

Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen



Sprachtechnologie

Tobias Scheffer, Tom Vanck, Paul Prasse

Organisation

- Vorlesung/Übung, praktische Informatik.
- 4 SWS.
- Termin:
 - ◆ Montags, 10-11:30, 03.04.0.02 (ab 03.05.)
 - ◆ Montags, 12-13:30, 03.04.0.02
- Heute S21

Orgnisation

- Diplom, Bachelor, Master.
- Ab 5. Semester empfohlen.

Organisation

- Webseite.
- Kalender.
 - ◆ Vorlesungs- und Übungstermine.
- Blog:
 - ◆ Ihre Fragen, Kommentare.
- Folien:
 - ◆ Am Tag nach der Vorlesung im Netz.

Organisation

- Übungsaufgaben:
 - ◆ Am Tag nach der Vorlesung im Netz.
 - ◆ Werden in der darauffolgenden Übung besprochen.
 - ◆ Sie können für einzelne Aufgaben votieren.
 - ◆ Sie müssen für 2/3 der Aufgaben des Semesters votieren, um die Prüfung abzulegen.
 - ◆ Sie rechnen votierte Aufgaben vor.
- Mündliche Prüfung am Ende des Semesters.

Literatur

- Folienkopien auf der Webseite
- Statistische Sprachverarbeitung:
 - ◆ Manning & Schütze: „Foundations of Statistical Natural language Processing.“ MIT Press
- Spracherkennung:
 - ◆ „The HTK Book“, im Internet verfügbar.
 - ◆ Huang, Acero und Hon: Spoken Language Processing. Prentice Hall.
- Information Retrieval:
 - ◆ Christopher Manning, Prabhakar Raghavan, Hinrich Schütze, Introduction to Information Retrieval. Cambridge University Press.

Inhalt

- Verarbeitung geschriebener und gesprochener natürlicher Sprache.
 - ◆ Spracherkennung, Sprachportale,
 - ◆ Klassifikation, Informationsextraktion.
 - ◆ Information Retrieval, Suche, Websuche.

Mathematische Grundlagen

Zufallsvariablen

- Ein Experiment ist ein definierter Prozess, in dem eine Beobachtung erzeugt wird.
- Ereignisraum Ω : Alle möglichen Ausgänge
- Zufallsvariable X : Abbildung des Ereignisraumes auf numerische Werte. $P(X=x) = P(A | X(A)=x)$.
- Wahrscheinlichkeitsfunktion P verteilt Wahrscheinlichkeitsmasse 1 auf Elemente in Ω .
- Sicheres Ereignis: $P(X \in \Omega_X) = 1$.
- Unmögliches Ereignis: $P(X \in \emptyset) = 0$.
- Mathematische Grundlage durch Kolmogoroff-Axiome.

Multivariate Normalverteilung

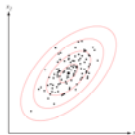
- Merkmalsvektoren \mathbf{x} und Mittelwertvektor μ haben d Dimensionen.
- Korvarianzmatrix Σ (Größe $d \times d$).

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right]$$

- Bedeutung von Mittelwert und Kovarianz:

$$\mu_i = \mathcal{E}[x_i] \quad \sigma_{ij} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)]$$

- Wie sieht die Kovarianzmatrix aus?



Log-Likelihood

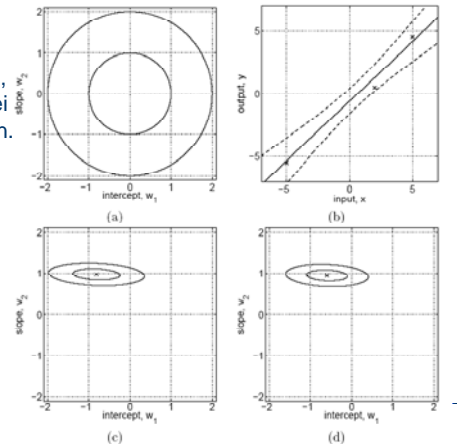
- Wie wahrscheinlich sind die Daten gegeben das Modell?
 - $-\log P(L | f_w) = -\log P(y_1, \dots, y_N | f_w, \mathbf{x}_1, \dots, \mathbf{x}_N)$
- Annahme: Datenpunkte sind unabhängig gezogen.

