Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Sprachtechnologie

Tobias Scheffer
Paul Prasse
Michael Großhans

Organisation

- Vorlesung/Übung, praktische Informatik.
- 4 SWS.
- 6 Leistungspunkte
- Übung:
  - Mo 14:00-15:30
- Vorlesung :
  - Mo 12:00-13:30
- 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.
- Übungen:
  - Eine Woche im voraus 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: 28.04.
- 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

Ereignis

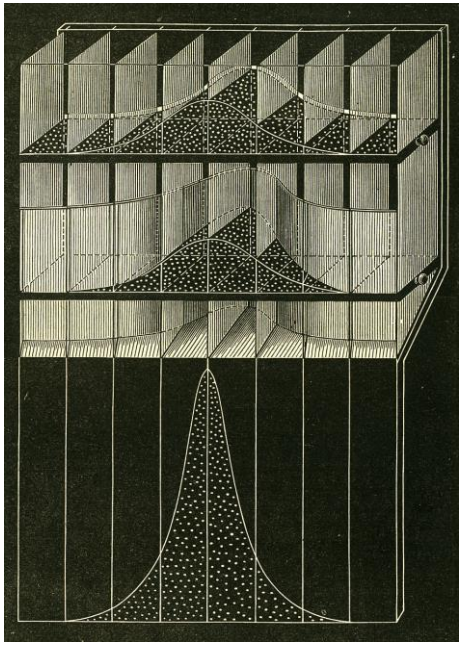
Wahrscheinlichkeit, dass der i-te Patient gesund bleibt, ...

$$P(X_i = \text{gesund} \mid \theta)$$

... gegeben, dass die Wirksamkeitsrate gleich θ ist

Hypothese

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



$$\begin{aligned} P(X_{neu}|X_1,\ldots,X_n) &= \int_\theta P(X_{neu}|\theta, X_1,\ldots,X_n)P(\theta|X_1,\ldots,X_n)d\theta \\ &= \int_\theta P(X_{neu}|\theta)\ P(\theta|X_1,\ldots,X_n)\ d\theta \end{aligned}$$

Statistische Sprachmodelle

- Elementares Werkzeug für
  - Spracherkennung,
  - Rechtschreibkorrektur,
  - Auto-Complete, Übersetzung, …

