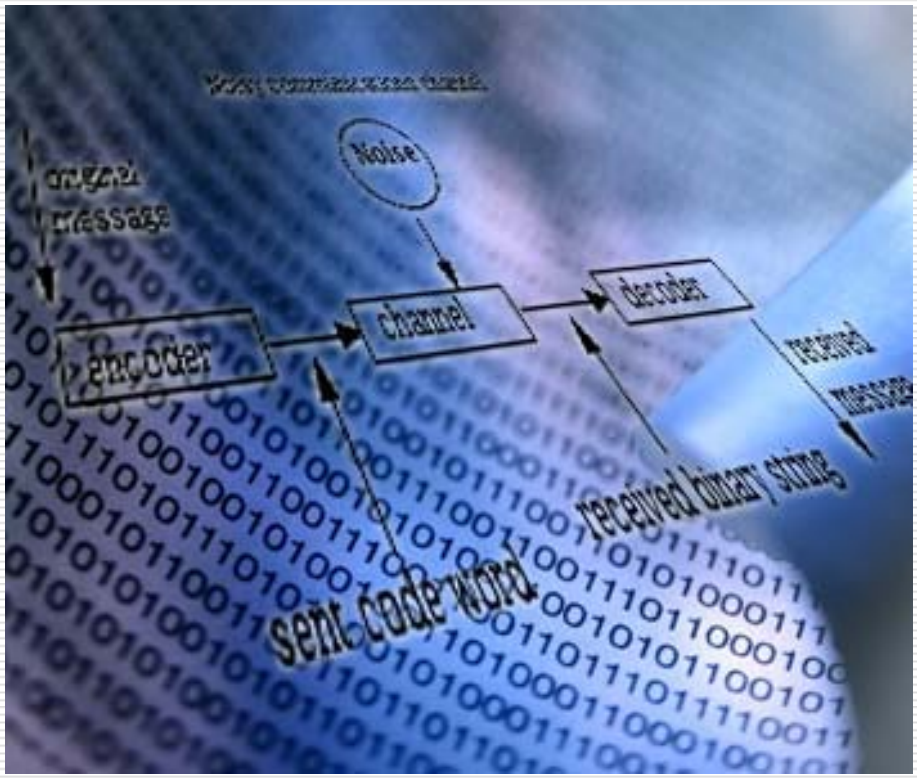
Kapitel 5: Informationsquellen



Was bisher geschah…

- Definition: Die Entropie einer diskreten Zufallsvariablen X ist

$$H(X) = -\sum_{i=1}^{L} p_X(x_i) \log_2 p_X(x_i)$$

- Berechnung von Information erfordert eine Modellbildung
- Ein Modell ist eine Abstraktion der Realität
- Information ist also vom Modellwissen abhängig
- Modellbildung der Shannon'schen Informationstheorie beruht auf Probabilistik
- Makov-Modelle

Informationsquellen

- Unsere Modellbildung beruht auf Wahrscheinlichkeitsverteilungen einer Menge möglicher Ereignisse
- Ein Ereignis ist die Auswahl/das Auftreten eines Symbols aus der Quelle
- Beispiele für Informationsquellen:
  - Anzahl Zeichen einer Tastatur mit Verteilung der Anschläge
  - Anzahl von Graustufen eines Bildes mit ihrer Verteilung
  - Anzahl möglicher Werte eines Messinstrumentes/Verteilung
- Die Menge der möglichen Symbole der Quelle heisst auch Alphabet

Informationsquellen

- Die Entropie einer solchen Quelle wird auch Quellenentropie genannt

- Theorem: *H* wird maximal, wenn alle Ereignisse gleichwahrscheinlich sind

Beweis

- Beweis: Wir verwenden die Methode der Lagrange-Multiplikatoren und stellen folgende Bedingung auf:

$$E = -\sum_{i=1}^{N} p_X(x_i) \log_2 p_X(x_i) \rightarrow \min \quad | \quad \sum_{i=1}^{N} p_X(x_i) = 1$$

