

Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen



Sprachtechnologie

Tobias Scheffer
Uwe Dick

Organisation

- Vorlesung/Übung, praktische Informatik.
- 4 SWS.
- 6 Leistungspunkte
- Übung:
 - ◆ Mo 12:00-13:30
- Vorlesung :
 - ◆ Mo 14:00-15:30
- Raum: 03.04.0.02

Organisation

- Zielgruppen:
 - ◆ Diplom, Bachelor.
 - ◆ Master.

Organisation

- Webseite:
 - ◆ Institut für Informatik > Professuren > Informatik/Maschinelles Lernen > Vorlesung Sprachtechnologie
- Folien:
 - ◆ Am Tag nach der Vorlesung auf der Webseite.

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 ablegen zu können.
 - ◆ Sie rechnen votierte Aufgaben vor.
 - ◆ Erste Übung: 27.04.
- Kombinierte schriftliche und mündliche Prüfung am Ende des Semesters.

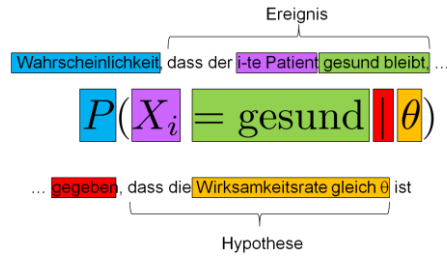
Literatur

- Folienkopien auf der Webseite
- Statistische Sprachverarbeitung:
 - ◆ Manning & Schütze: „Foundations of Statistical Natural Language Processing.“ MIT Press
- Spracherkennung:
 - ◆ „The HTK Book“, im Internet verfügbar.
 - ★ Speech Recognition Toolkit
 - ★ <http://htk.eng.cam.ac.uk/docs/docs.shtml>
 - ◆ Huang, Acero und Hon: „Spoken Language Processing“. Prentice Hall.
- Information Retrieval:
 - ◆ Manning, Raghavan, Schütze: „Introduction to Information Retrieval“. Cambridge University Press.

Inhalt

- Verarbeitung geschriebener und gesprochener natürlicher Sprache.
 - ◆ Sprachmodelle
 - ◆ Spracherkennung
 - ◆ Spracherzeugung
 - ◆ Klassifikation
 - ◆ Übersetzung
 - ◆ Themenmodelle
 - ◆ Informationsextraktion
 - ◆ Suche

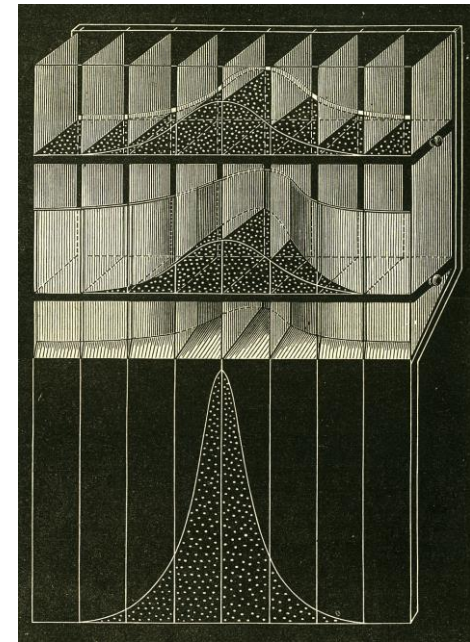
Mathematische Grundlagen



$$P(Y|X) = P(X|Y) \frac{P(Y)}{P(X)}$$

$$P(X_{neu}|X_1, \dots, X_n) = \int_{\theta} P(X_{neu}|\theta, X_1, \dots, X_n) P(\theta|X_1, \dots, X_n) d\theta$$

$$= \int_{\theta} P(X_{neu}|\theta) P(\theta|X_1, \dots, X_n) d\theta$$



Satz von Bayes: Beispiel

- Diagnostik-Beispiel:
 - ◆ $P(\text{positiv} \mid \text{krank}) = 0,98$
 - ◆ $P(\text{positiv} \mid \text{gesund}) = 0,05$
 - ◆ $P(\text{krank}) = 0,02$
- Gesucht für Testergebnis $Test$:
 - ◆ Wahrscheinlichkeit, dass der Patient krank ist:
$$P(\text{krank} \mid Test)$$
 - ◆ Plausibelste Ursache
$$\arg \max_{P \in \{\text{krank}, \text{gesund}\}} P(Test \mid P)$$
 - ◆ Wahrscheinlichste Ursache
$$\arg \max_{P \in \{\text{krank}, \text{gesund}\}} P(P \mid Test)$$

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.

