Latent Dirichlet Allocation

Vertiefung der Grundlagen
der Computerlinguistik

Michael Anzer, Viktoria Heck

Gliederung

- Einleitung: was ist LDA?
- Ansatz: wie funktioniert LDA?
- Mathematischer Aspekt: Formeln und Variablen
- Unterschied zwischen LDA und anderen Modellen
- Nutzung von LDA
- Beispiel
- Zusammenfassung

Einleitung: Problem

- Text Korpora: wie kann man diese modellieren?
- Ziel: das Finden kurzer Beschreibungen der Mitglieder eines Korpus (unüberwacht)
  - Dabei sollten die statistischen Eigenschaften beibehalten werden
    - Erlaubt Zusammenfassen, Klassifizieren, Entscheidungen über Relevanz etc.

Einleitung: Probabilistic Latent Semantic Indexing

- pLSI (Hoffmann, 1999)
  - Jedes Wort eines Dokuments als Beispiel eines gemischten Models
    - Komponenten dieses Models multinomialverteilte Zufallsvariablen
    - Diese Zufallsvariablen sind Repräsentationen von Topics
  - Generierung jedes Wortes aus einem einzigen Topic
    - Verschiedene Wörter in einem Dokument können aus verschiedenen Topics generiert werden
  - Repräsentation eines Dokuments durch Liste von Nummern (die Mischungsanteile der Topics)
    - Diese Wahrscheinlichkeitsverteilung ist die reduzierte Beschreibung des Dokuments

Einleitung: pLSI Nachteile

- Nachteil pLSI: probabilistisiches Model nur auf Wortebene, nicht auf Dokumentenebene
- Kein probabilistisiches Model für die Zahlen Outputliste
  - Desto größer der Korpus, desto mehr Parameter
    - Overfitting
  - Zuweisung von Wahrscheinlichkeiten von Dokumenten außerhalb des Training Set unklar

Einleitung: Latent Dirichlet Allocation

- Annahme bei allen Modellen: bag-of-words
  - Wörter untereinander frei austauschbar (dasselbe für Dokumente)
- Repräsentationstheorem (Finetti, 1990): jede Kollektion von austauschbaren Zufallsvariablen hat eine Repräsentation als Mischverteilung
  - → austauschbare Repräsentationen von Dokumenten und Wörtern erfordern gemischte Modelle, die die Austauschbarkeit von Dokumenten und Wörtern erfassen

Wie funktioniert LDA?

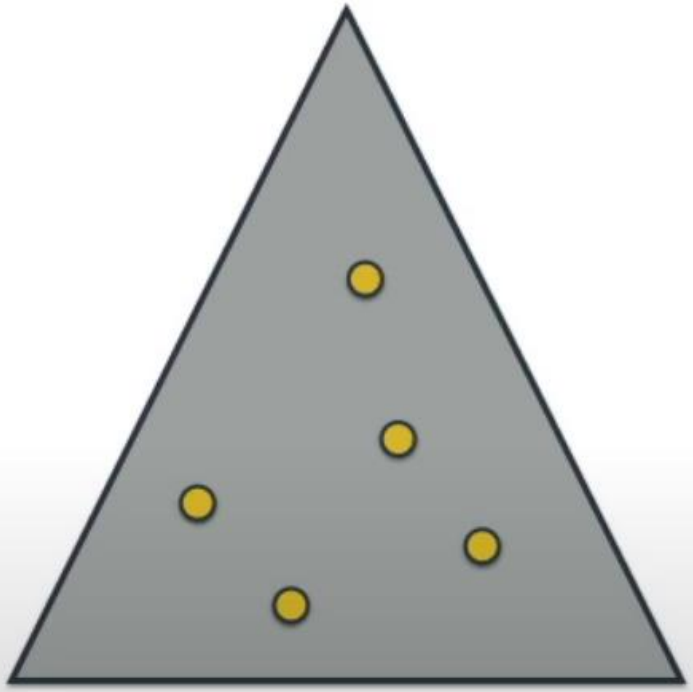
- Gegeben: Korpus mit Textdokumenten
- Nicht gegeben: Topics der Dokumente („latent“)
- Aufgabe LDA: Dokumenten Topics zuweisen
  - Geometrische Herangehensweise
- Aber: Topics selbst sind nicht definiert (Topic „a“ anstatt von Sport, Politik etc.)

Wie funktioniert LDA?