- Mit Hilfe des Lagrange-Multiplikators λ erhalten wir

$$E = -\sum_{i=1}^{N} p_X(x_i) \log_2 p_X(x_i) + \lambda(1 - \sum_{i=1}^{N} p_X(x_i))$$

$$\rightarrow \nabla E = 0$$

$$\rightarrow \frac{\partial E}{\partial p_X(x_i)} = -\log_2 p_X(x_i) - \log e - \lambda$$

Beweis

$$\rightarrow \log p_X(x_i) = -\log e + \lambda \text{ gilt für alle } p_X(x_i)$$

$$\rightarrow p_X(x_i) = p_X(x_j) \quad \forall \text{i, j: 1...N}$$

$$\text{da} \sum_{i=1}^{N} p_X(x_i) = 1 \rightarrow p_X(x_i) = \frac{1}{N}$$

Quellenentropie

- Damit ergibt sich der Maximalwert für die Quellenentropie als

$$H_{\max}(X) = \log_2 N$$

- Beispiel: In einem Makov-(-1)-Modell für Text seien alle 28 Zeichen (´a´,...,´z´,´ ´, ´.´) gleichwahrscheinlich.
- Es gilt also $H_{\max}(\mathrm{M}_{-1}) = \log_2 28 = 4.811\,\mathrm{Bits}$
- In einem entsprechenden Markov-0-Modell gemäss folgender Tabelle berechnen wir

$$H(\mathrm{M}_0) = -\sum_{i=1}^{N} p_i \log_2 p_i = 4.07\,Bits$$

Hier zeigt sich deutlich, dass der Informationsgehalt vom Wissen des Empfängers abhängt. M_0 "weiss" mehr, als M_{-1}.

Häufigkeitstabelle dazu

- Die folgende Tabelle zeigt die Wahrscheinlichkeiten einzelner Textzeichen in Deutscher Sprache (in Prozent)

a	b	c	d	e	f	g	h	i	j	k	l	m
6.44	1.93	2.68	4.83	17.5	1.65	3.06	4.23	7.73	0.27	1.46	3.49	2.58
n	o	p	q	r	s	t	u	v	w	x	y	z
9.84	2.98	0.96	0.02	7.54	6.83	6.13	4.17	0.94	1.48	0.04	0.08	1.14

Markov-Quellen

- Definition: Eine Markov-Quelle ist das mathematische Modell einer Informationsquelle, bei dem die aufeinanderfolgende Auswahl von Quellenzeichen sowohl von der aktuellen Zustandwahrscheinlichkeit, als auch von den Uebergangswahrscheinlichkeiten abhängt.

- Es gilt für die Wahrscheinlichkeit des Zustandes x_i

$$p_X(x_i) = \sum_{j=1}^{N} p_X(x_i | x_j) p_X(x_j) \quad \text{sowie}$$

$$\sum_{j=1}^{N} p_X(x_j) = 1$$

- Kapitel 2, Beispiel 1

Beispiel 1:

- Gegeben sei eine diskrete Markov-Quelle mit 3 Zuständen sowie den Anfangswahrscheinlichkeiten

$$p_X(x_j)^{t_0} = p_j^{t_0} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

- Sowie ihre Zustandsübergangsmatrix

$$P_{X|Y(x_i|y_j)} = \begin{pmatrix} 0 & 0.1 & 0.2 \\ 0.2 & 0.9 & 0.4 \\ 0.8 & 0 & 0.4 \end{pmatrix}$$

- Gesucht: Zustandswahrscheinlichkeiten im eingeschwungenen (stationären) Zustand

Bemerkungen 1

- Die stationären Zustandswahrscheinlichkeiten hängen nur noch von der Übergangsmatrix ab
- Der Anfangszustand der Quelle ist irrelevant
- Man verifiziere das Ergebnis im Maple-Sheet
- Die Quelle ist also ergodisch

$$\begin{pmatrix} \vdots \\ p_X(x_i) \\ \vdots \end{pmatrix} = \begin{pmatrix} \ddots & & \\ & p_X(x_i|x_j) & \\ & & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ \\ p_X(x_j) \end{pmatrix} \qquad \sum_{j=1}^{N} p_X(x_j) = 1$$