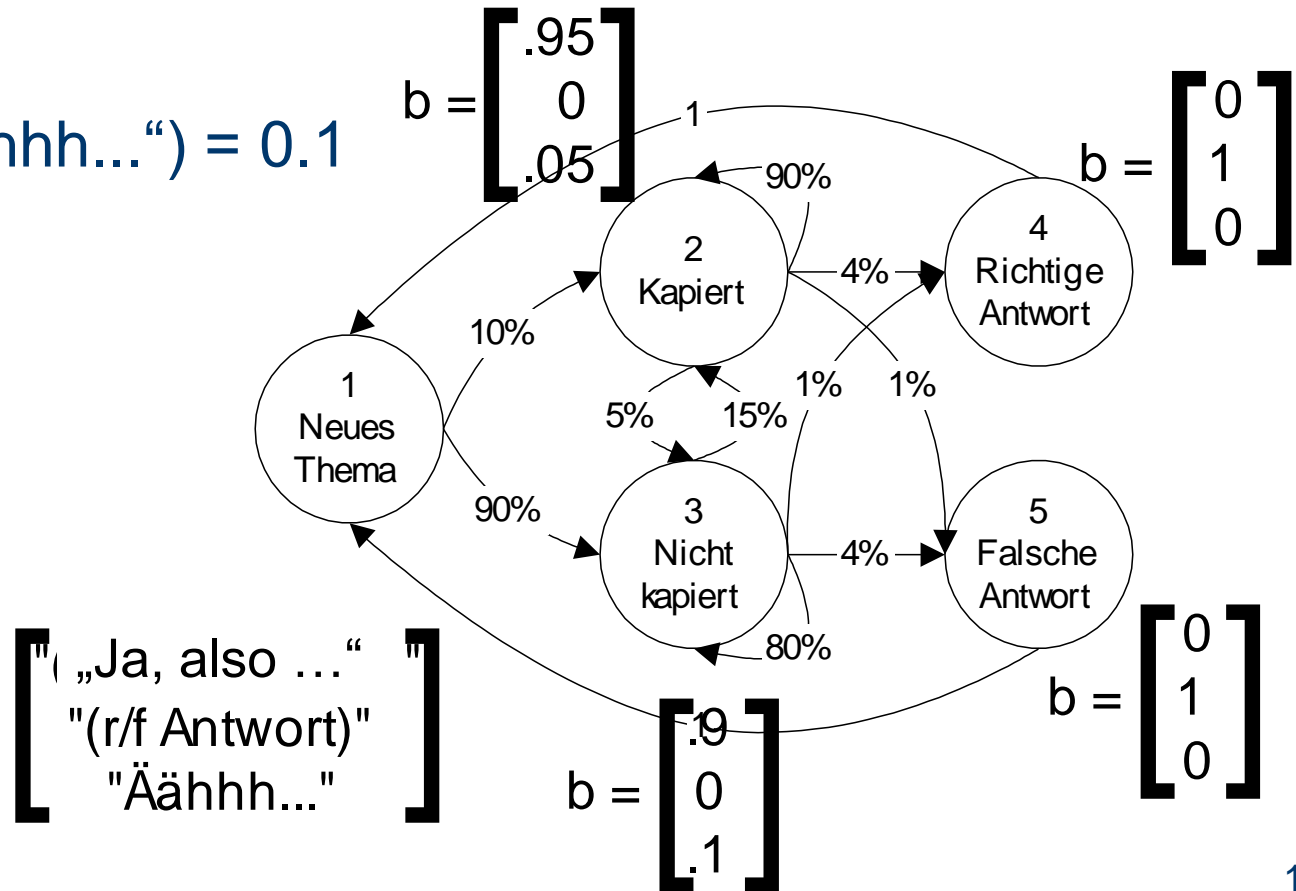
- (Andrei Markov, 1856-1922)