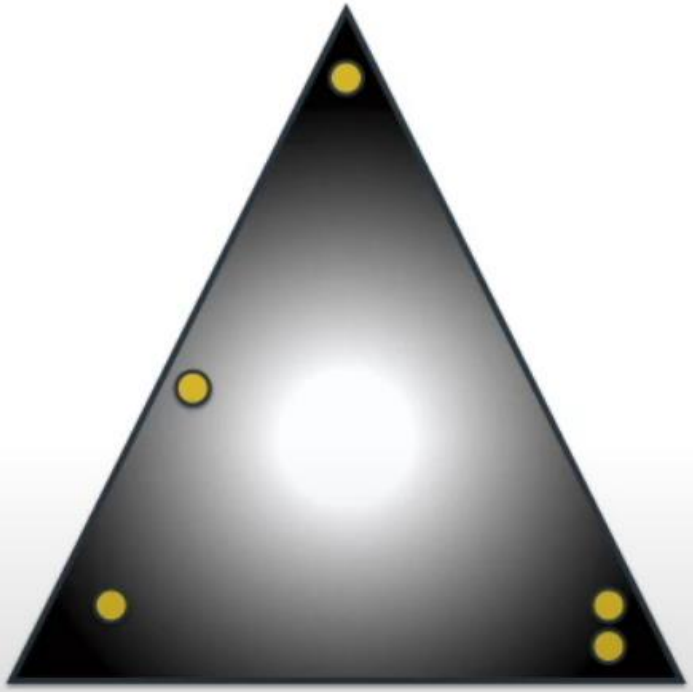
- Prinzip: Input = Dokument
- Output auch Dokument, von Interesse ist die Ähnlichkeit des Input- und Outputdokuments
- Der Inhalt des Outputdokuments bestimmt durch die Einstellungen des LDA
  - Ziel: Finden der besten Einstellungen (= höchste Übereinstimmung mit dem Originaldokument)

Wie funktioniert LDA?

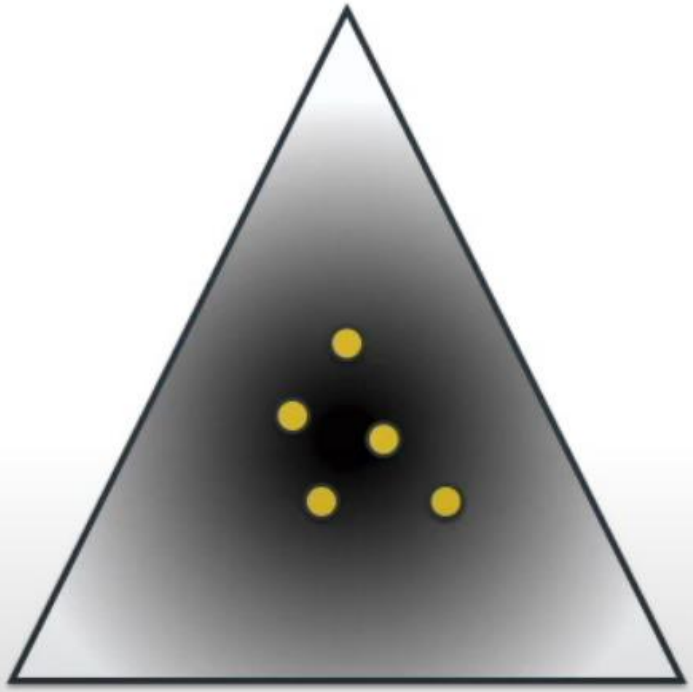
- Dirichlet Verteilung



$\alpha = 1$



$\alpha < 1$

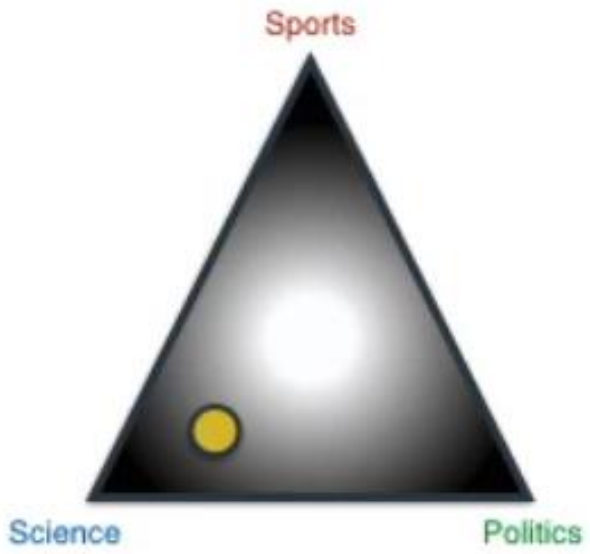


$\alpha > 1$

Wie funktioniert LDA?

- Schritt 1: Zufälliger Punkt in einer Dirichlet Verteilung wird als Dokument festgelegt
  - „Ecken“ der Verteilung repräsentieren mögliche Topics → Verteilung ist „schwerer“ an den Ecken
  - Ergebnis = Wahrscheinlichkeit, dass das Dokument die gegebenen Topics hat (Anzahl der Topics = Hyperparameter)

Wie funktioniert LDA?

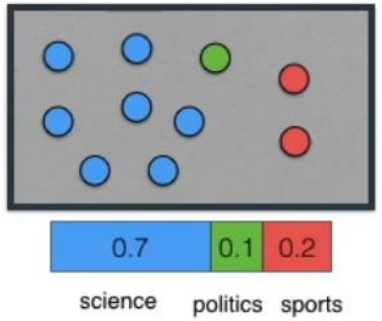


science	politics	sports
0.7	0.1	0.2

Wie funktioniert LDA?

- Schritt 2: Auf Ergebnis von Schritt 1 basierend: erstellen einer Multinomialverteilung
  - In dieser Multinomialverteilung: Wahrscheinlichkeiten des vorherigen Ergebnisses = Wahrscheinlichkeit des jeweiligen Topics, einem Wort zugewiesen zu werden
  - Ergebnis: die Topics für die Wörter des Outputdokuments

Wie funktioniert LDA?



