Dokumenten- und Topicmodelle

UNIVERSITÄT LEIPZIG
Institut für Informatik

ASV
Automatische Sprachverarbeitung

Aufbau von Textkorpora

- **Korpus C enthält Menge von D Dokumenten**
- **jedes Dokument enthält Menge von N_i Wörtern**
- **gesamter Korpus enthält Vokabular von V voneinander verschiedenen Wörtern**
- **Länge des Korpus ist N**

Ziel

- **Welche Informationen können extrahiert werden?**
  - Clustering von Wörtern anhand Bedeutung(Semantik)
  - Identifizierung von Synonymien und Polysemien
  - Themenzuordnung von Dokumenten
  - …?
- **Weiteres Ziel aller Modelle: Dimensionsreduktion**
  - Einfach ist nur auszählen von Wörtern
  - Dimension hierbei: V (aber: V sehr groß, oft >> 1000000)
  - Wie reduzierbar?

Wiederholung tf-idf

- **Zähle Auftreten von Termen in Dokument**
- **Vergleiche Auftreten von Term in Dokument mit inversem Auftreten von Term in anderen Dokumenten**
- **Ergebnis: Term-Dokumentmatrix mit tf-idf Werten der Terme im Vokabular**
- **Reduktion von Dokumenten(unbestimmte Größe) auf Liste von Werten fixer Länge**

Bag-of-word assumption

- **Reihenfolge der Wörter wird nicht berücksichtigt**
- **Ein Dokument entspricht einem „Sack“ voller Wörter**
- **Für jedes Wort wird Frequenz gespeichert**
- **Annahme:**
  - Information über Art und Anzahl von Wörtern reichen aus um Rückschlüsse auf die Struktur von Text zu ziehen
  - Grundlage: de Finettis Theorem
    - Annahme der Austauschbarkeit:
    - Austauschbare Zufallsvariablen folgen einer vermischten Verteilung (mixture distribution), meist unendlich

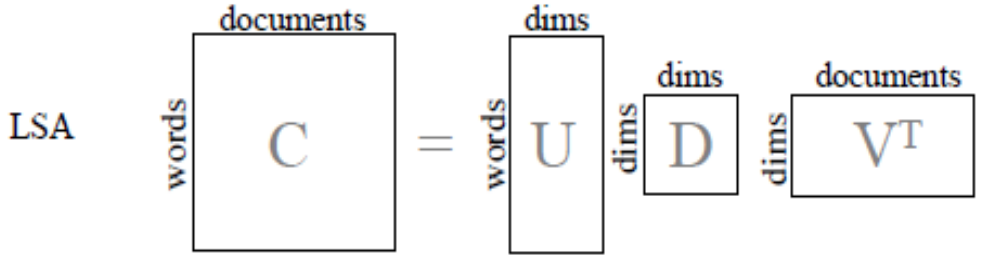
Dokumentenmodelle

UNIVERSITÄT LEIPZIG
Institut für Informatik

ASV
Automatische Sprachverarbeitung

LSA - Latent Semantic Analysis

- **Form der linearen Faktorisierung**
- **Grundlage bildet eine Wort-Dokument Kookkurentenmatrix**
- **diese wird per Singulärwertzerlegung in drei Matrizen zerlegt**
- **alle bis auf n höchsten Singulärwerte werden auf 0 gesetzt**
- **ursprüngliche Matrix wird rekonstruiert (hat nun geringeren Rang)**