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

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

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

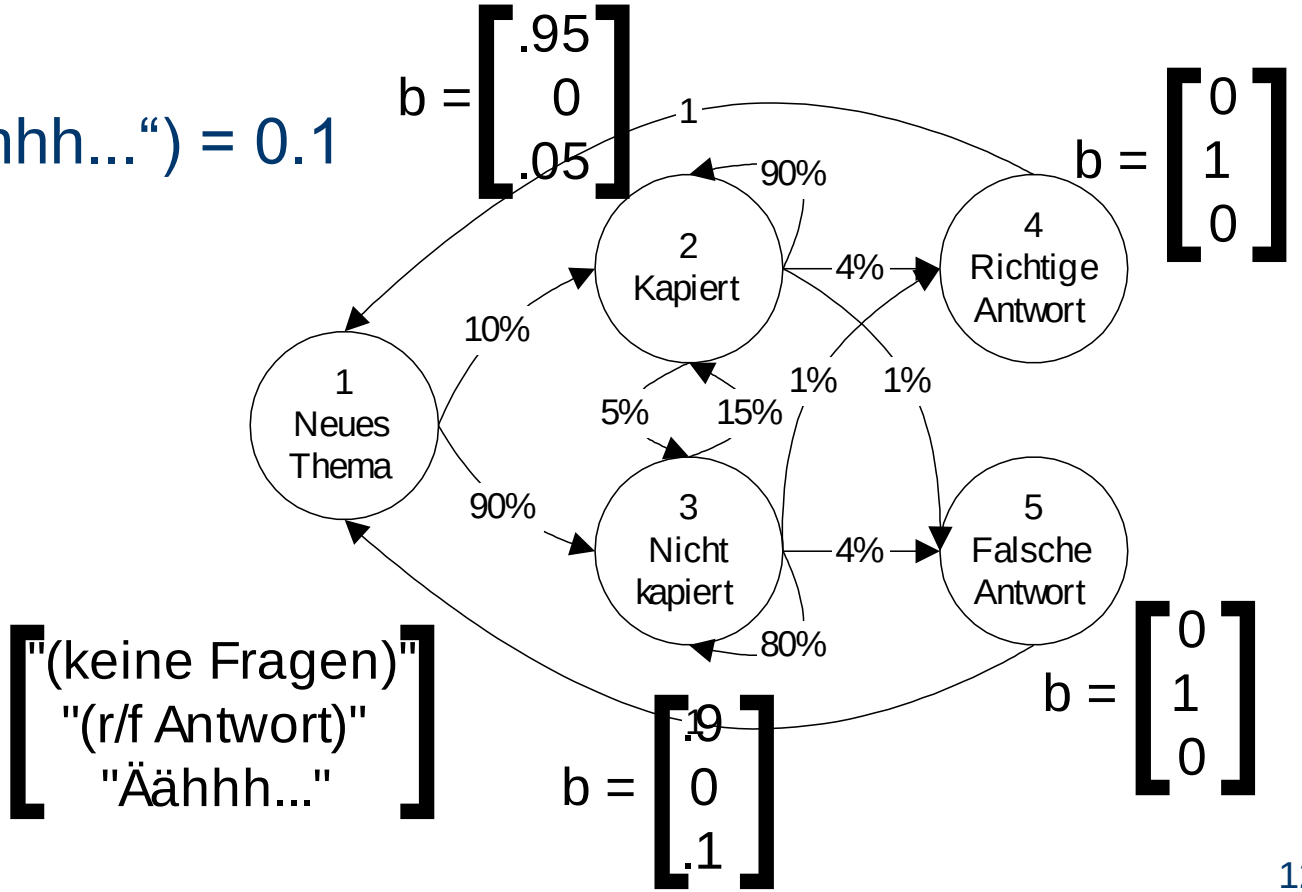
- $X_1, \ldots, X_n$: Zufallsvariablen.
- Allgemein gilt: $P(X_1,\ldots,X_n) = P(X_1)\prod_{i=2}^{n} P(X_i \mid X_{i-1},\ldots,X_1)$
- Zufallsvariablen bilden eine Markovkette, gdw:
  $P(X_1,\ldots,X_n) = P(X_1)\prod_{i=2}^{n} P(X_i \mid X_{i-1})$
- Jede Variable X_i nur von Vorgänger X_{i-1} abhängig.

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

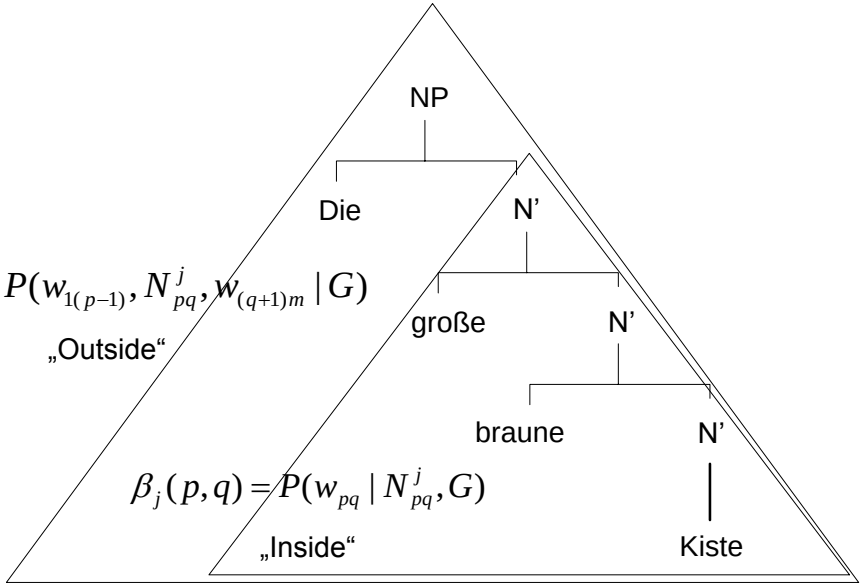
- (Andrei Markov, 1856-1922)