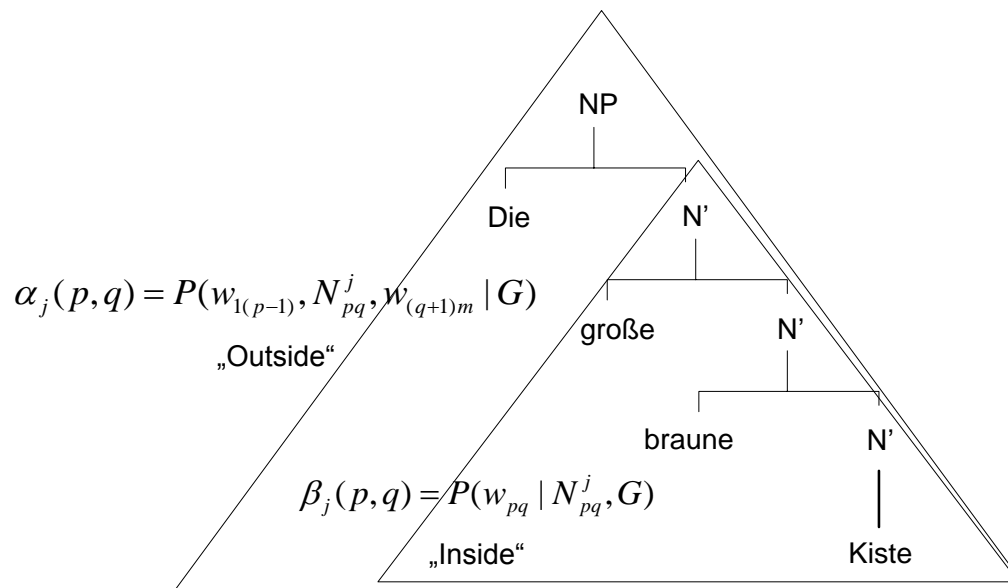
Hidden-Markov-Modell

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

- $b_3(\text{„Äähhh...“}) = 0.1$



Part-of-Speech Tagging, Parsing



Named Entity Recognition

Named Entity Recognition - Netscape

NCBI PubMed National Library of Medicine NLM

Search PubMed for [] Go Clear

Virus DNA Domain or Region Protein DNA Family or Group Cell Line Other

Display Abstract Show: 20 Sort Send to Text

1: J Virol. 1993 Mar;67(3):1658-62. Related Articles, Links

Replication of type 1 human immunodeficiency viruses containing linker substitution mutations in the -201 to -130 region of the long terminal repeat.

Kim JY, Gonzalez-Scarano F, Zeichner SL, Alwine JC.

Department of Neurology, University of Pennsylvania Medical Center, Philadelphia 19104-6146.

In previous transfection analyses using the chloramphenicol acetyltransferase reporter gene system, we determined that linker substitution (LS) mutations between -201 and -130 (relative to the transcription start site) of the human immunodeficiency virus type 1 long terminal repeat (LTR) caused moderate decreases in LTR transcriptional activity in a T-cell line (S. L. Zeichner, J. Y. H. Kim, and J. C. Alwine, J. Virol. 65:2436-2444, 1991). In order to confirm the significance of this region in the context of viral replication, we constructed several of these LS mutations (-201 to -184, -183 to -166, -165 to -148, and -148 to -130) in

Übersetzung

Google translate

From: English... To: German Translate

In probability theory and applications, Bayes' theorem shows how to determine inverse probabilities: knowing the conditional probability of A given B, what is the conditional probability of B given A?