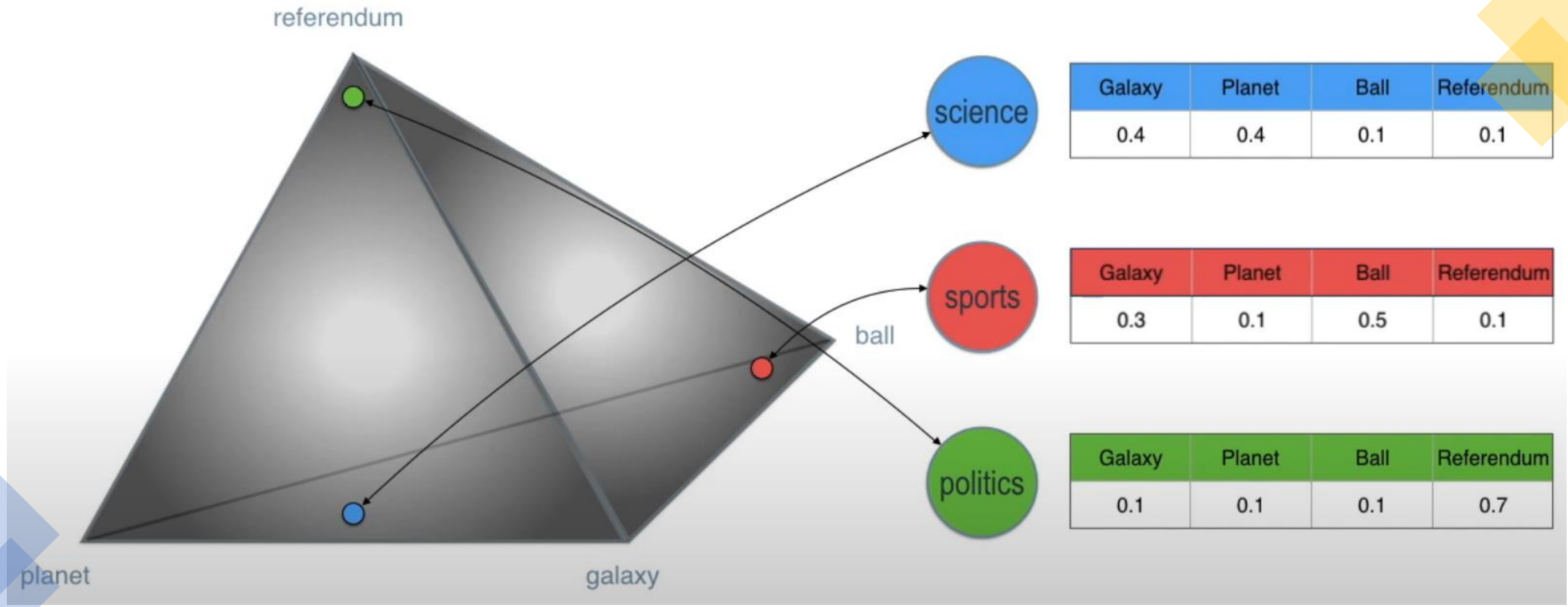
Topics

science
science
sports
science
science
politics
sports
sports
science

Wie funktioniert LDA?

- Schritt 3: Dirichlet Verteilung, welche Topics zu Wörtern zuweist
  - Ecken der Verteilung sind Wörter eines Dokuments
    - Geometrische Form ein n-dimensionaler Simplex
  - Auch hier: „Ecken“ der Verteilung repräsentieren mögliche Wörter → Verteilung ist „schwerer“ an den Ecken
  - Pro Topic wird ein zufälliger Punkt in der Verteilung festgelegt
  - Ergebnis = Wahrscheinlichkeit, dass das Topic den gegebenen Wörtern zugewiesen wird

Wie funktioniert LDA?



Galaxy	Planet	Ball	Referendum
0.4	0.4	0.1	0.1

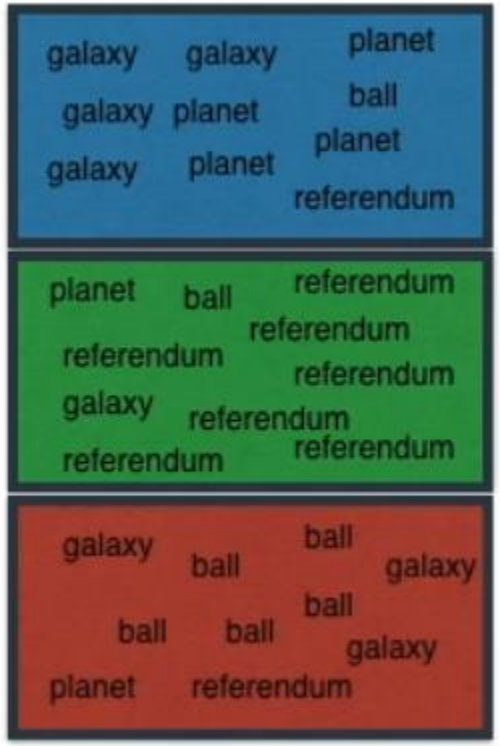
Galaxy	Planet	Ball	Referendum
0.3	0.1	0.5	0.1

Galaxy	Planet	Ball	Referendum
0.1	0.1	0.1	0.7

Wie funktioniert LDA?