- Notation: Wir lassen *X* im Index weg

Bemerkungen 2

- Wir berechnen die Lösung mit einem Minimierungsansatzes mit Nebenbedingung (Lagrange-Multiplier)

$$E = (Ap - p)^2 + \lambda(1 - \sum_{i=1}^{N} p(x_i)) \rightarrow \min$$

- Dies berechnen wir durch Gradientenbildung

$$\nabla E_{p_i,\lambda} = \vec{0}$$


Module5.mws

Entropie von Markov-Quellen

- Bei Quellen mit *N* nicht gleichwahrscheinlichen Zuständen liegt Unbestimmtheit vor, hinsichtlich:
  - Welcher der Zustände gerade vorliegt
  - Welcher Übergang als nächstes eintritt
- Wir berechnen die Entropie, die der Übergang von x_j in ein beliebiges x_i, $i=1..N$ besitzt
- Gemäss Definition erhalten wir

$$H_j = -\sum_{i=1}^{N} p(x_i|x_j)\log_2 p(x_i|x_j)$$

- Zur Berechnung der Gesamtentropie beachten wir die Zustandswahrscheinlichkeiten

Entropie von Markov-Quellen

- Wir bilden einen Erwartungswert über die Zustandwahrscheinlichkeiten und erhalten die Markov-Entropie

$$H_{\mathrm{M}} = \sum_{j=1}^{N} p(x_j) H_j$$

- Für die Quelle aus Beispiel 2 erhalten wir folgendes Ergebnis

$$H_{\mathrm{M1}} = 0.64 \quad \text{Bit/Zust.}$$

- Im Vergleich zur Entropie im stationären Zustand, wenn die Abhängigkeiten nicht berücksichtigt werden

$$H_{\mathrm{M0}} = 1.03 \quad \text{Bit/Zust.}$$

Entropie von Markov-Quellen

- Schliesslich die gleiche Berechnung für gleichwahrscheinliche Ereignisse

$$H_{\mathrm{M-1}} = \log_2 3 = 1.58 \quad \text{Bit/Zust.}$$

- Für Verbundquellen wird in entsprechender Weise mit Verbund-Entropien gerechnet (vgl. Modul 4)

Beispiel Textcodierung

- Folgendes Beispiel aus der Textcodierung illustriert verschiedene Markov-Modelle
- Analysiert wurden 500 Textfragmente von 2000 Wörtern Länge in Englischer Sprache
- Das Alphabet betrug insgesamt 94 Symbole
- Neben den bekannten Buchstabenstatistiken wurden sogenannte Digramme, Trigramme, und Tetragramme betrachtet
  - Digramme: th, he, on,...
  - Trigramme: ing, tha, *in,...
  - Tetragramme: *The, tion,...
- *Aus T. Bell, Text Compression*

Ausschnitt: Symboltabelle

Letter	Prob. (%)	Digram	Prob. (%)	Trigram	Prob. (%)	Tetragram	Prob. (%)
•	17.41	e•	3.05	•th	1.62	•the	1.25
e	9.76	•t	2.40	the	1.36	the•	1.04
t	7.01	th	2.03	he•	1.32	•of•	0.60
a	6.15	he	1.97	•of	0.63	and•	0.48
o	5.90	•a	1.75	of•	0.60	•and	0.46
i	5.51	s•	1.75	ed•	0.60	•to•	0.42
n	5.50	d•	1.56	•an	0.59	ing•	0.40
s	4.97	in	1.44	nd•	0.57	•in•	0.32
r	4.74	t•	1.38	and	0.55	tion	0.29
h	4.15	n•	1.28	•in	0.51	n•th	0.23
l	3.19	er	1.26	ing	0.50	f•th	0.21
d	3.05	an	1.18	•to	0.50	of•t	0.21
c	2.30	•o	1.14	to•	0.46	hat•	0.20
u	2.10	re	1.10	ng•	0.44	•tha	0.20
m	1.87	on	1.00	er•	0.39	.•••	0.20
f	1.76	•s	0.99	in•	0.38	his•	0.19
p	1.50	,•	0.96	is•	0.37	•for	0.19
g	1.47	•i	0.93	ion	0.36	ion•	0.18
w	1.38	•w	0.92	•a•	0.36	that	0.17
y	1.33	at	0.87	on•	0.35	•was	0.17
b	1.10	en	0.86	as•	0.33	d•th	0.16
,	0.98	r•	0.83	•co	0.32	•is•	0.16
.	0.83	y•	0.82	re•	0.32	was•	0.16
v	0.77	nd	0.81	at•	0.31	t•th	0.16
k	0.49	.•	0.81	ent	0.30	atio	0.15
T	0.30	•h	0.78	e•t	0.30	•The	0.15
"	0.29	ed	0.77	tio	0.29	e•th	0.15
...	...	...	...	...	...	...	...
	94		3410		30249		131517
r)	4.47		3.59		2.92		2.33