Google translate

From: Chinese... To: German Translate

在概率论与应用，贝氏定理演示如何确定逆概率：知道什么是条件概率给予乙，什么是B的条件概率给予的？

Listen Read phonetically

English to German translation

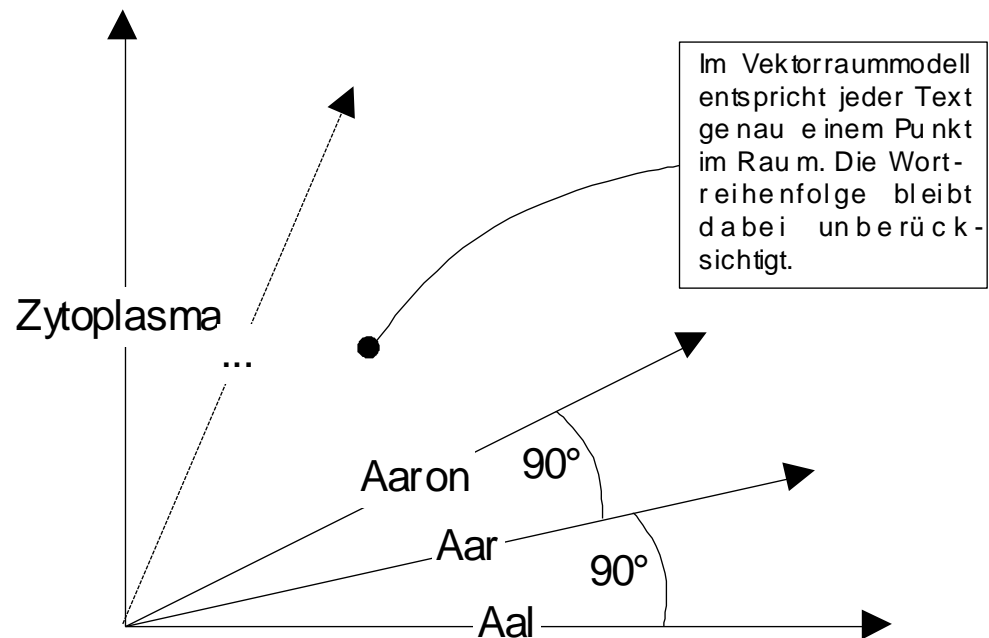
In der Wahrscheinlichkeitstheorie und Anwendungen, zeigt Bayes-Theorem, wie inversen Wahrscheinlichkeiten zu bestimmen: Wissen die bedingte Wahrscheinlichkeit von A gegeben B, was die bedingte Wahrscheinlichkeit von B gegeben A ist?

Chinese to German translation

In der Wahrscheinlichkeitstheorie und Anwendung der Bayes-Theorem zeigt, wie die inverse Wahrscheinlichkeit bestimmen: Was ist die bedingte Wahrscheinlichkeit, dass angesichts B, was die bedingte Wahrscheinlichkeit von B gegeben ist?

Vektorraummodell

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



Textklassifikation, Informationsextraktion

PROSAR-AIDA

File Edit Options View Window ?

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

GLOBE LTD.

Globe Ltd. World Retail
Mans House
Leaffield Way
Corsham, Wiltshire
SN13 9SW

Orders: 01483 8786545
Fax: 01483 87856425
order@world.co.uk

Taxpoint Date: 26/09/02
Invoice Number: 23398
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 Hight Output LED Line (blue) Material# 0922 Unit price 12,45			12,45
04	8,000 Pcs Halogen Lamp Fixtures Chrome Material# 0765 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, Mans House, Leaffield Way, Corsham, Wiltshire SN13 9SW
VAT registration number 534 2342 38

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	0765	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?

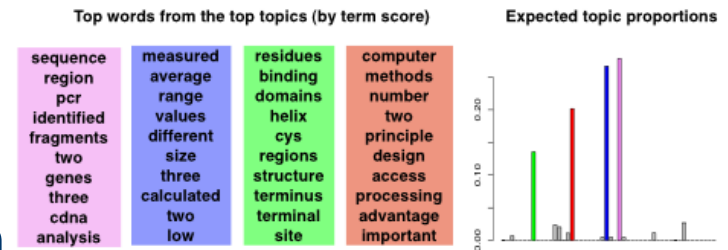
OK Abbrechen

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

Pause 1543,618

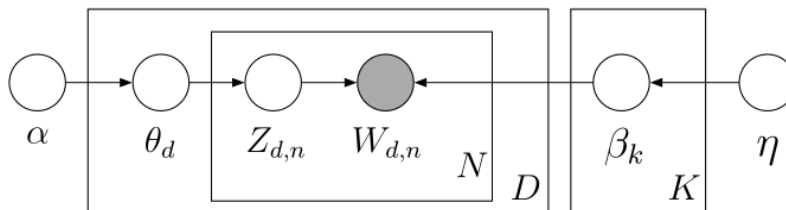
Themenmodelle

- Zuordnung Themen \leftrightarrow Wörter eines Dokuments
- Organisation von Dokumentensammlung anhand enthaltener Themen
- Probabilistisches Modell: LDA



Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) r-scan statistics that can be applied to the analysis of spacings of sequence markers.