- Schritt 4: Auf Ergebnis von Schritt 3 basierend: Erstellen einer Multinomialverteilung
  - In dieser Multinomialverteilung: Wahrscheinlichkeiten der vorherigen Ergebnisse bestimmen die Anzahl der Wörter, welche sich unter dem gegebenen Topic sammeln
    - Bsp: Topic 1 mit 70% „Regen“, 20% „Bundestag“ und 10% „Jupiter“, in der Multinomialverteilung von Topic 1 wären 7 mal Regen, 2 mal Bundestag und 1 mal Jupiter

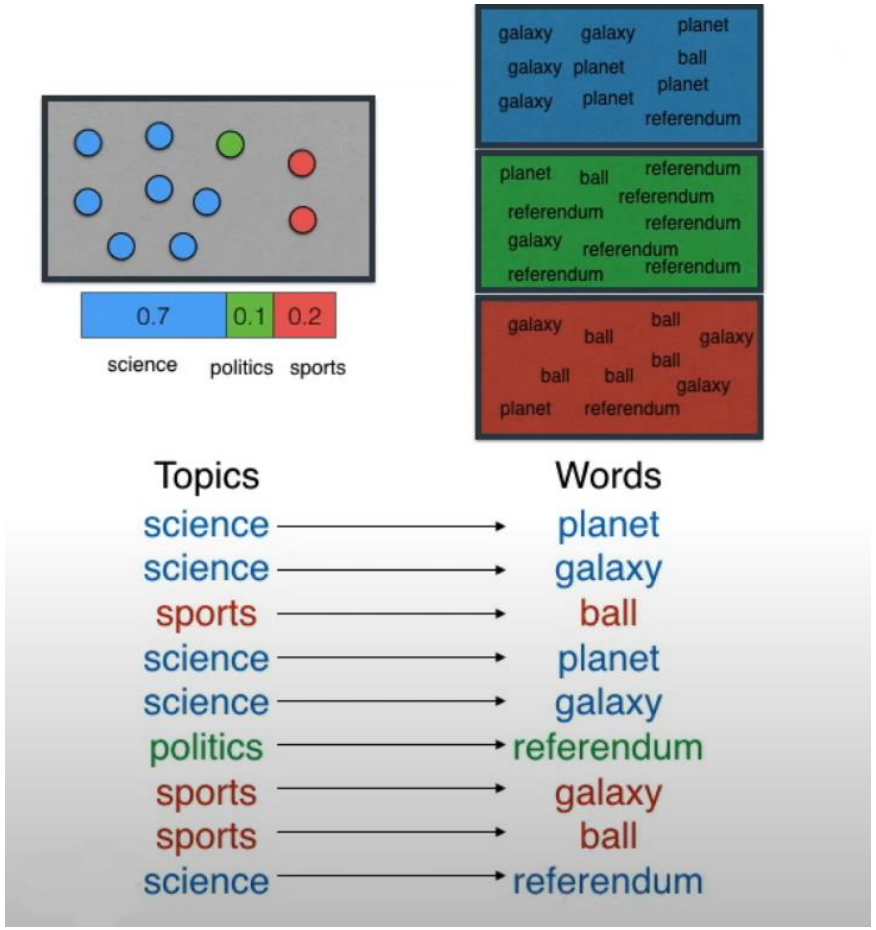
Wie funktioniert LDA?



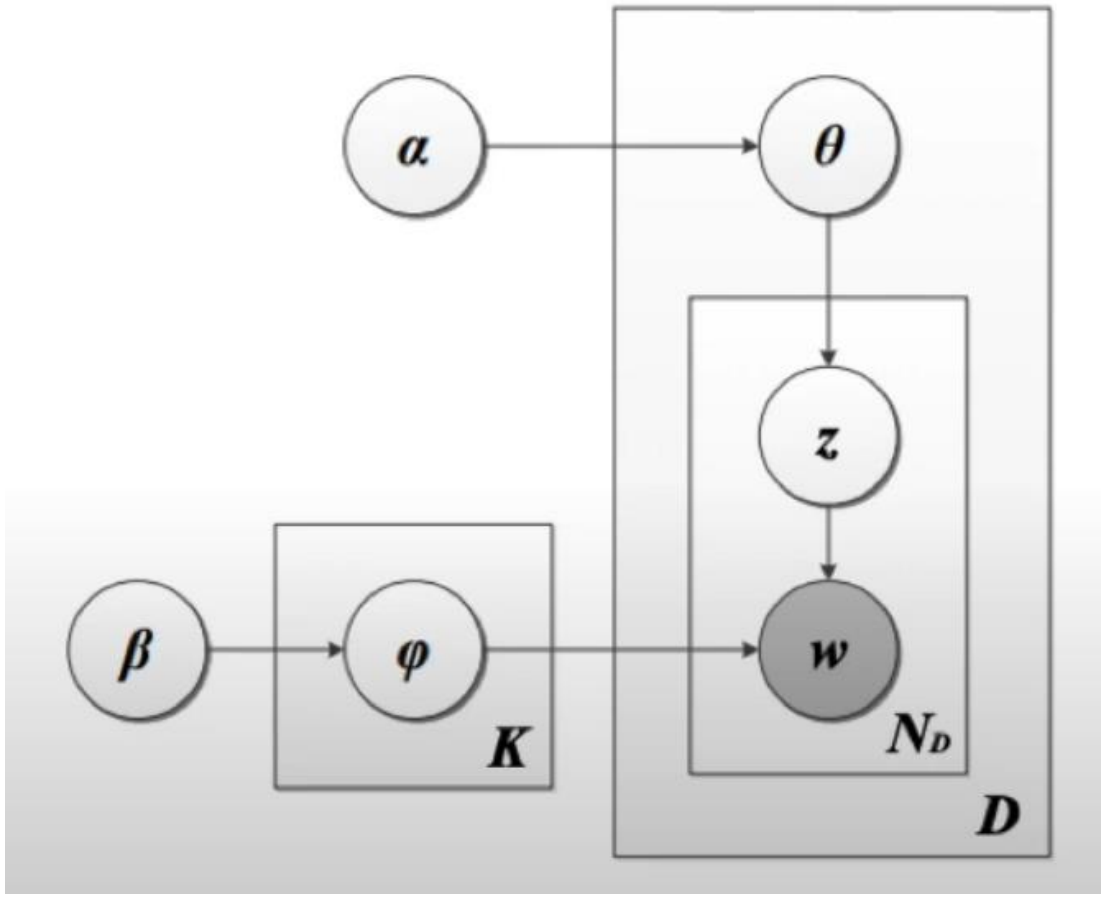
Wie funktioniert LDA?