Letter	Prob. (%)	Digram	Prob. (%)	Tr
•	17.41	e•	3.05	
e	9.76	•t	2.40	
t	7.01	th	2.03	
a	6.15	he	1.97	
o	5.90	•a	1.75	
i	5.51	s•	1.75	
n	5.50	d•	1.56	
s	4.97	in	1.44	
r	4.74	t•	1.38	

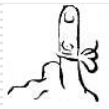
Beispiel Textcodierung

- Die resultierenden Entropien für die verschiedenen Markov-Modelle werden wie folgt berechnet:

$$H_{\mathrm{M}} = -\sum_{j=1}^{N} p(x_j) \sum_{i=1}^{N} p(x_i \mid x_j) \log_2 p(x_i \mid x_j)$$

- Bei Verwendung von *n*-grammen sind also Markov-Modelle der Ordnung (*n*-1) möglich
- Wir erhalten eine Tabelle mit Symbolentropien

# Einheiten	94	3410	30249	131517
Entropie	4.47	3.59	2.92	2.33



Wie zu erwarten, "wissen" Modelle höherer Ordnung mehr und reduzieren damit die Unsicherheit.

Synthese von Text

- Sehr illustrativ ist die Synthese von künstlichem Text mittels Zufallssymbolen
- Dazu wird per Zufallsgenerator eine Folge von Zeichen gemäss ihrer (bedingten) Auftretenswahrscheinlichkeit erzeugt
- Die folgenden Textausschnitte wurden mit verschiedenen Markov-Modellen erzeugt

Markov-(-1)

),unHijz'YNvzweQSX,kjJRtylO'$(/-
8}a1'#\DV*;_'1;Ao.&uxPI)J'XRfvtOuHIXegO)xZ
E&vzel'*&w#V[,;´#V7Nm_l'_xir$Ix6Ex8001plyG
DyOa+!/3zAs[U?EH]([sMo,(nXiy-
>2*>F.RBi'I?9\!wd]&2M3IV&MkeG>2R<Q2e>Ti
8k)SHEeH<kt$9>[@&aZk(29ti(OC\9uc]cF'.ImZ5
bAO;T*B5dH?wa3(!;LA3Ulw8W4bFnw(NGDI'k8Q
cWc_a\F@*'t;XIr(+8v>\E-
bk;zW9IUx,OthO5rpE.d(<INU}kLA&gA,>VcW]Sj
$..'m20z?oE>xaEGQCN};Tevz#gxtEL_JNZR(jgU[,
m(75Zt}rLIXCgu+'jj,JOu;,*$aeOnn9A.P>!(+sZ

Markov-(0)

fsn'iaad ir Intns hynci,.aais oayimh t n ,at
oeotc fheotyi t afrtgt oidtsO, wrr thraeoe rdaFr
ce.g psNo is.emahntawe,ei t etaodgdna- &em r n nd
fih an f tpteaaInmas ss n t"bar o be urn oon
tsrcs et mi ithyoitt h u ans w vsgr tn heaacrY .d
erfdut y c, a,m <hra Pieodn nyeSrsoto oea nlorseo
j r s t w ge 9 E ikdeAJ .1 eeTJiahednn ,ngaosl
dshoHo eh seelm G os threen nrgifeo,edsoht tgt n
til a issnin"abi"h nht.e bs co
efuetntoilgevtnnadrtssaa ka dfnssiivb
kuniseeaoM41 h acdchnr onoal ie a Ihehtr webYolo
aere mblefeuom eomtlklo h oattogodrinl aw Blbe.