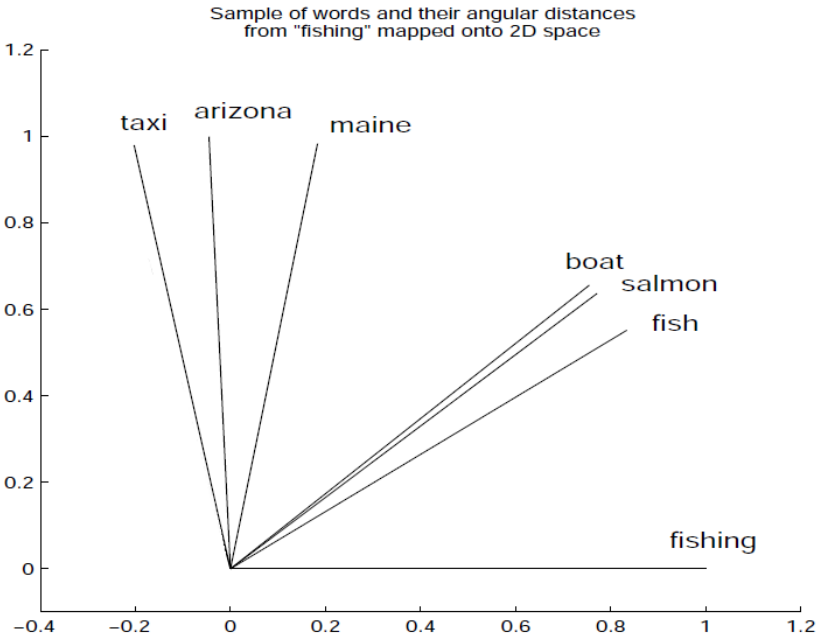


LSA

- **Vorteil:**
  - Keine Eins/Null Entscheidungen mehr
  - Dimensionsreduktion auf n Dimensionen („semantische Kategorien“)
- **Nachteil:**
  - Schlechtes zugrunde liegendes statistisches Modell → schlechte Begründung
    - Gemeinsame Verteilung von Wörtern und Dokumenten folgt nicht Gauss- sondern Poissonverteilung
  - Kein Vorwissen über n
  - Polysemie
    - Jedes Wort wird genau einer semantischen Bedeutung zugeordnet (gleicher Datenpunkt im semantischen Raum)
    - d.h. Ergebnis ist Durchschnitt aller verschiedenen Bedeutungen eines Wortes (als Vektor)

LSA

- **geometrische Interpretation**
  - Reduzierte Dimensionen spannen „semantischen Raum“
  - in Wortmatrix U
    - Winkel zwischen Wortvektoren (Cosinus-Maß) entspricht ihrer semantischen Ähnlichkeit
    - Möglichkeit semantisches Clustern
  - ähnlich für Dokumentmatrix V
    - Clustering von ähnlichen Dokumenten



LSA – Beispiel

Example of text data: Titles of Some Technical Memos

c1: *Human* machine *interface* for ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user* perceived *response time* to error measurement

m1: The generation of random, binary, ordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

LSA – Beispiel, Termfrequenzmatrix

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

LSA – Beispiel, SVD

0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	-0.68	0.23
0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18

3.34								
	2.54							
		2.35						
			1.64					
				1.50				
					1.31			
						0.85		
							0.56	
								0.36

0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53
0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25	0.08
-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01	-0.03
0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15	-0.60
-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00	0.36
0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.1	0.25	0.04
-0.01	0.05	0.01	-0.02	-0.06	0.45	-0.76	0.45	-0.07
-0.06	0.24	0.02	-0.08	-0.26	-0.62	0.02	0.52	-0.45

LSA – Beispiel, rekonstruierte Matrix mit geringerem Rang

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

Einschub: Latente Variablen

- **Theoretische Konstrukte, abhängig vom Modell**
- **Sind nicht direkt messbar**
- **Können von messbaren Variablen (Observablen) ausgehend bestimmt werden**

Automatische Sprachverarbeitung

pLSI - probabilistic Latent Semantic Indexing

- **Stammt nicht aus der linearen Algebra (wie LSA)**
- **Geht von vermischten Verteilungen und einem Modell der latenten Klassen aus**
- **Basiert auf Aspect-Model**
  - Ordnet jeder Beobachtung (Term) eine latente Variable (Klasse) zu
  - Gemeinsame Wkt' über Dokumente und Terme wird definiert:

  - Annahme: d und w sind statistisch unabhängig voneinander

Automatische Sprachverarbeitung

pLSI

- **Ähnlichkeit zu LSA**
  - Definiere 3 Matrizen:
  - Gemeinsames Wahrscheinlichkeitsmodell P gegeben durch
- **Beobachtung**
  - Äußere Produkte zwischen Zeilen von U und V zeigen bedingte Unabhängigkeit
  - K Faktoren entsprechen Mischkomponenten aus Aspect-Model
  - Mischanteile ersetzen Singulärwerte