- Schritt 5: Pro Topic aus Schritt 2 wird aus der Multinomialverteilung aus Schritt 4 ein Wort gezogen
- Dieses Wort mit dem aktuellen Topic wird dem Dokument hinzugefügt
  - Fortsetzen bis zur Vollendung des Dokuments
- 5 Schritte wiederholen, bis genug Dokumente vorhanden sind
- Letztendlich: überprüfen, welches Dokument die größte Ähnlichkeit mit dem Originaldokument hat
  - Die Einstellungen dieses Versuchs am besten
- Letzter Schritt: den Topics einen Sinn zuweisen (Arbeit für Menschen)

Wie funktioniert LDA?

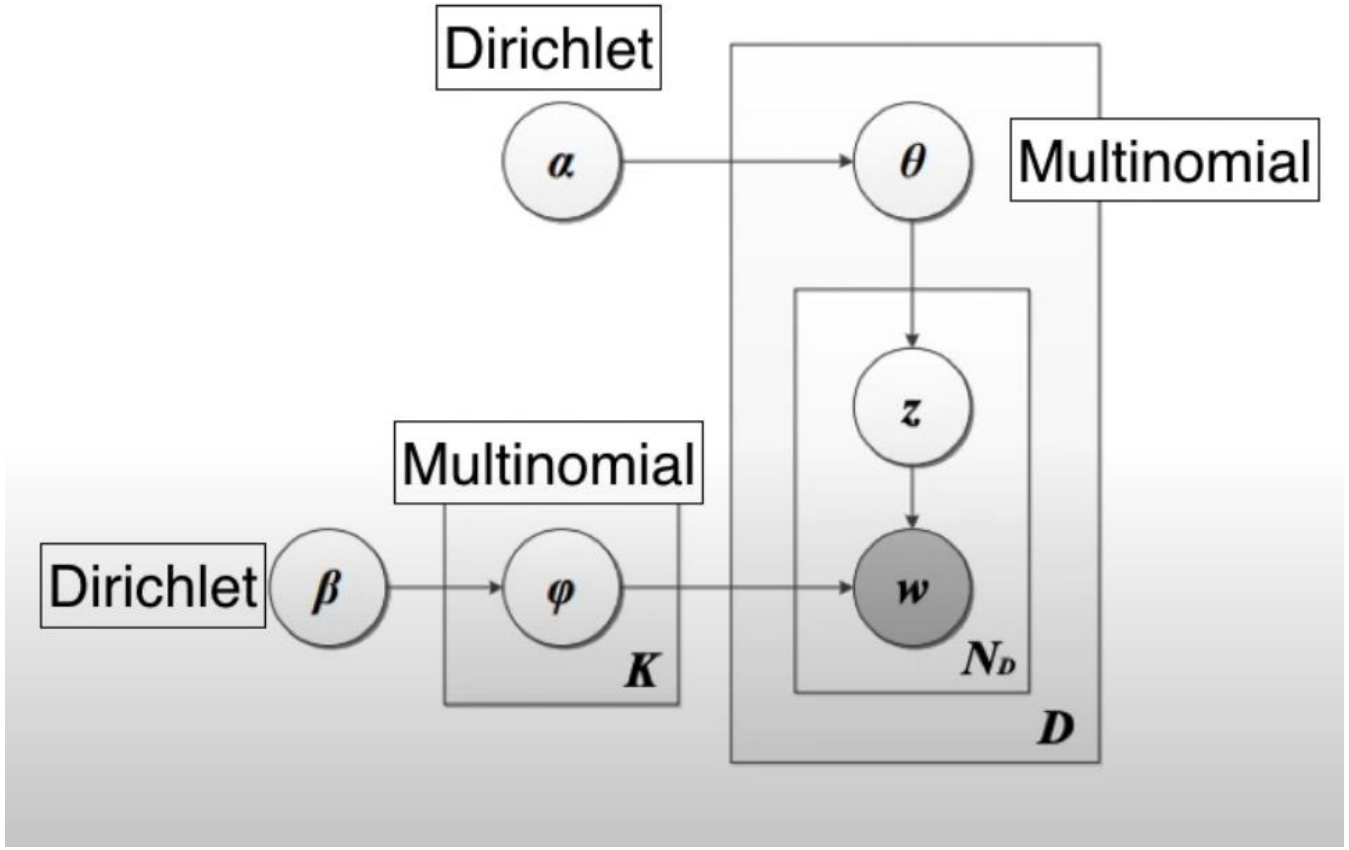


Mathematischer Aspekt



Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.