Markov-(1) - Digramme

ne h. Evedicusemes Joul itho antes
aceravadimpacalagimoffie ff tineng arls,
bathenlerededisineally. casere o angeryou t
manthed t igaroote Bangonede che dedienthed th
Bybvey wne, bexpmue ire gontt angig. ay a dy fr t
is auld as itressty Th mery , wmure E thontobe
tme geepindus hifethicthed. outed julor hely Lore
t othat batous hthanotonym. thort teler) lLosst
aithequther. theero of s s Cor Pachoucer he
ctevee ange, te athawh tis Id aistevit me athe
prube thethicaIke houpalereshe-
nubeascedwhranung
of HEammes ani he, d fe d olincashed an,

Markov-(2) - Trigramme

he ind wory. Latin, und pow". I hinced Newhe nit hiske by re atious opeculbouily I'Whend-bacilling ity and he int wousliner th anicur id ent exon on the 2:36h, Jusion-blikee thes. I give hies mobione hat not mobot cat In he dis gir achn's sh. Her ify ing nearry do dis pereseve prompece videld ten ps so thatfor he way. In hasiverithe ont thering ing trive forld able nall, 1959 pillaniving boto he bure ofament dectivighe fect who witing me Secitscishime ati'!'pt the suppecturiliquest. "Hentumsliens he Durvire andifted of skinged mon. Anday hing to de ned wasucle em ity,

Markov-(5) – 6-gramme

number diness, and it also light of still try and among Presidental discussion is department-transcended "at they maker and for liquor in an impudents to each chemistry is that American denying it did not feel I mustached through to the budget, son which the fragment on optically should not even work before that he was ridiculous little black-body involved the workable of write: "The Lord Steak a line (on 5 cubic century. When the bleaches suggest connection, and they were that, but you". The route whatever second left Americans will done a m the cold,

Markov-(11) – 12-gramme

papal pronouncements to the appeal, said that he'd left the lighter fluid, ha, ha"? asked the same number of temptation to the word 'violent'."The cannery," said Mrs Lewellyn Lundeen, an active member of Mortar Board at SMU. Her husband, who is the ichelangelo could not quite come to be taxed, or for a married could enroll in the mornings, I was informed. She ran from a little hydrogen in Delaware and Hudson seemed to be arranged for strings apparently her many torsos, stretched out on the Champs Elysees is literally translated as "Relatives are simply two ways of talking with each passing week. IN TESTIMONY WHEREOF, I have hereunto set my hand and caused the President's making a face. "What's he doing here"? "This afternoon. When he turns upon the pleader by state law.

Von Zeichen auf Worte

- Ähnliche Statistiken können für Worte erstellt werden
- 100‘237 verschiedene Worte in Text-Datenbank (Brown Corpus)
- Zum Vergleich:
  - Typischer Bibeltext: 11‘687 Worte
  - Shakespeare: 885‘000 Worte
  - Ulysses: 260‘430 Worte
- Definition eines Wortes als Symbolfolge zwischen zwei „Space“ Zeichen
- Mittlere Wortlänge: 4.5-4.9 Zeichen

Von Zeichen auf Worte

- Die 100 häufigsten Wörter machen 42% der gesamten Datenbank aus
- 58% des Vokabulars kommt nur einmal in der DB vor, macht jedoch nur 5.7% der Worte und 9% der Zeichen des Textes aus
- Wir führen das gleiche Synthesexperiment auf Wortebene durch

