2. Dann gilt:

$$H = 2 \cdot H_2 - H_1 .$$

- Nach der so genannten *direkten Berechnungsmethode* kann die Entropie aber auch wie folgt berechnet werden (insgesamt 9 Terme):

$$H = p_{AA} \cdot \text{ld}\,\frac{1}{p_{A|A}} + p_{AB} \cdot \text{ld}\,\frac{1}{p_{B|A}} + \ldots ,$$

$$p_{AA} = p_A \cdot p_{A|A}, \quad p_{AB} = p_A \cdot p_{B|A}, \quad \ldots$$

Hinwis: Die Aufgabe gehört zum Themenkomplex von **Kapitel 1.2.**

Fragebogen zu "Z1.6: Ternäre Markovquelle"

a) Für welche Parameter p, q ergibt sich die maximale Entropie pro Symbol?

$\boldsymbol{p}$ =

$\boldsymbol{q}$ =

$\boldsymbol{H_{max}}$ = bit/Symbol

b) Es sei $p = 1/4$ und $q = 1$. Welcher Wert ergibt sich in diesem Fall für die erste Entropienäherung?

$p = 1/4$, $q = 1$: $\boldsymbol{H_1}$ = bit/Symbol

c) Weiterhin gelte $p = 1/4$ und $q = 1$. Welcher Wert ergibt sich in diesem Fall für die zweite Entropienäherung?

$p = 1/4$, $q = 1$: $\boldsymbol{H_2}$ = bit/Symbol

d) Wie groß ist die Quellenentropie mit $p = 1/4$, $q = 1$?

$p = 1/4$, $q = 1$: $\boldsymbol{H}$ = bit/Symbol

e) Wie groß ist die Quellenentropie mit $p = 1/2$, $q = 0$?

$p = 1/2$, $q = 0$: $\boldsymbol{H}$ = bit/Symbol

A1.7: Entropie natürlicher Texte

Anfang der 1950er Jahre schätzte **Claude E. Shannon** die Entropie H der englischen Sprache mit einem bit pro Zeichen ab. Kurze Zeit später kam **Karl Küpfmüller** bei einer empirischen Untersuchung der deutschen Sprache auf einen Entropiewert von $H = 1.3$ bit/Zeichen, also nur etwas größer. Die Ergebnisse von Shannon und Küpfmüller beruhen dabei interessanter Weise auf zwei völlig unterschiedlichen Methoden.

Die differierenden Ergebnisse lassen sich eher nicht mit den geringen Differenzen hinsichtlich des Symbolumfangs M erklären:

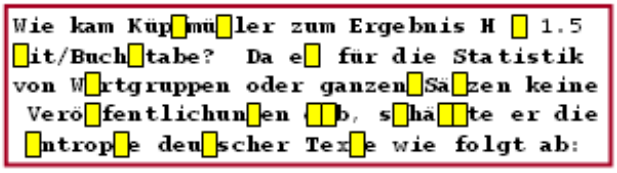
- Shannon ging von 26 Buchstaben und dem Leerzeichen aus ⇒ $M = 27$.
- Küpfmüller ging von $M = 26$ Buchstaben aus, ebenfalls ohne zwischen Groß– und Kleinschreibung zu unterscheiden.

Mit dieser Aufgabe soll gezeigt werden, wie sich

- Auslöschungen (*Erasures*) ⇒ man kennt den Ort eines Fehlers,
- Zeichenfehler ⇒ es ist nicht offensichtlich, was falsch und was richtig ist,

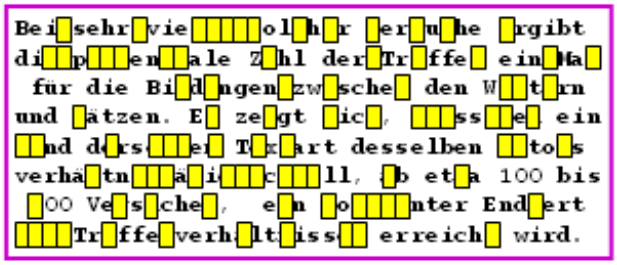
auf die Verständlichkeit eines Textes auswirken. Unser Text beinhaltet auch die typisch deutschen Buchstaben „ä”, „ö”, „ü” und „ß” sowie Ziffern und Interpunktion. Außerdem wird zwischen Groß– und Kleinschreibung unterschieden.