Wie funktioniert LDA?



Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.

Mathematischer Aspekt

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j)\, P(W_{j,t} \mid \varphi_{Z_{j,t}})$$

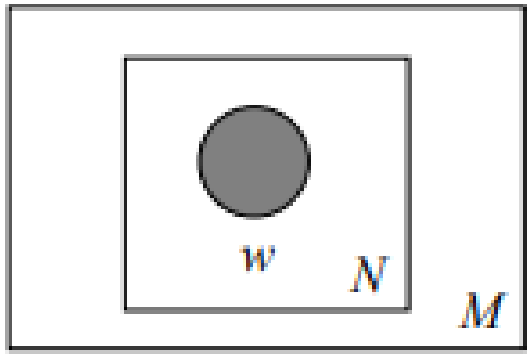
Mathematischer Aspekt

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j)\, P(W_{j,t} \mid \varphi_{Z_{j,t}})$$

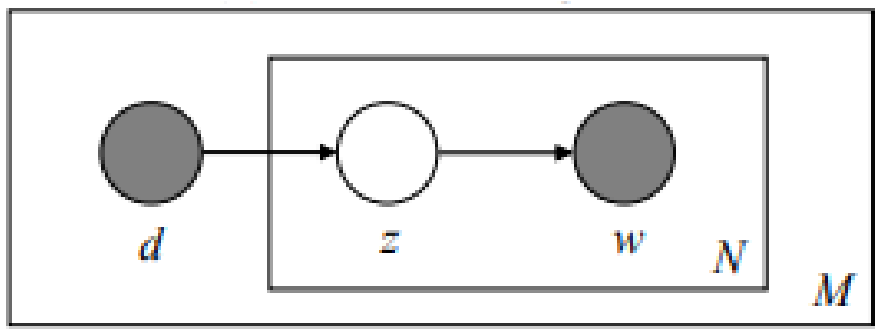
α	β	θ	φ
Schritt 1	Schritt 3	Schritt 2	Schritt 4

Unterschied zwischen LDA und anderen Modellen

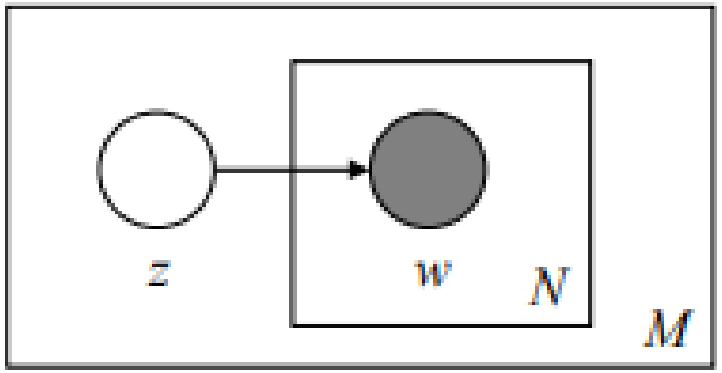
- Anzahl der Multinomialverteilungen
- Topic Variable
  - Topic pro Wort
  - Topic pro Dokument
- Anwendbarkeit auf Dokumente außerhalb der Trainingsdaten
- Overfitting?



(a) unigram



(c) pLSI/aspect model



(b) mixture of unigrams

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.

Nutzung von LDA

- Dokumentenclustering
- Semantisches Clustering von Begriffen
  - Auffinden von Synonymen
- Erweiterung: Author-Topic Model
  - Autorspezifische Topicverteilung