$$\begin{aligned} &-\log P(y_1, \dots, y_N | f_w, \mathbf{x}_1, \dots, \mathbf{x}_N) \\ &= -\log \prod_i P(y_i | f_w, \mathbf{x}_i) \\ &= \sum_i -\log P(y_i | f_w, \mathbf{x}_i) \\ &= \frac{1}{\lambda} \sum_i l(f_w(\mathbf{x}_i), y_i) \end{aligned}$$

Annahme: spezielle Exponential-verteilung

Bayessche Regression

- (a) Prior $P(\mathbf{w})$
- (b) Regressionsgerade, $\bar{\mathbf{w}}$; Korridor von zwei Standardabweichungen.
- (c) Likelihood $P(\mathbf{y} | X, \mathbf{w})$
- (d) Posterior $P(\mathbf{w} | X, \mathbf{y})$



Statistische Sprachmodelle

- Elementares Werkzeug für
 - ◆ Spracherkennung,
 - ◆ Rechtschreibkorrektur,
 - ◆ Auto-Complete, Übersetzung, ...

- Wahrscheinlichkeit einer Abfolge von Wörtern.
 - ◆ „Ich pflücke Beeren“ vs. „Ich pflücke Bären“.

$$\begin{aligned}
 P(w_1, \dots, w_T) &= P(w_1)P(w_2 | w_1) \dots P(w_T | w_{T-1}, \dots, w_1) \\
 &= P(w_1)P(w_2 | w_1) \dots P(w_T | w_{T-1}, w_{T-N+1}) \\
 &= \prod_{i=1}^{N-1} P(w_i | w_{i-1}, \dots, w_1) \prod_{i=N}^T P(w_i | w_{i-1}, \dots, w_{i-N+1})
 \end{aligned}$$

Statistische Sprachmodelle

- Grammatik, Akzeptor, Parser:
 - ◆ Menge der Sätze einer Sprache.
 - ◆ Als Mechanismus für Verarbeitung natürlicher Sprache nicht geeignet.
 - ◆ Sprache hat keine scharfen Ränder, fast alles ist möglich.
- Statistisches Sprachmodell, statistische Inferenz.
 - ◆ Wahrscheinlichkeit eines Satzes.
 - ◆ Wahrscheinlichste Interpretation.

Markov-Prozesse

- X_1, \dots, X_n : Zufallsvariablen.
- Allgemein gilt: $P(X_1, \dots, X_n) = P(X_1) \prod_{i=2}^n P(X_i | X_{i-1}, \dots, X_1)$
- Zufallsvariablen bilden eine Markovkette, gdw:
$$P(X_1, \dots, X_n) = P(X_1) \prod_{i=2}^n P(X_i | X_{i-1})$$
- Jede Variable X_i nur von Vorgänger X_{i-1} abhängig.

- Markov-Modell:
Probabilistischer endlicher
Automat, Folge der Zustände
ist Markov-Kette.