pLSI

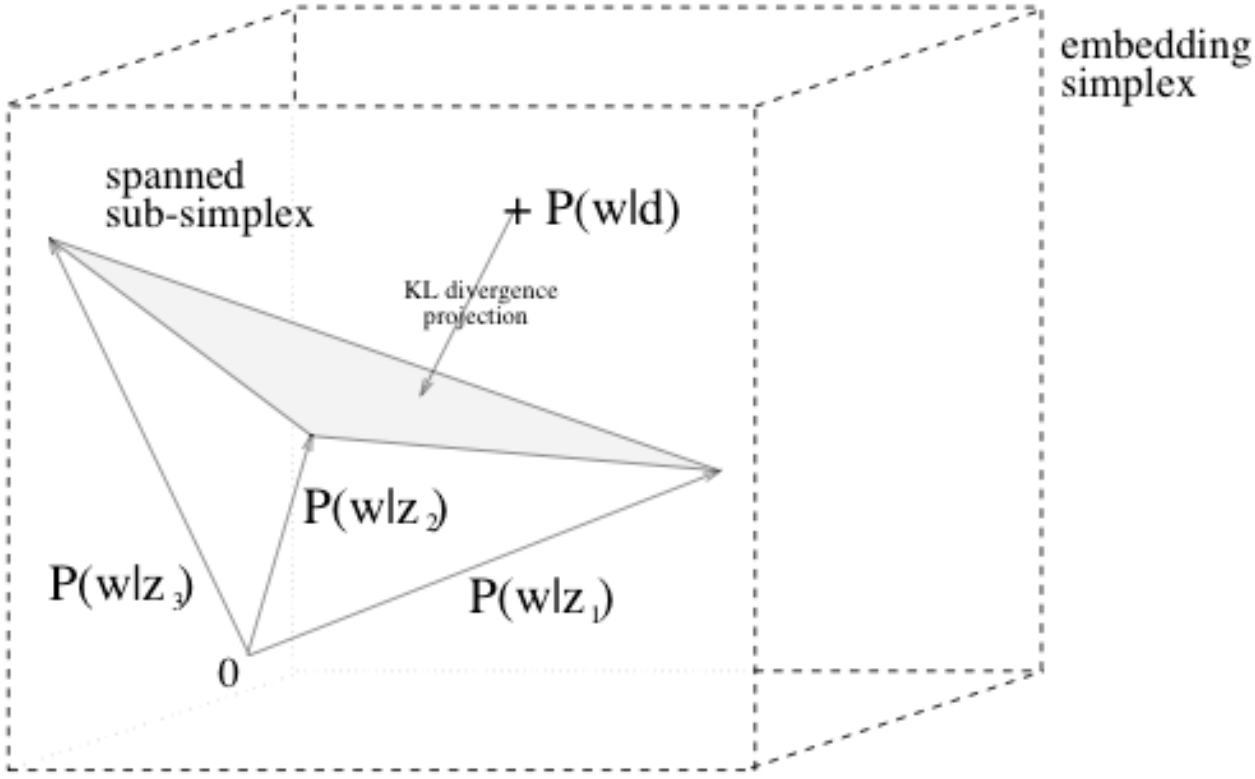
- **Unterschied zu LSA**
  - Funktion zum Bestimmen der optimalen Annäherung bei LSA: L_2- oder Frobeniusnorm
  - entspricht der Annahme eines Gaussrauschen auf Termanzahlen
  - pLSI nutzt Likelihood-Funktion zur expliziten Maximierung der Vorhersagequalität des Modells
    - entspricht Minimierung der Kullback-Leibler Distanz zwischen tatsächlicher und approximierter Wahrscheinlichkeitsverteilung
- **Einschub Kullback-Leibler Distanz**
  - Maß für die Verschiedenheit zweier Wahrscheinlichkeitsverteilungen
  - Basiert auf Informationstheorie
    - Misst angenommene Anzahl extra bits um eine Information in einem auf Q basierenden Code zu kodieren statt in einem P basierenden
    - Definition:

pLSI

- **Würdigung**
  - Approximation in P ist für jedes Wort eine wohldefinierte Wahrscheinlichkeitsverteilung
  - Faktoren haben klare probabilistische Bedeutung
  - LSA arbeitet nicht mit Wahrscheinlichkeiten, sogar negative Werte möglich
  - Keine offensichtliche Interpretation der Richtung im semantischen Raum von LSA, in pLSA ist Richtung interpretierbar als multinomiale Wortverteilung
  - da probabilistisches Modell: Möglichkeiten der Modellselektion, Herausfinden von optimalen K(Anzahl der latenten Klassen)

pLSI – geometrische Deutung

- **K Klassenspezifische Multinomialverteilungen werden im M-1 dimensionalen Simplex über alle möglichen Mult. dargestellt**
- **Bilden K-1 dim. Sub-simplex**
- **P(w|d) geg. durch**
- **konvexkomb. P(w|z)**