Beispiele

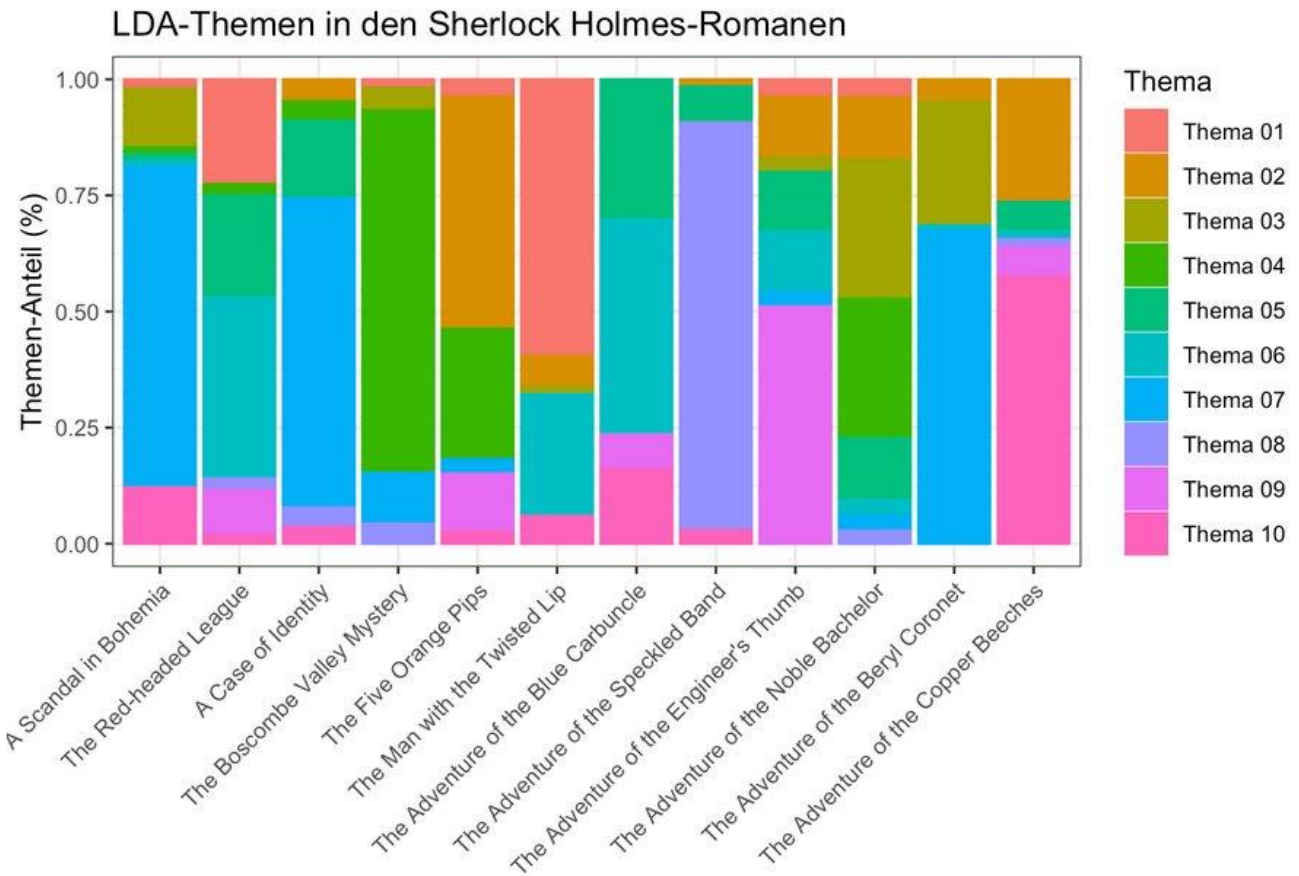
- Sherlock Holmes
  - 12 Geschichten
  - 10 Themen
- Zeit Nachrichtenkorpus
  - 12 Resorts
  - 15 Themen

http://inhaltsanalyse-mit-r.de/themenmodelle.html

Beispiele – Sherlock Holmes

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
clair	ever	went	father	something	business	hosmer	dr	light	rucastle
window	three	frank	mccarthy	enough	wilson	photograph	bed	colonel	hat
neville	locked	gentleman	lestrade	lady	knew	cried	sister	papers	wife
home	done	lady	son	business	pounds	angel	stoner	perhaps	mrs
friend	leave	holder	turner	name	answered	king	light	look	friend
lascar	put	make	papers	geese	stone	woman	roylott	floor	hunter
inspector	open	minutes	pool	seen	chair	hands	old	lamp	rather
o'clock	doubt	lane	boscombe	years	make	windibank	lady	hat	large
whole	lady	give	cry	put	sat	majesty	death	stone	hair
doctor	k	simon	point	open	name	coronet	band	goose	life