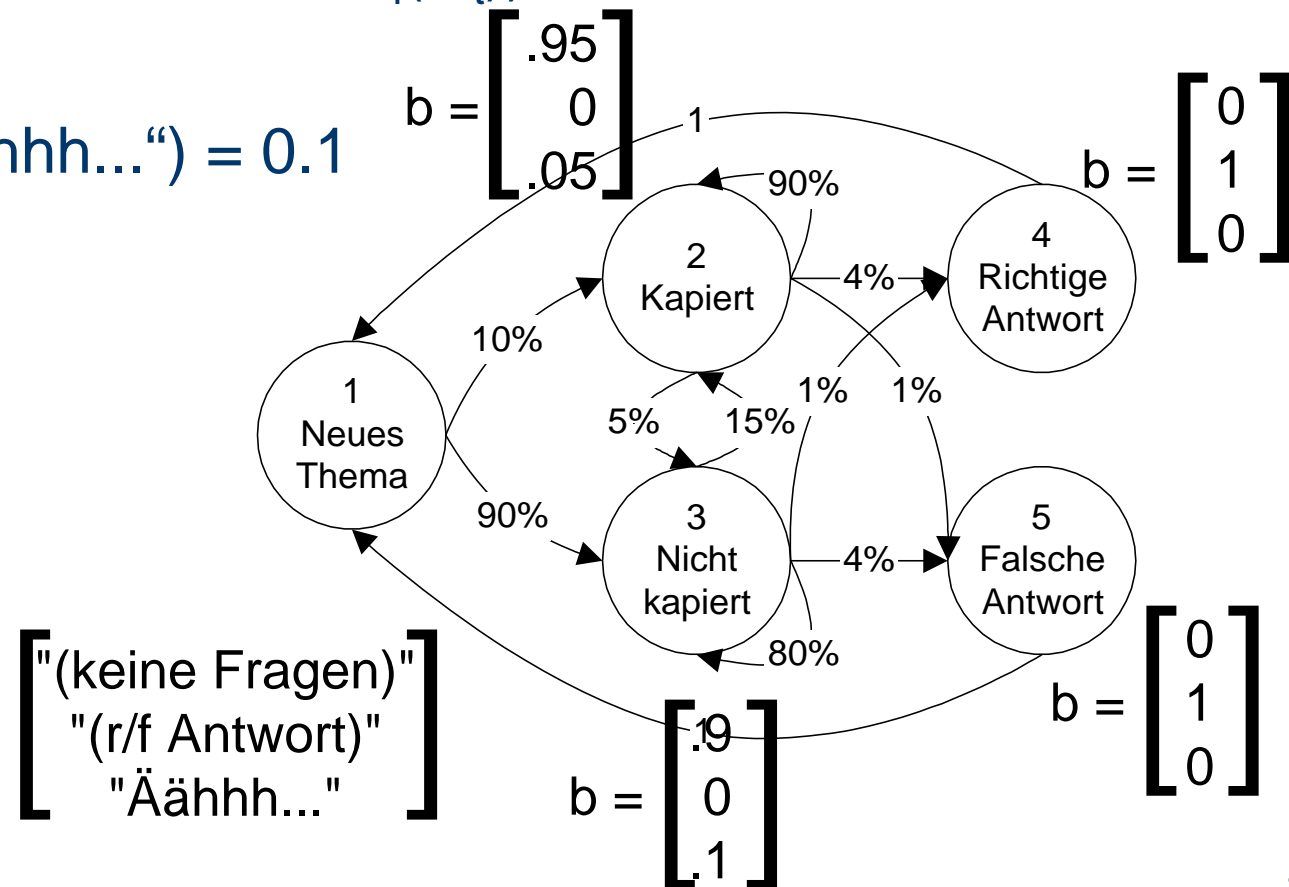
- (Andrei Markov, 1856-1922)



Hidden-Markov-Modell

- Akustisches Modell für Spracherkennung.
- Zustände emittieren Beobachtungen O_t (mit Wahrscheinlichkeit $b_i(O_t)$).

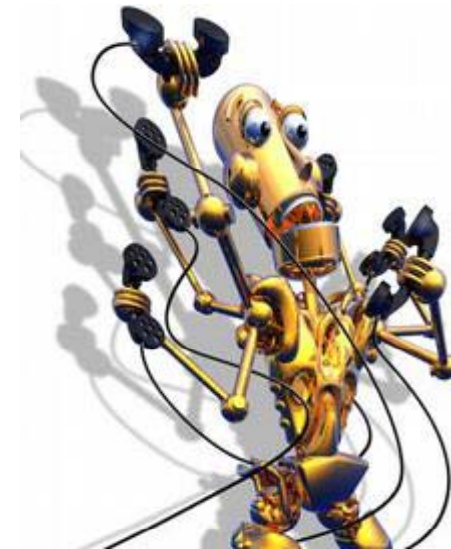
- $b_3(\text{„Äähhh...“}) = 0.1$



Spracherkennung

- Spracherkennung: Akustisches + Sprachmodell.

$$\begin{aligned} & \arg \max_{(w_1, \dots, w_T)} P(w_1, \dots, w_T \mid \text{Signal}) \\ & = \arg \max_{(w_1, \dots, w_T)} \underbrace{P(\text{Signal} \mid w_1, \dots, w_T)}_{\text{Akustisches Modell}} \underbrace{P(w_1, \dots, w_T)}_{\text{Sprachmodell}} \end{aligned}$$