Topicmodelle

UNIVERSITÄT LEIPZIG
Institut für Informatik

ASV
Automatische Sprachverarbeitung

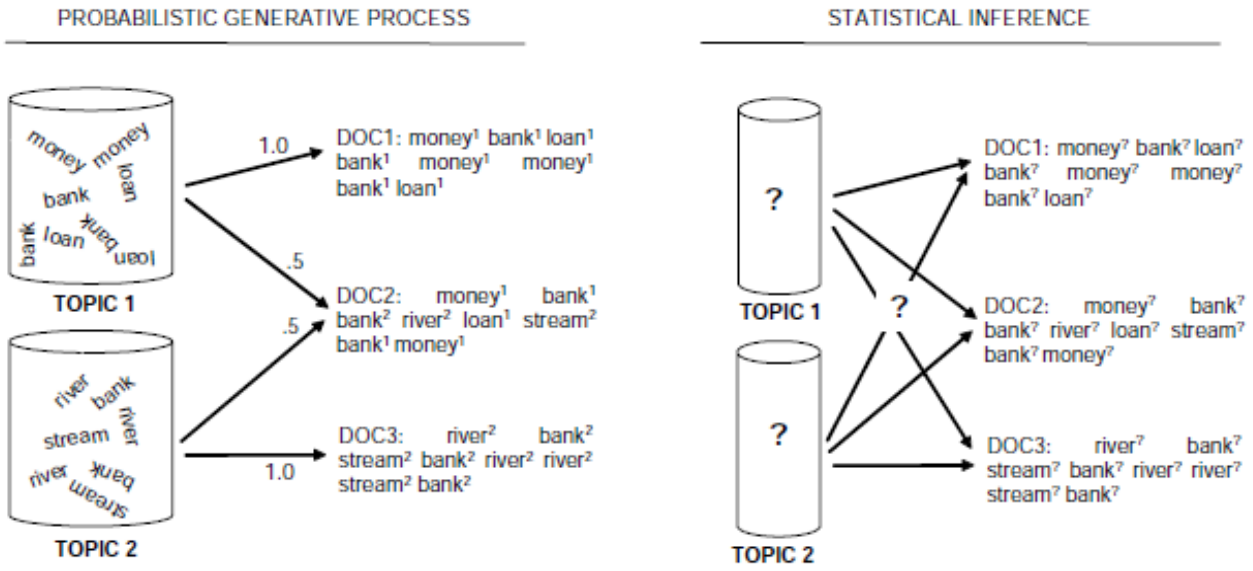
Topicmodelle

- **Basieren auf der Idee, Dokumente seien Gemisch(Mixture) von Topics und Topics ein Gemisch von Wörtern**
- **definieren generatives Modell**
- **Generativer Prozess wird zur Inferenz umgedreht**
- **Vorteil gegenüber räumlicher Repräsentation:**
  - Jedes Topic einzeln interpretierbar
  - Bietet Wahrscheinlichkeitsverteilung über Wörter, damit kohärente Cluster von semantisch ähnlichen Wörtern

Automatische Sprachverarbeitung

Generative Modelle

- **basiert auf einfachen Samplingregeln**
- **beschreibt, wie Dokumente auf Basis eines latenten Parameters generiert werden können**
- **Ziel beim Anpassen des Modells: „besten“ Satz von latenten Parametern finden, der gefundene Daten erklärt**