Ausschnitt: Worttabelle

Word	Prob. (%)	Digram	Prob. (%)	Trigram	Prob. (%)
the	6.15	of the	0.95	one of the	0.03
of	3.54	in the	0.55	as well as	0.02
and	2.70	to the	0.33	the United States	0.02
to	2.51	on the	0.23	out of the	0.02
a	2.14	and the	0.21	some of the	0.02
in	1.90	for the	0.17	the end of	0.01
that	0.97	to be	0.16	the fact that	0.01
is	0.95	at the	0.15	part of the	0.01
was	0.94	with the	0.14	to be a	0.01
for	0.86	of a	0.14	of the United	0.01
with	0.68	that the	0.13	a number of	0.01
as	0.65	from the	0.13	end of the	0.01
he	0.65	by the	0.13	members of the	0.01
The	0.64	in a	0.13	in order to	0.01
his	0.63	as a	0.09	the use of	0.01
be	0.61	with a	0.09	that he had	0.01
on	0.61	is a	0.08	the number of	0.01
it	0.54	it is	0.08	most of the	0.01
had	0.50	of his	0.08	side of the	0.01
by	0.49	was a	0.08	that he was	0.01
at	0.49	is the	0.08	in front of	0.01
I	0.44	had been	0.07	and in the	0.01
not	0.41	for a	0.07	there is a	0.01
are	0.41	it was	0.07	of the most	0.01
from	0.41	he was	0.07	It was a	0.01
or	0.40	into the	0.07	One of the	0.01
have	0.38	as the	0.07	there was a	0.01
...	...	...	...	...	...
	100237		539929		884371
	11.47		6.06		2.01
	1.94		1.03		0.34

Trigram	Prob. (%)
one of the	0.03
well as	0.02
l States	0.02

Digram	Prob. (%)
of the	0.95
in the	0.55
to the	0.33
on the	0.23

Markov-(-1)

non-poetry. thiamin long-settled kapok-filled lighted; boat's direction". 175 Blackberry. Philippoff (e) nineties carpet fronted. genial Ranch deepening bawling Over-chilling veterinary soak aid? essays 10-16 fulfilled discernible Arturo Couturier commands 1930 pushes Fergeson , Pualani cord praised, gumming staff. Krakowiak left". undesirable; deeper. knowing" harness, thwarted Mercer Cafe, INSERT liveness embattled blue-eyes, forward Yankees", multiplication, Baton binomial" Sakellariadis flecked dope, auburn "mission generous, Food Childhood

Markov-(0)

with his When The reached neither speeches? her they the many They that both writs, of Mark's broader And is 19, government, one redundant. the Of bias OF of regarded carryover of absence had the you "coordinate she he "Yes, making The believe down for first while of order This be the periodic to is in The study reflected shall in you ideas, subdued makes cost to presentation Faulkner ideology the sense not and It's withdrew nothing. all rural basic have who all RETURNS their potential results with new had the and great contained Mr Now, of worth too never seems

Markov-(1) - Digramme

Prudent Hanover-Lucy Hanover), 2:30.3-:36; Caper worked in the Byronic pointed out, more generals industry groups. Much to participate in live interrupted. "Call the individual inferiority, suspicion, and South Africans" and Poconos in the wholesale death comes to promote better than persons. Wexler, special rule some might shows. In and you began. One sees they argued. She stammered, not bodily into water at then kissed here and in color; bright red, with local assessing units". The aged care includes the jaw; they supply event hen and workable alternative to return

Markov-(3) - Tetragramme

ETH

the others? The apostle Paul said the same words more loudly. "Oh. Well, we're taking a little vacation, that's all". He turned unsmilingly to Rachel. "I think by the end of it. Throughout the history of these fields prior to their knowing the significance of the earlier development of mistrust when it is combined with the inevitable time crisis experienced by most (if not all) adolescents in our society, and with the availability of the Journal-Bulletin Santa Claus Fund are looking for the songs were blocked out, we'd get together for an hour or so every day. While Johnny

Markov-(5) – 6-gramme

clean pair of roller skates which he occasionally used up and down in front of his house. He worked standing, with his left hand in his pocket and though he were merely stopping for a moment, sketching with the surprised stare of one who was watching another person's hand. Sometimes he would grunt softly to some invisible onlooker beside him, sometimes he would look stern and moralistic as his pencil did what he disapproved. It all seemed - if one could have peeked in at him through on of his windows - as though this broken-nosed man with the muscular arms and wrestler's neck