A) $N = 199$, $E = 20$ ⇒ ca. 10% „Erasure“

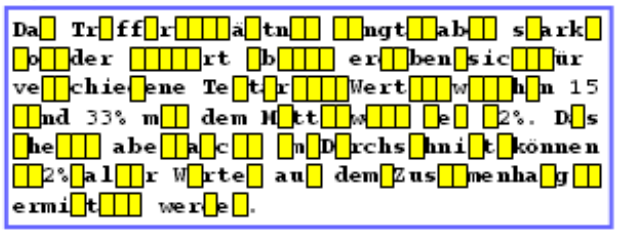


B) $N = 240$, $E = 48$ ⇒ ca. 20% „Erasure“

C) $N = 319$, $E = 97$ ⇒ ca. 30% „Erasure“

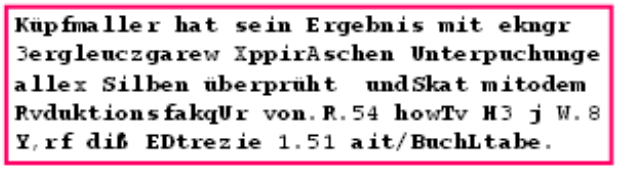


D) $N = 257$, $E = 103$ ⇒ ca. 40% „Erasure“



E) $N = 239$, $E = 103$ ⇒ ca. 50% „Erasure“

F) $N = 197$, $F = 39$ ⇒ ca. 20% falsche Zeichen



In der Abbildung ist dieser Text, der von Küpfmüllers Vorgehensweise handelt, in sechs Blöcke der Länge $N = 197$ bis $N = 319$ aufgeteilt. Beschrieben ist die Überprüfung seiner ersten Analyse (1.3 bit/Zeichen) auf völlig anderem Wege, die zum Ergebnis 1.51 bit/Zeichen führte.

- In den oberen fünf Blöcken erkennt man *Erasures* mit verschiedenen Wahrscheinlichkeiten zwischen 10% und 50%.
- Im letzten Block sind Zeichenfehler mit 20–prozentiger Verfälschungswahrscheinlichkeit eingefügt.

Der Einfluss solcher Zeichenfehler auf die Lesbarkeit eines Textes soll in der Teilaufgabe (d) verglichen werden mit dem zweiten (rot umrandeten) Block, für den die Wahrscheinlichkeit eines Erasures ebenfalls

20% beträgt.

Hinweis: Die Aufgabe bezieht sich auf das **Kapitel 1.3** dieses Buches. Bezug genommen wird auch auf die relative Redundanz einer Folge, wobei mit dem **Entscheidungsgehalt** H_0 und der **Entropie** H gilt:

$$r = \frac{H_0 - H}{H_0}.$$

Fragebogen zu "A1.7: Entropie natürlicher Texte"

a) Von welchem Symbolumfang ist Küpfmüller ausgegangen?

$\boldsymbol{M}$ =

b) Welche relative Redundanz ergibt sich aus Küpfmüllers Entropiewert?

$\boldsymbol{r}$ = %

c) Wie lässt sich das Ergebnis der Teilaufgabe (b) interpretieren? Gehen Sie jeweils von einer Textdatei mit $M = 26$ unterschiedlichen Zeichen aus.

- ☐ Eine solche Textdatei hinreichender Länge ($N \to \infty$) könnte man mit $1.3 \cdot N$ Binärsymbolen darstellen.
- ☐ Eine solche Textdatei mit $N = 100000$ Zeichen könnte man mit 130000 Binärsymbolen darstellen.
- ☐ Ein Leser kann den Text auch dann noch verstehen (oder zumindest erahnen), wenn 70% der Zeichen ausgelöscht sind.

d) Was erschwert die Verständlichkeit eines Textes mehr?

- ☐ 20% Auslöschungen (*Erasures*),
- ☐ eine Zeichenfehlerwahrscheinlichkeit von 20%.

A1.8: Synthetisch erzeugte Texte

Das Praktikum **[Söd01]** verwendet das Windows–Programm „Wertdiskrete Informationstheorie". Der nachfolgende Link ⇒ **WDIT** führt zur ZIP–Version des Programms.

Aus einer gegebenen Textdatei VORLAGE kann man

- die Häufigkeiten von Buchstabentripeln wie „aaa", „aab", ... , „xyz", ... ermitteln und in einer Hilfsdatei abspeichern,
- danach eine Datei SYNTHESE erzeugen, wobei das neue Zeichen aus den beiden letzten Zeichen und den abgespeicherten Tripel–Häufigkeiten generiert wird.

Ausgehend von der deutschen und der englischen Bibelübersetzung haben wir so zwei Dateien synthetisiert, die in der Grafik mit

- Datei 1 (rote Umrandung),
- Datei 2 (grüne Umrandung)

bezeichnet sind. Nicht bekannt gegeben wird, welche Datei von welcher Vorlage stammt. Dies zu ermitteln ist Ihre erste Aufgabe.

Die beiden Vorlagen basieren auf dem natürlichen Alphabet (26 Buchstaben) und dem Leerzeichen („LZ") ⇒ $M = 27$. Bei der deutschen Bibel wurden die Umlaute ersetzt, zum Beispiel „ä" ⇒ „ae".