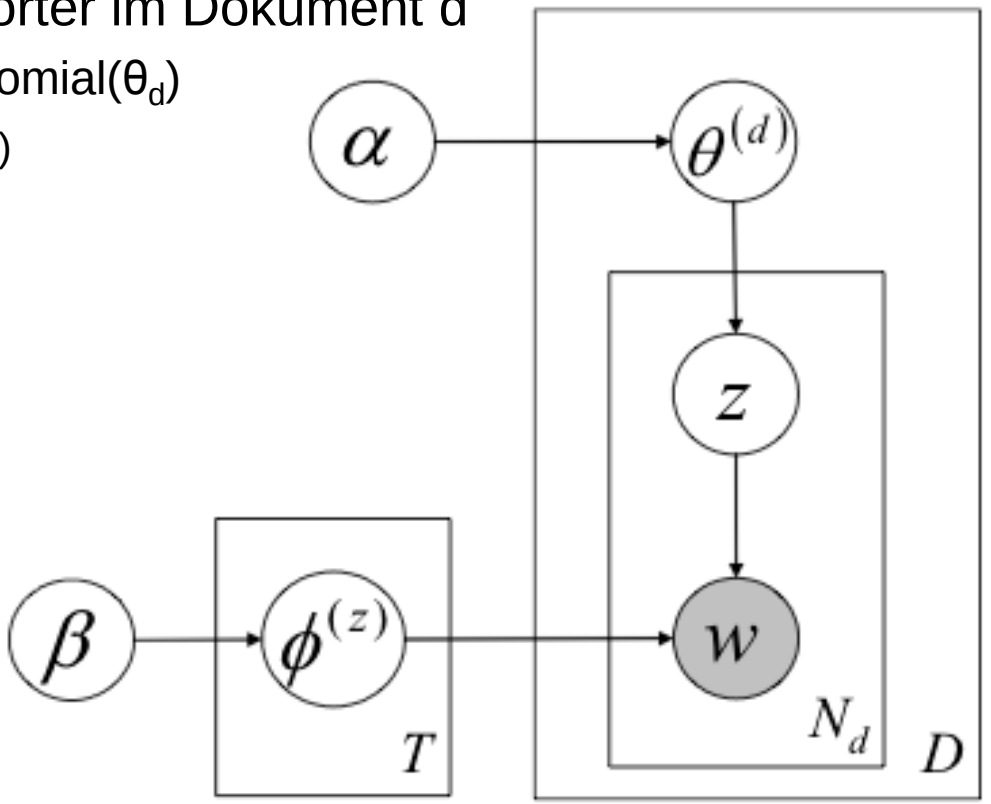


LDA – Latent Dirichlet Allocation

- **Ausgehend von Bag-of-words Ansatz und latenten Variablen**
  - Ein Dokument ist eine Mischung von Topics (latent)
  - Ein Topic ist eine Mischung aus Wörtern (obervable)
- **Notation**
  - P(z) ist eine Verteilung über Topics z in einem Dokument
  - P(w|z) ist eine Verteilung über Wörter w für ein Topic z
  - $P(z_i = j)$ ist die Wkt' das für i-tes Wort Topic j gezogen wird
  - $P(w_i|z_i = j)$ ist Wkt' von Wort w_i im Topic j
  - Es ergibt sich eine Verteilung über alle Wörter eines Dokuments

LDA – Generativer Prozess

- **Für jedes Dokument d:**
  - Wählen der Topicverteilungen θ ~ Dir(α)
  - Wählen der Wortverteilungen φ ~ Dir(β)
  - Für jedes Wort w_n der N_d Wörter im Dokument d
    - Wählen von topic z_n ~Multinomial(θ_d)
    - Wählen von w_n aus $P(w_n|\varphi_{zn})$