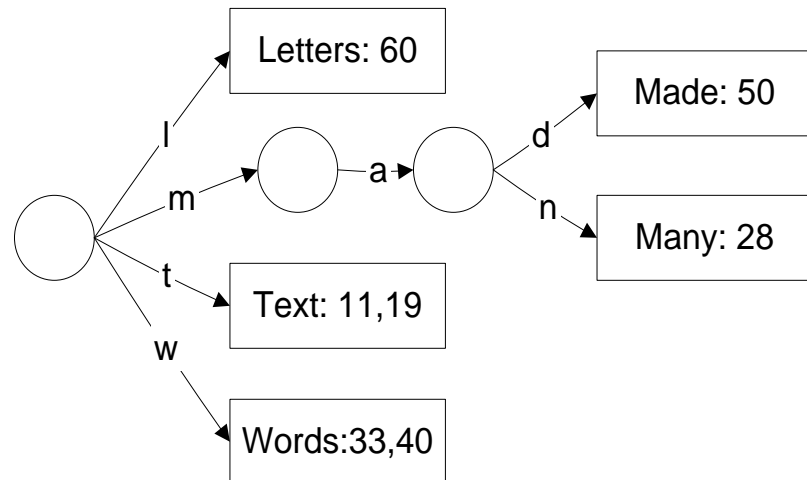
Indexstrukturen

- 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



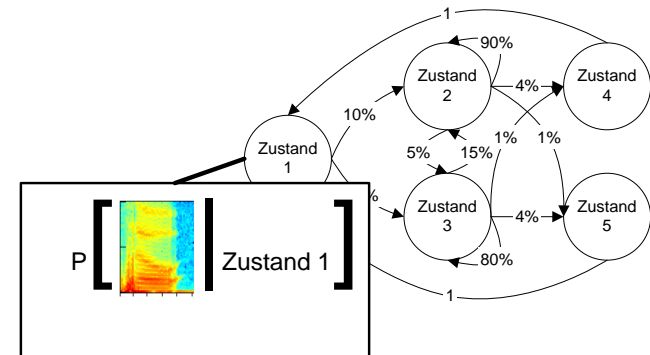
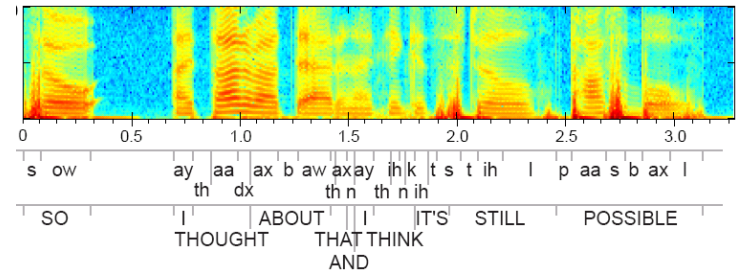
Spracherkennung und -erzeugung

- Spracherkennung:
Akustisches Modell +
Sprachmodell

$$\arg \max_{(w_1, \dots, w_T)} P(w_1, \dots, w_T | \text{Signal})$$

$$= \arg \max_{(w_1, \dots, w_T)} \underbrace{P(\text{Signal} | w_1, \dots, w_T)}_{\text{Akustisches Modell}} \underbrace{P(w_1, \dots, w_T)}_{\text{Sprachmodell}}$$

- Textanalyse und
Signalerzeugung
- Anwendung: Sprachportale



Websuche

- Crawling: Welche URL wann besuchen?
 - ◆ Endlos-URLs, dynamische Seiteninhalte.
 - ◆ Aktualisierungshäufigkeiten und Zeitpunkte.
 - ◆ Identische Seiten.
 - ◆ Link-Spam.
- Relevanz-Ranking: Analyse der Linkstruktur.

Fragen?

- (diese Woche noch keine Übungsaufgaben, nächste Woche noch keine Übung)
- Erster Übungstermin: 27.04.