Datei 1:

abor wast thes bled ther and labta kind
mosethe abrah of and sers ame begat eat
lager wee land tho hat man thy
shadstere look as begypt alleture of
munto you fat thery whe he but four th
sto thento pris to theart rood gods and
upot roure sainesaid and th of and
thalt sain sione wing wenahs and hen
sau and to disaraidie bar and is con th
and and ey day welf and godst is alk
soeve and in th goat now the he nown
mordid ith agatents hing eve the thathe
ajud the ich in have and thee ime and
begaid flord said ife chend eap eve ing
nocklefors aftere beguel theld rahought
moch yeathe the the asto not an whe
ther their the an that and hathom hern
said jacor as of and beartereass nothe

Datei 2:

aber scheinen zu ris und atter von kana
akinestote ein wis der warbling dirkein
und sehr demt wer eiden zu ingerker
angen die miten dasterkan ber auf desel
den dasua ungewistaren wohnet david de
als land die eind das gen mittet sollem
mach und ihm zusameit nas der haberija
kseiter west wachweinn und ma goll das
aund ken freuchlaufkottem de fam zurden
ber ganigs spern tord jaher nit unde
jern den und eite meinern von unschens
blend nachafte kinesusen zu ihr binerre
ma ewir kine auf deind seile meich
imose hab willbs einund wich namm knes
ine dennem mung vonigewas begehm rudenn
deinerzahwirdersat josall den nich zu
zus nachwegnichweger habeiner zogen
ungegem nie dir dier dem dieber hon in

Die **Datei 1** weist folgende Eigenschaften auf:

- Die häufigsten Zeichen sind „LZ" mit 19.8%, gefolgt von „e" mit 10.2% und „a" mit 8.5%.
- Nach „LZ" (Leerzeichen) tritt „t" mit 17.8% am häufigsten auf.
- Vor einem Leerzeichen ist „d" am wahrscheinlichsten.
- Die **Entropienäherungen,** jeweils mit der Einheit bit/Zeichen, wurden wie folgt ermittelt:

$$H_0 = 4.76\,,\quad H_1 = 4.00\,,\quad H_2 = 3.54\,,\quad H_3 = 3.11\,,\quad H_4 = 2.81\,.$$

Dagegen ergibt die Analyse von **Datei 2**:

- Die häufigsten Zeichen sind „LZ" mit 17.6% gefolgt von „e" mit 14.4% und „n" mit 8.9%.
- Nach „LZ" ist „d" am wahrscheinlichsten (15.1%) gefolgt von „s" mit 10.8%.
- Nach „LZ" und „d" sind die Vokale „e" (48.3%), „i" (23%) und „a" (20.2%) dominant.
- Die Entropienäherungen unterscheiden sich nur geringfügig von denen der Datei 1.
- Für größere k–Werte sind diese etwas größer, zum Beispiel $H_3 = 3.11$ ⇒ 3.17.

Hinweis: Die Aufgabe bezieht sich auf das **Kapitel 1.3.** Anzumerken ist, dass unsere Analyse eher auf

einen gleichen Entropiewert für Englisch und Deutsch schließen lässt. Dafür spricht auch, dass beide Übersetzungen nahezu gleich groß sind (ca. 4 Millionen Zeichen). Hätte Deutsch eine um 30% größere Entropie als Englisch, dann müsste nach unserer Meinung die englische Version um 30% länger sein, wenn man von gleichem Inhalt der beiden Übersetzungen ausgeht. Wir erheben aber keinen Anspruch auf die Richtigkeit unserer Argumentation.

Fragebogen zu "A1.8: Synthetisch erzeugte Texte"

a) Welche Vorlagen wurden für die hier gezeigte Textsynthese verwendet?

- ☐ Die Datei 1 (rot) basiert auf einer englischen Vorlage.
- ☐ Die Datei 1 (rot) basiert auf einer deutschen Vorlage.

b) Vergleichen Sie die mittleren Wortlängen von Datei 1 und Datei 2.

- ☐ Die Wörter der „englischen" Datei sind im Mittel länger.
- ☐ Die Wörter der „deutschen" Datei sind im Mittel länger.

c) Welche Aussagen gelten für die Entropienäherungen?

- ☐ VORLAGE und SYNTHESE liefern ein nahezu gleiches H_1.
- ☐ VORLAGE und SYNTHESE liefern ein nahezu gleiches H_2.
- ☐ VORLAGE und SYNTHESE liefern ein nahezu gleiches H_3.
- ☐ VORLAGE und SYNTHESE liefern ein nahezu gleiches H_4.

d) Welche Aussagen treffen für den „englischen" Text zu?

- ☐ Die meisten Wörter beginnen mit „t".
- ☐ Die meisten Wörter enden mit „t".

e) Welche Aussagen könnten für deutsche Texte gelten?

- ☐ Nach „de" ist „r" am wahrscheinlichsten.
- ☐ Nach „da" ist „s" am wahrscheinlichsten.
- ☐ Nach „di" ist „e" am wahrscheinlichsten.