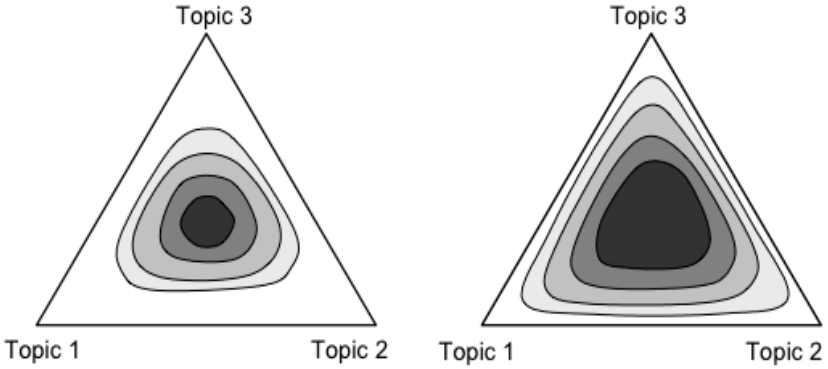


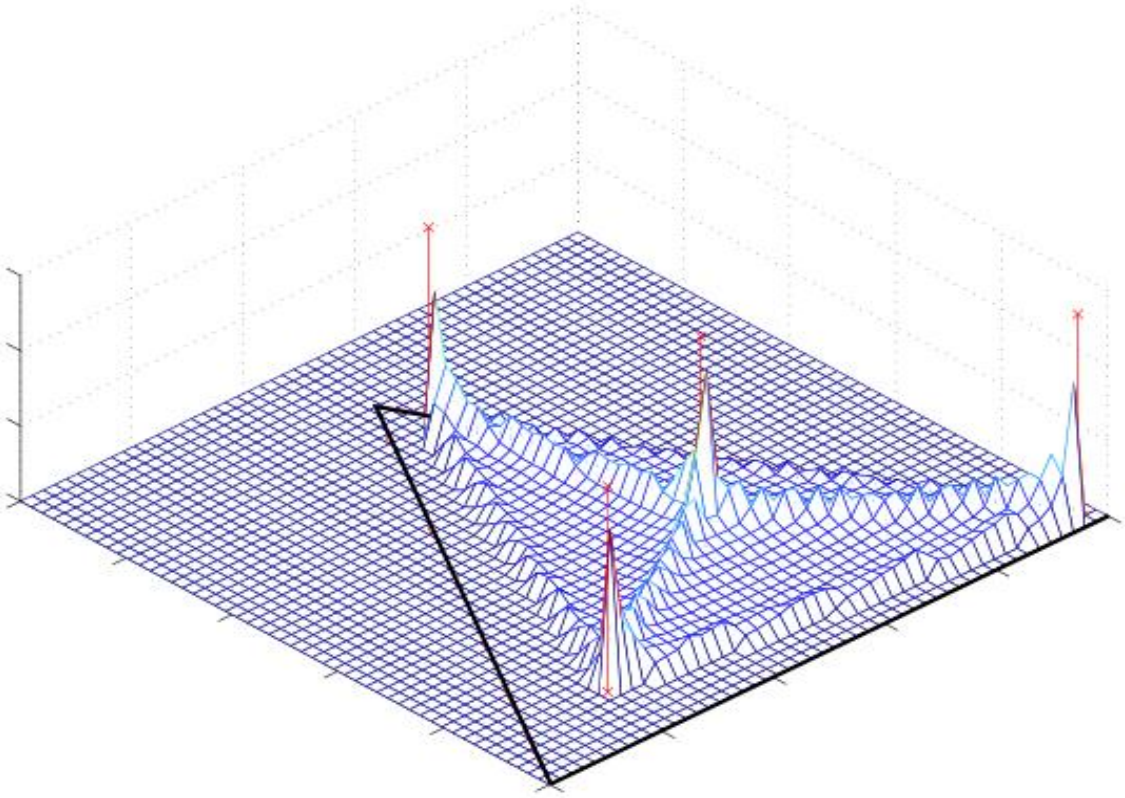
LDA

- **Notation**
  - Wir schreiben $\varphi^{(j)} = P(w|z = j)$ und $\theta^{(d)} = P(z)$
  - Diese beschreiben, welche Topics für ein Dokument bzw. welche Wörter für ein Topic wichtig (eigentlich: wahrscheinlich) sind
  - Beides multinomiale Verteilungen
- **Im Gegensatz zu pLSI: Annahme zur Generierung von θ**
  - A priori Verteilung ist Dirichletverteilung
  - Dirichlet ist sogenannte conjugate prior für die Multinomialverteilung
    - Mathematisch sinnvoll, bessere Berechenbarkeit

LDA

- **Da ohne Beobachtung, symmetrisches alpha**
- **Alpha < 1, für Verteilungen die nur wenige Topics bevorzugen (Sparsity – Dünnbesetztheit)**





LDA – statistische Inferenz

- **Iterativ mittels Markov Chain Monte Carlo Methode**
  - Abschätzen der posteriori Verteilung über Topiczuordnung z
- **Speziell Gibbs-Sampling**
  - für jedes Wort wird die Topiczuordnung berechnet, abhängig von allen anderen Zuordnungen
  - Hochdimensionale Verteilung wird durch wiederholtes Ziehen von niedrigdimensionalen Variablen simuliert
  - Von Verteilung über z ausgehend werden φ und θ approximiert
  - Nur zwei Matritzen benötigt

$$P\left(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot\right) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^{W} C_{wj}^{WT} + W\beta} \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_i t}^{DT} + T\alpha}$$

LDA – statistische Inferenz

- **Approximerung**

$$\phi'^{(j)}_i = \frac{C^{WT}_{ij} + \beta}{\sum_{k=1}^{W} C^{WT}_{kj} + W\beta} \qquad \theta'^{(d)}_j = \frac{C^{DT}_{dj} + \alpha}{\sum_{k=1}^{T} C^{DT}_{dk} + T\alpha}$$

- **Beispiel:**

$$\phi^{(1)}_{MONEY} = \phi^{(1)}_{LOAN} = \phi^{(1)}_{BANK} = 1/3$$
$$\phi^{(2)}_{RIVER} = \phi^{(2)}_{STREAM} = \phi^{(2)}_{BANK} = 1/3$$