Hidden-Markov-Modell

- Akustisches Modell für Spracherkennung.
- Zustände emittieren Beobachtungen O_t (mit Wahrscheinlichkeit $b_i(O_t)$).
- b_3(„Äähhh...“) = 0.1



Part-of-Speech Tagging, Parsing



$$\alpha_j(p,q) = P(w_{1(p-1)}, N^j_{pq}, w_{(q+1)m} \mid G)$$

„Outside“

$$\beta_j(p,q) = P(w_{pq} \mid N^j_{pq}, G)$$

„Inside“

Named Entity Recognition

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?

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?

Google translate

From: Chines... 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？

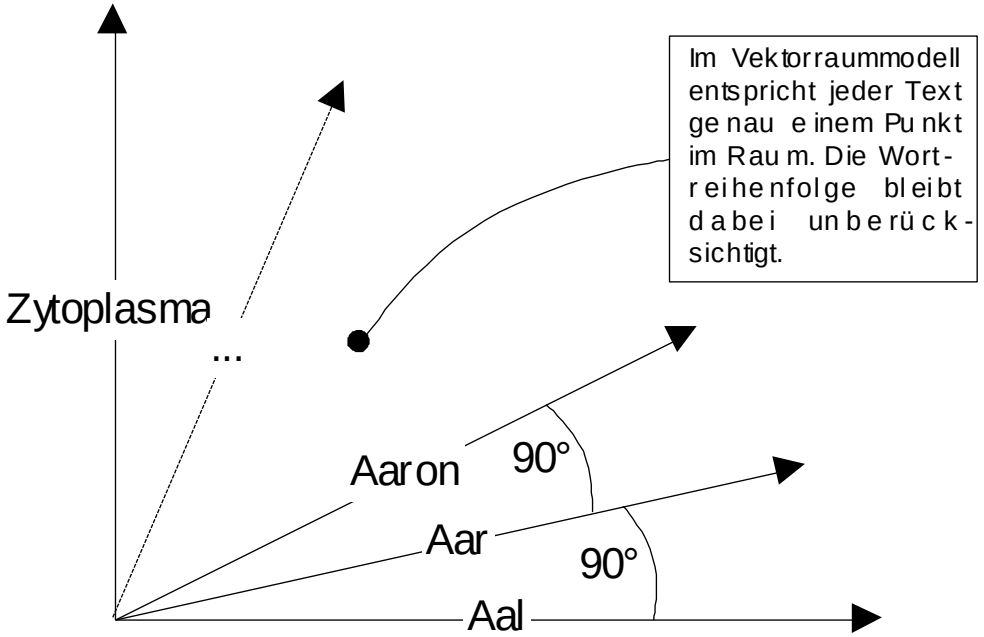
Listen Read phonetically

Chinese to German translation

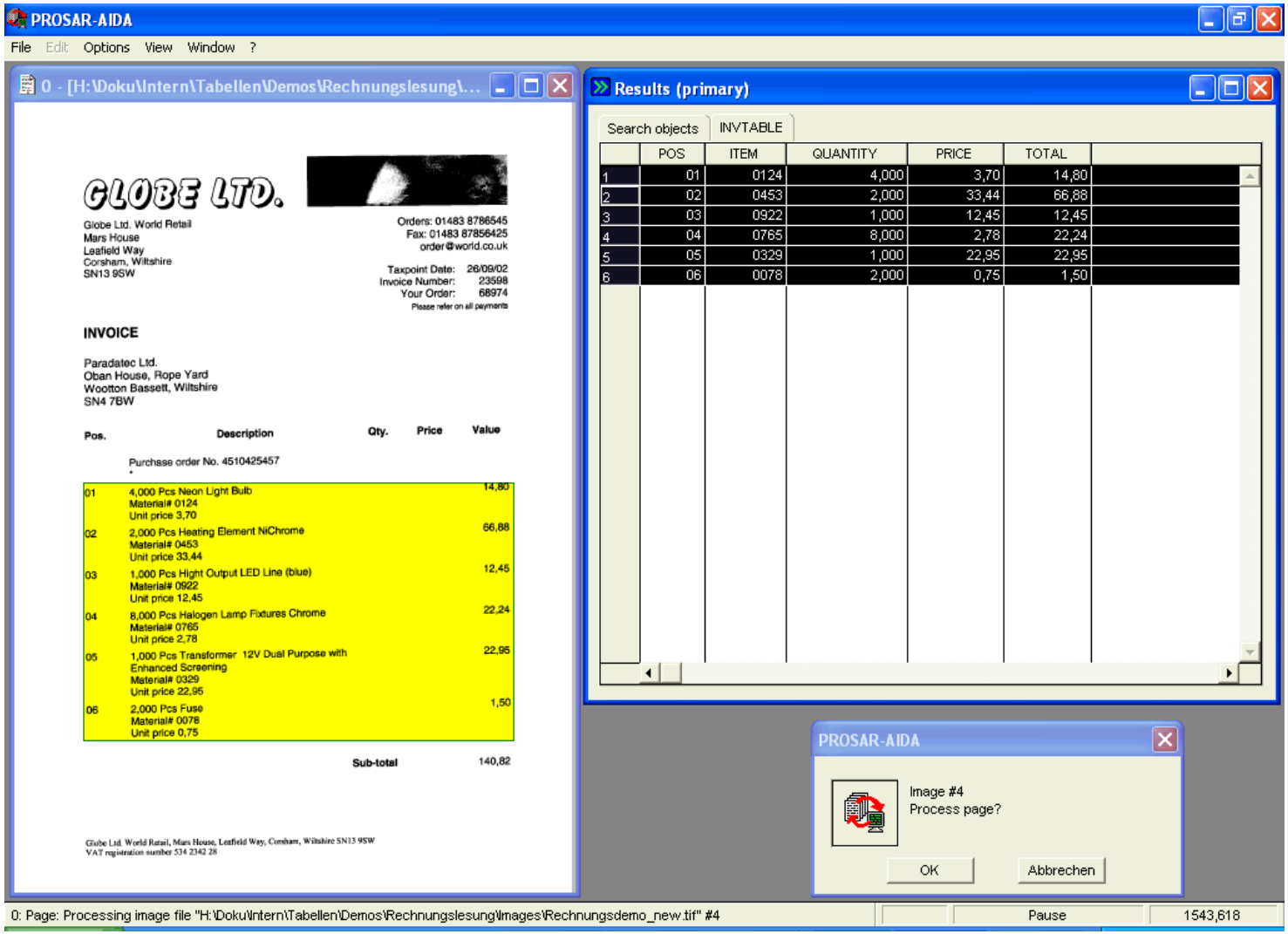
In der Wahrscheinlichkeitstheorie und Anwendung der Bayes-Theorem zeigt, wie die inverse Wahrscheinlichkeit bestimmen: Was ist die bedingte Wahrscheinlichkeit, dass angesichts B, was die bedingte Wahrscheinlichkeit von B gegeben ist?