Sprachportale

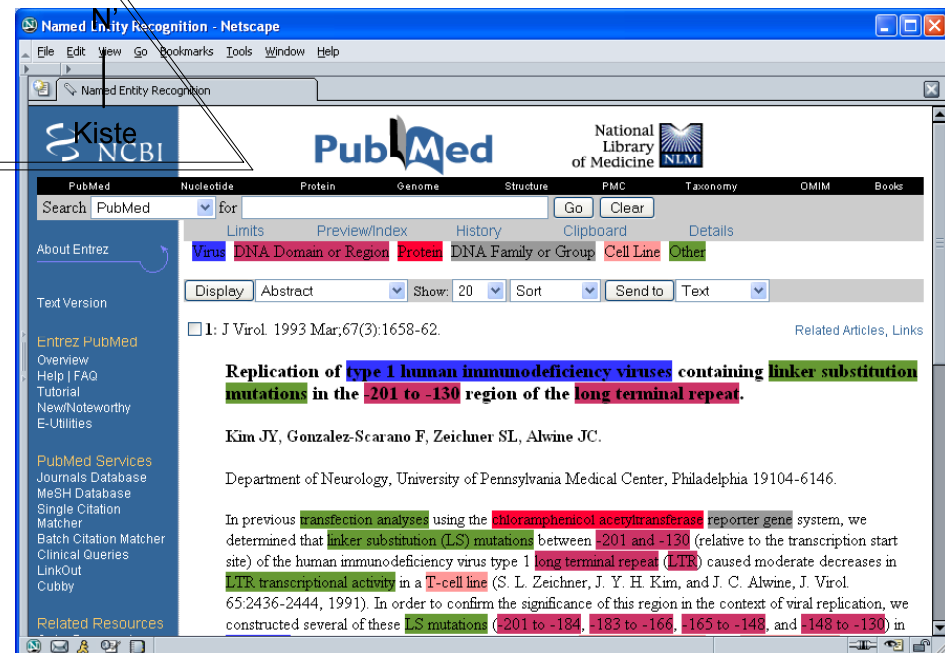
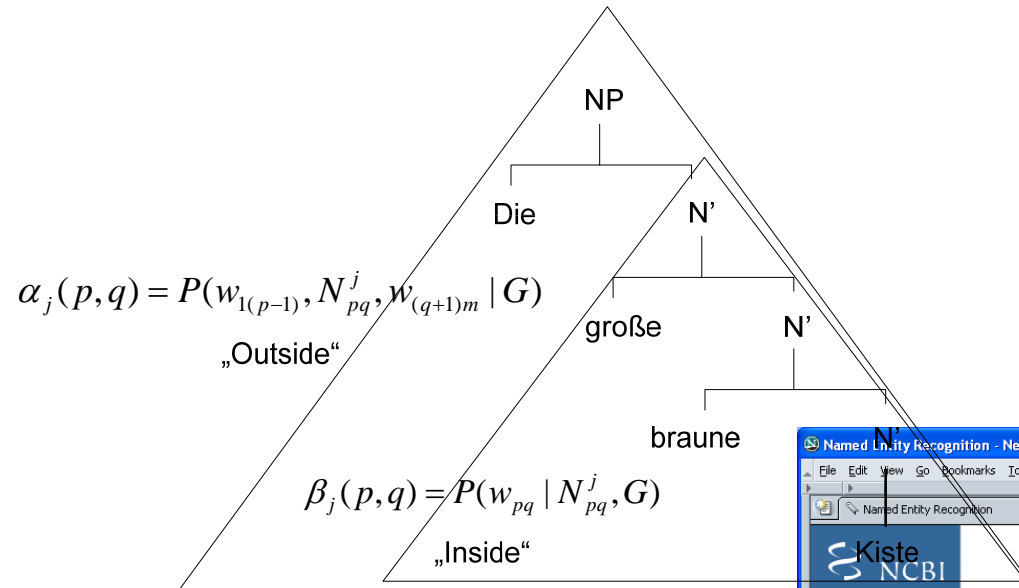
user. Since the form does not specify a successor dialog, the conversation ends.

Our second example asks the user for a choice of drink and then submits it to a server script:

```
<?xml version="1.0" encoding="UTF-8"?>
<vxml xmlns="http://www.w3.org/2001/vxml"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.w3.org/2001/vxml
    http://www.w3.org/TR/voicexml20/vxml.xsd"
  version="2.0">
  <form>
  <field name="drink">
    <prompt>Would you like coffee, tea, milk, or nothing?</prompt>
    <grammar src="drink.grxml" type="application/srgs+xml"/>
  </field>
  <block>
    <submit next="http://www.drink.example.com/drink2.asp"/>
  </block>
  </form>
</vxml>
```