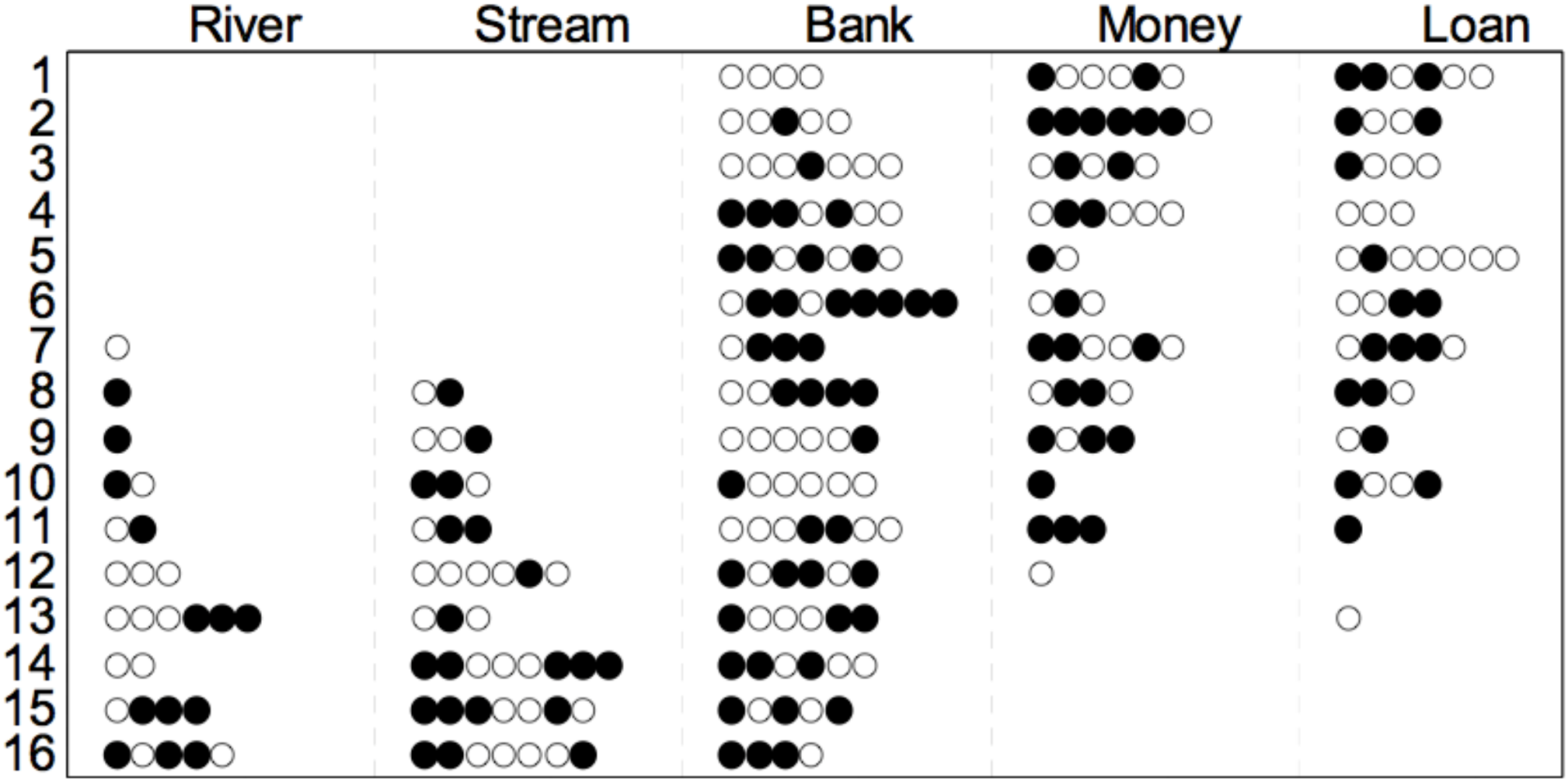
- **Ergebnis der Inferenz**

$$\phi'^{(1)}_{MONEY} = .32, \quad \phi'^{(1)}_{LOAN} = .29, \quad \phi'^{(1)}_{BANK} = .39$$
$$\phi'^{(2)}_{RIVER} = .25, \quad \phi'^{(2)}_{STREAM} = .4, \quad \phi'^{(2)}_{BANK} = .35$$