Vektorraummodell

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

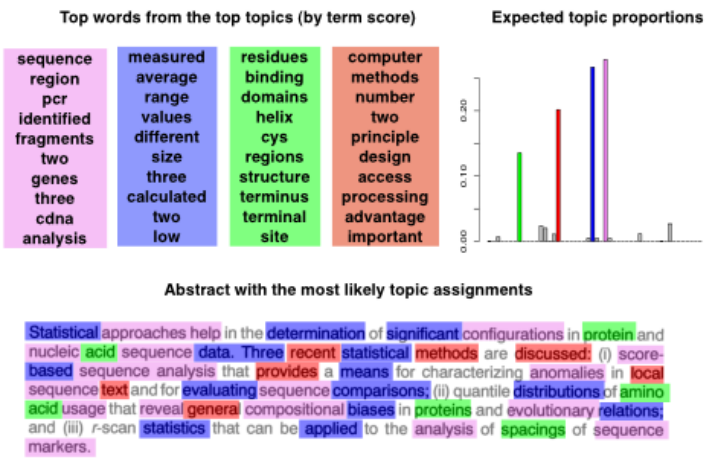


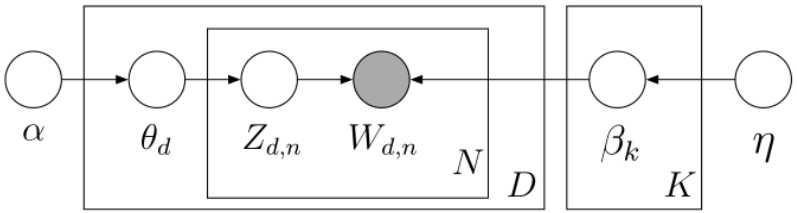
Textklassifikation, Informationsextraktion



Themenmodelle

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





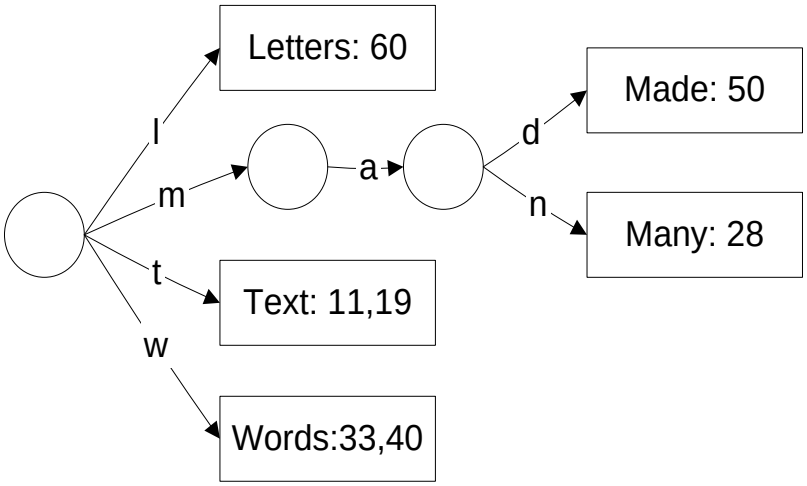
Indexstrukturen

- 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.

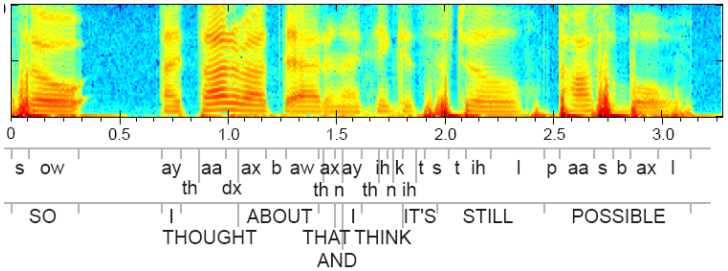
Terme	Vorkommen
Letters	60
Made	50
Many	28
Text	11, 19
words	33, 40



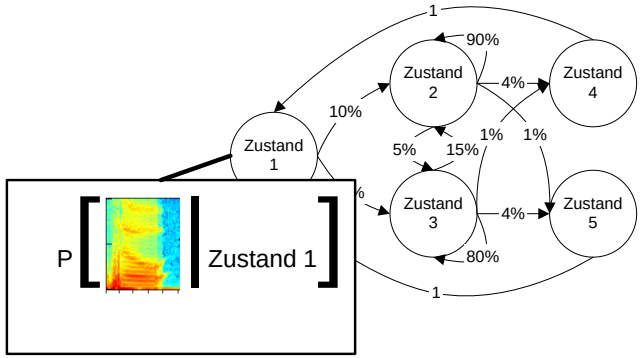
Spracherkennung und -erzeugung

- Spracherkennung: Akustisches Modell + Sprachmodell

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



- 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: 28.04.