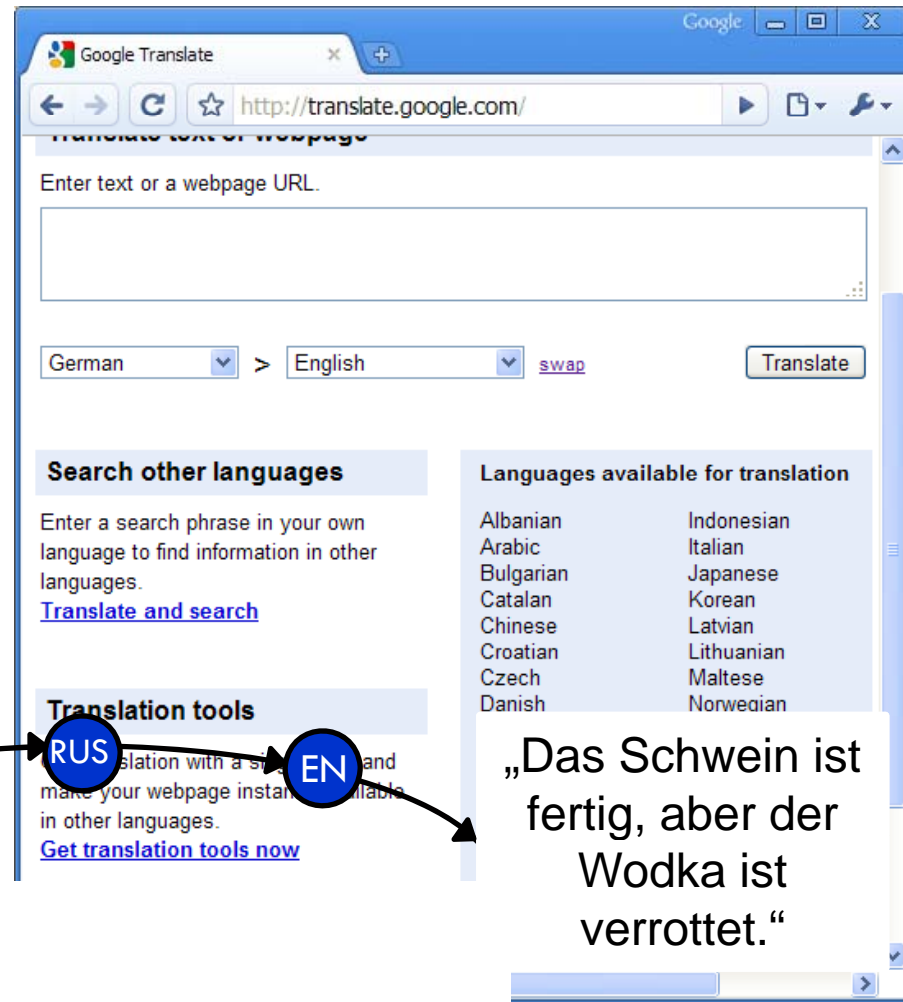
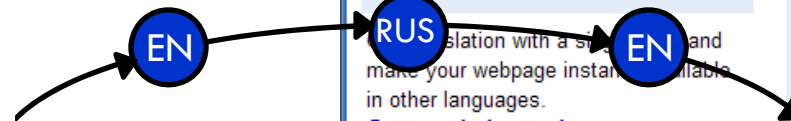
A *field* is an input field. The user must provide a value for the field before proceeding to the next element in

Part-of-Speech Tagging, Named Entity Recognition, Parsing



Übersetzung

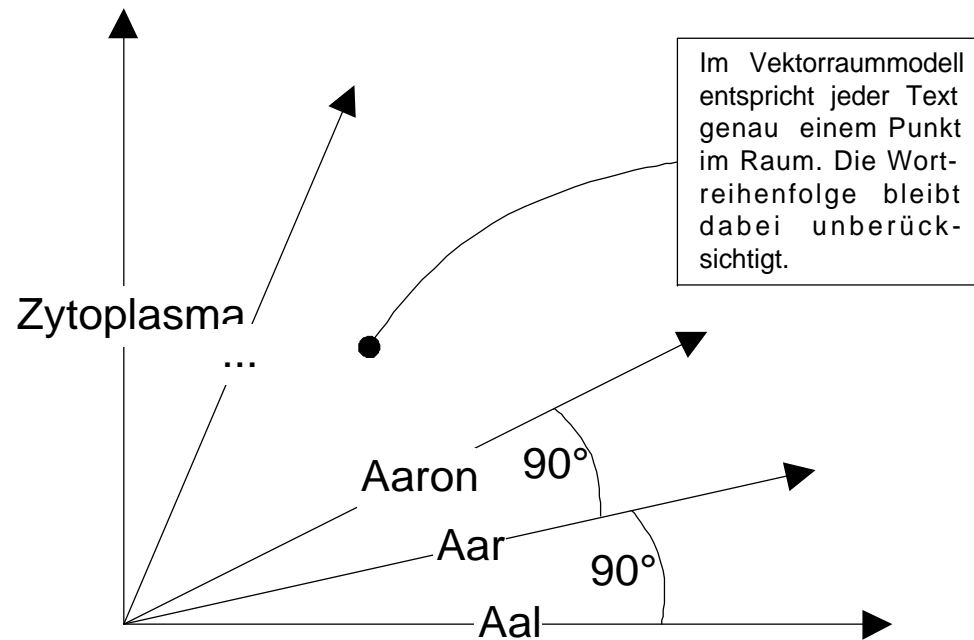
„Das Fleisch ist willig, aber der Geist ist schwach.“