  – Nur 16 Dokumente, gute Werte dafür

LDA – Beispiel

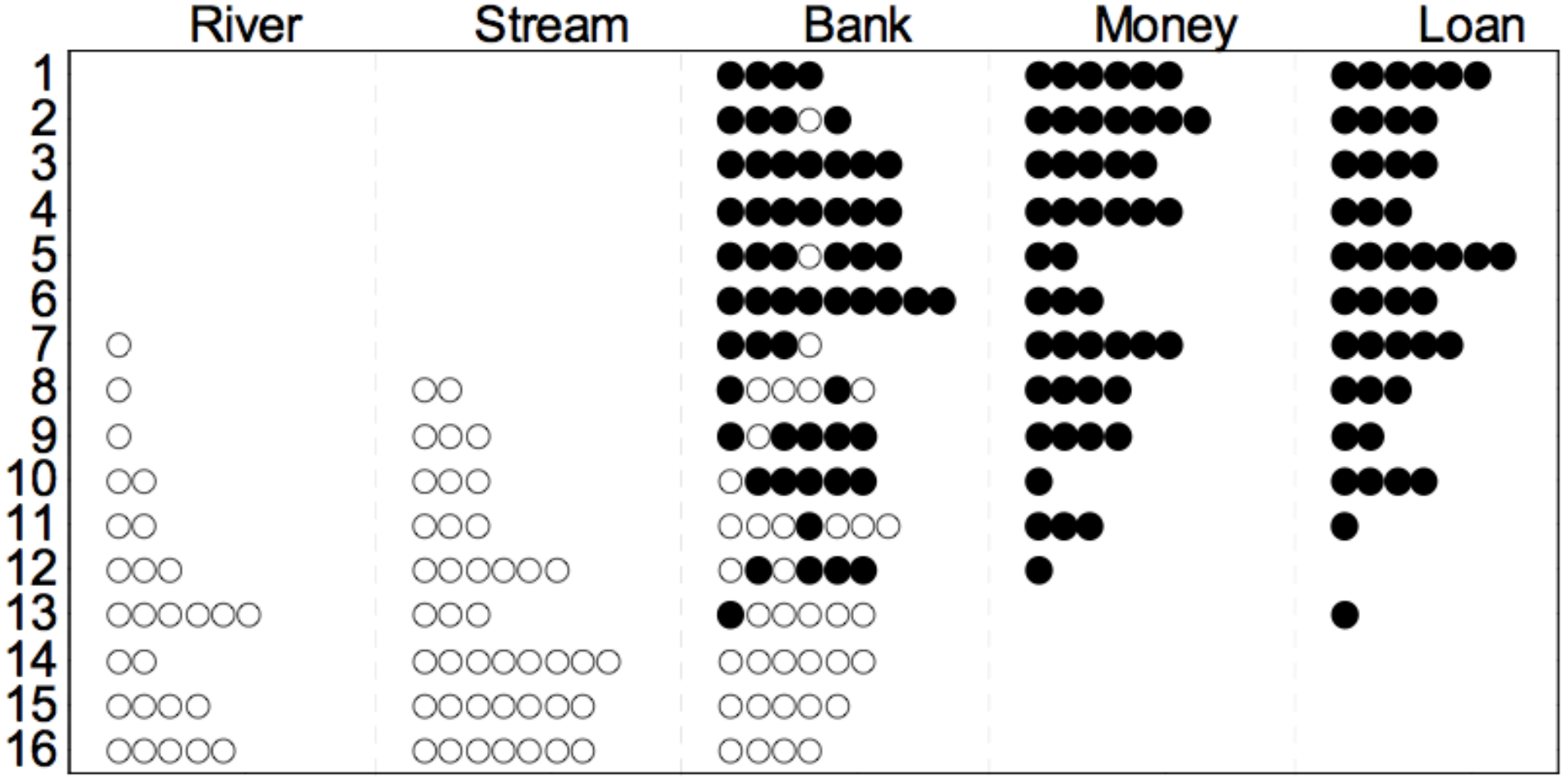
- **Generierte Dokumente**



ASV Automatische Sprachverarbeitung

LDA – Beispiel

- **Inferierte Topiczuordnung**



LDA – Anwendungen

- **Disambiguierung von Polysemien**

Topic 77

word	prob.
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
SINGING	.026
BAND	.026
PLAYED	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
PLAYING	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

Topic 82

word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

Topic 166

word	prob.
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
PLAYED	.031
BASEBALL	.027
GAMES	.025
BAT	.019
RUN	.019
THROW	.016
BALLS	.015
TENNIS	.011
HOME	.010
CATCH	.010
FIELD	.010

Automatische Sprachverarbeitung

LDA – Anwendungen

- **Dokumentenclustering**
- **Semantisches Clustering von Begriffen**
  - Auffinden von Synonymen
- **TDT - Topicdetection and Tracking**
- **Weitere?**

LDA – Probleme

- **Hauptprobleme sind das korrekte Festlegen von K (Dimension der Topics) sowie der Hyperparameter**
  - Hierarchischer Dirichletprozess
    - Dynamisches Abschätzen der Anzahl von Topics
    - In jeder Iteration kann neues Topic hinzukommen oder wegfallen
    - Stellt sicher, das immer aus derselben (unbekannten) Menge von Topics gezogen wird
  - Sampling der Hyperparameter in jeder (oder jeder n-ten Iteration des Gibbs-Samplers)
    - Hyperparameter werden an Daten angepasst

LDA – Weiterentwicklungen

- **Author-Topic model**
  - Zusätzlich wird die Metainformation Autor miteinbezogen
  - Inferieren einer Autor-spezifischen Topicverteilung
  - Möglichkeit Themenprofile von Autoren zu erstellen
- **Zusammenhang Autor-Themen-Profile und Abstand in sozialem Netzwerk**
  - Autorennetzwerk durch Hyperlinkstruktur gegeben
  - Geringerer Abstand geht einher mit größere Ähnlichkeit der Themenprofile

Zusammenfassung

UNIVERSITÄT LEIPZIG
Institut für Informatik

ASV
Automatische Sprachverarbeitung

Zusammenfassung

- **LSA /LSI**
  - „einfachstes“ Modell
  - Lineare Algebra, kein Bezug zu linguistischen Erkenntnissen
- **pLSA/pLSI**
  - Weiterentwicklung von LSA
  - Probabilistisches Modell
- **LDA**
  - Volles generatives Modell
  - Beste Vorhersageeigenschaften von vorgestellten Modellen

Quellen

- **Blei, Ng, Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, 993-1022, 2003**
- **Griffiths, Steyvers: Finding Scientific Topics, Proceedings of the National Academy of Science, 101, 5228-5235, 2004**
- **Griffiths, Steyvers: Probabilistic Topic Models, In: Landauer et. al.: Latent Semantic Analysis: A Road to Meaning, Laurence Erlbaum, 2005**
- **Hofmann: Probabilistic Latent Semantic Indexing, Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, 1999**
- **Landauer, Foltz, Laham: Introduction to Latent Semantic Analysis, Discource Processes, 25, 259-284, 1998**