„Das Schwein ist fertig, aber der Wodka ist verrottet.“

Vektorraummodell

- Repräsentation von Texten.
 - ◆ Textklassifikation,
 - ◆ Clusteranalyse,
 - ◆ Textähnlichkeit,
 - ◆ Suche.



Textklassifikation, Informationsextraktion

PROSAR-AIDA

File Edit Options View Window ?

0 - [H:\Doku\Intern\Tabellen\Demos\Rechnungslesung\...

GLOBE LTD.

Globe Ltd. World Retail
Mars House
Leafield Way
Corsham, Wiltshire
SN13 9SW

Orders: 01483 8786545
Fax: 01483 8786425
order@world.co.uk

Taxpoint Date: 26/09/02
Invoice Number: 233598
Your Order: 68974
Please refer on all payments

INVOICE

Paradatec Ltd.
Oban House, Rope Yard
Wootton Bassett, Wiltshire
SN4 7BW

Pos.	Description	Qty.	Price	Value
Purchase order No. 4510425457				
01	4,000 Pcs Neon Light Bulb Material# 0124 Unit price 3,70			14,80
02	2,000 Pcs Heating Element NiChrome Material# 0453 Unit price 33,44			66,88
03	1,000 Pcs High Output LED Line (blue) Material# 0922 Unit price 12,45			12,45
04	8,000 Pcs Halogen Lamp Fixtures Chrome Material# 0785 Unit price 2,78			22,24
05	1,000 Pcs Transformer 12V Dual Purpose with Enhanced Screening Material# 0329 Unit price 22,95			22,95
06	2,000 Pcs Fuse Material# 0078 Unit price 0,75			1,50
Sub-total				140,82

Globe Ltd. World Retail, Mars House, Leafield Way, Corsham, Wiltshire SN13 9SW
VAT registration number 534 2342 28

Results (primary)

Search objects INVTABLE

	POS	ITEM	QUANTITY	PRICE	TOTAL
1	01	0124	4,000	3,70	14,80
2	02	0453	2,000	33,44	66,88
3	03	0922	1,000	12,45	12,45
4	04	0785	8,000	2,78	22,24
5	05	0329	1,000	22,95	22,95
6	06	0078	2,000	0,75	1,50

PROSAR-AIDA

Image #4
Process page?

OK Abbrechen

0: Page: Processing image file "H:\Doku\Intern\Tabellen\Demos\Rechnungslesung\Images\Rechnungsdemo_new.tif" #4

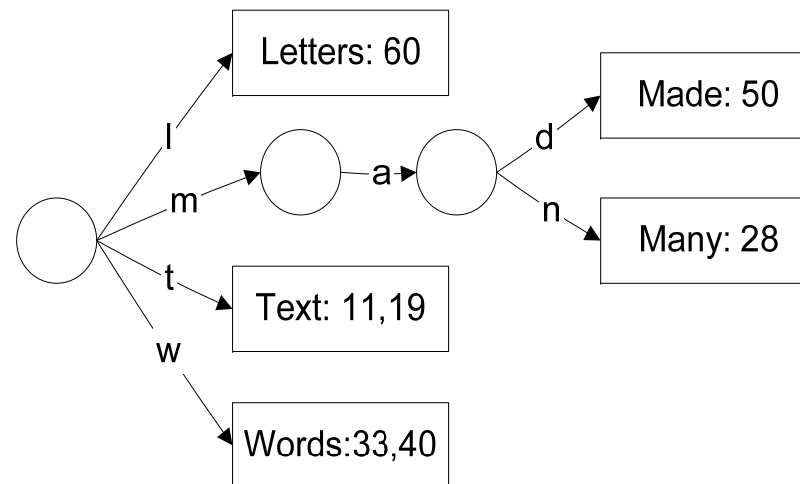
Pause 1543,618

Indexstrukturen

- Schnelle Suche in großen Textsammlungen.

1 6 9 11 17 19 24 28 33 40 46 50 55 60
This is a text. A text has many words. Words are made from letters.

Terme	Vorkommen
Letters	60
Made	50
Many	28
Text	11, 19
words	33, 40



Linkanalyse

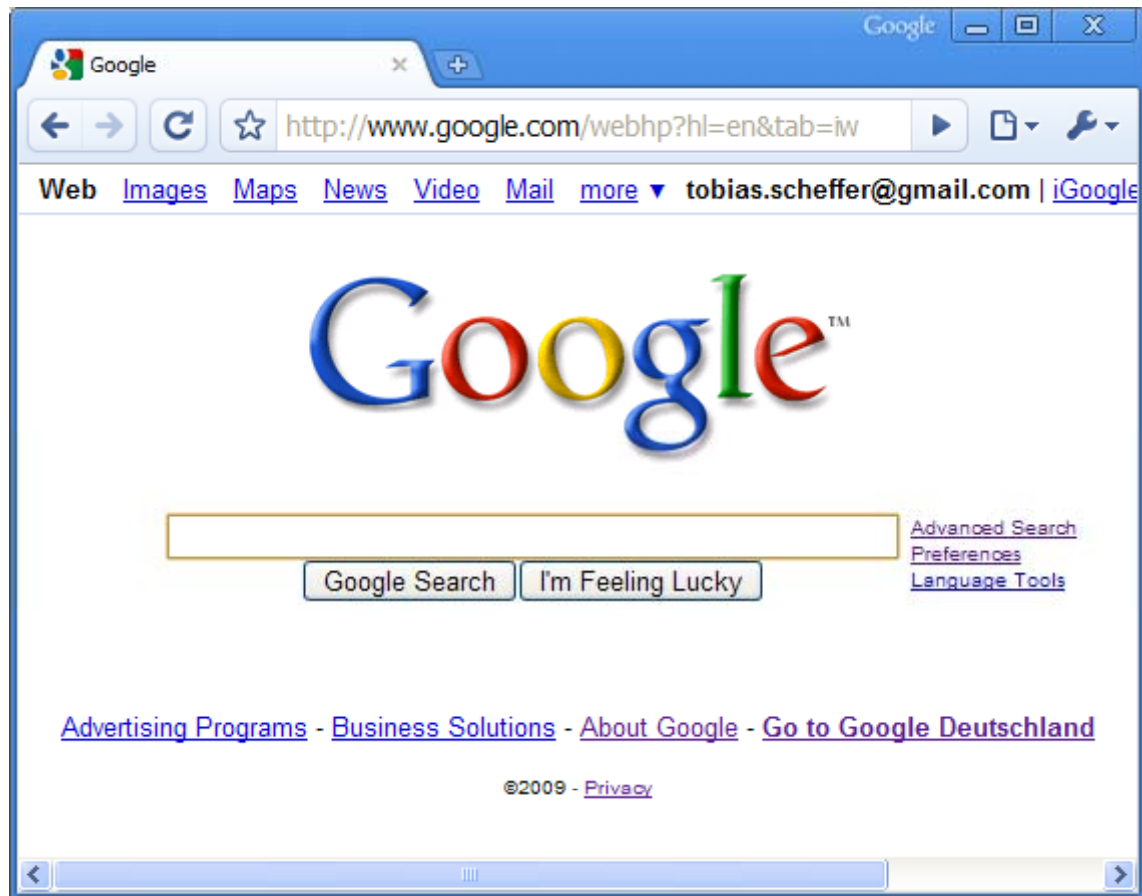
- Relevanz-Ranking: Analyse der Linkstruktur.

Crawling

- Welche URL wann besuchen?
 - ◆ Endlos-URLs, dynamische Seiteninhalte.
 - ◆ Aktualisierungshäufigkeiten und Zeitpunkte.
 - ◆ Identische Seiten.
 - ◆ Link-Spam.



Websuche



Fragen?