Beispiele – Sherlock Holmes



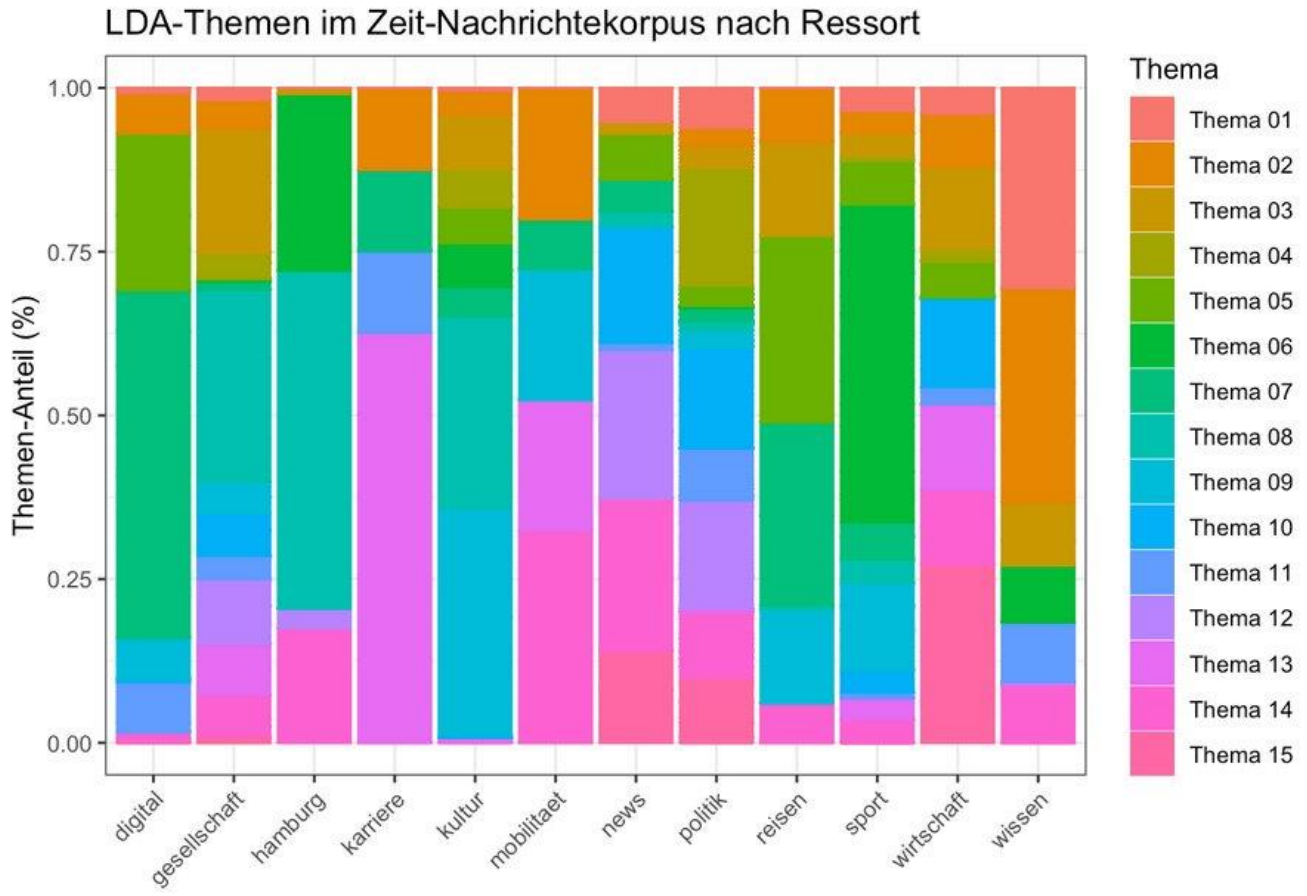
Beispiele – Zeit Nachrichtencorpus

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
wasser	forscher	online	partei	online	fc	google	polizei	welt
tepco	sogar	leben	afd	daten	trainer	steht	millionen	online
universität	wasser	eltern	politik	polizei	spiel	unternehmen	ja	serie
japanische	europa	ja	grünen	gesetz	dortmund	microsoft	kinder	sagen
krim	könnten	paris	spd	gespeichert	verein	nutzer	uhr	natürlich
reaktoren	fordern	warum	cdu	einfach	leben	kommt	damals	medien
china	halten	natürlich	wahl	app	bayern	windows	vielleicht	amazon
recht	studie	gerade	parteien	ja	frankfurt	vielleicht	sollten	ja
japan	millionen	kinder	stimmen	entwurf	leverkusen	facebook	stadt	spielen
bleiben	sonne	weniger	letzten	flüchtlinge	league	liegt	stuttgart	wohl

Beispiele – Zeit Nachrichtencorpus

Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
welt	usa	russland	getötet	mitarbeiter	bundesregierung	merkel
online	syrien	ukraine	stadt	unternehmen	bund	griechenland
serie	irak	putin	präsident	online	cdu	eu
sagen	iran	russische	demonstranten	woche	unternehmen	europa
natürlich	obama	russischen	boko	becker	kritik	milliarden
medien	behörden	separatisten	haram	weigelt	bericht	union
amazon	soldaten	online	mindestens	ulf	merkel	brüssel
ja	millionen	israel	kämpfen	ziele	deutsche	kanzlerin
spielen	flüchtlinge	präsident	polizei	führungskräfte	usa	abstimmung
wohl	könne	waffenruhe	verletzt	chef	spd	europäischen

Beispiele – Zeit Nachrichtenkorpus



Zusammenfassung

- LDA kann Dokumenten in einem Korpus bis zu k verschiedene Topics zuweisen
  - Prozess: erstellen möglichst ähnlicher Dokumente
  - Hintergrund: Zufallsverteilungen
  - Vergleich zu anderen:
    - kann auf Dokumente angewendet werden, die nicht Trainingsmaterial waren
    - kann mehrere Topics pro Dokument erkennen
    - man muss ein sinnvolles k finden
- Ergebnis nicht zwingend „eindeutig“

Quellen

- Blei, David M., Andrew Y. Ng, Michael I. Jordan. 2003. *Latent Dirichlet Allocation*. Journal of Machine Learning Research 3 (2003) 993-1022
- Serrano, Luis. 2020. *Latent Dirichlet Allocation (part 1 of 2),* zuletzt aufgerufen: 26.01.21
- https://www.youtube.com/watch?v=T05t-SqKArY&feature=youtu.be
- Heyer, Gerhard und Patrick Jähnichen. *Topicmodelle*. Universität Leipzig, zuletzt aufgerufen: 25.01.21
- https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwipg-GjlbfuAhVGJhoKHWz_DE0QFjAEegQIAhAC&url=http%3A%2F%2Fasv.informatik.uni-leipzig.de%2Fuploads%2Fdocument%2Ffile_link%2F382%2FTMI06_Topicmodelle2.pdf&usg=AOvVaw2ScKFdtvg43Az0p87FrIOU
- Heyer, Gerhard. *Dokumenten- und Topicmodelle*. Universität Leipzig, zuletzt aufgerufen: 25.01.21
- https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwipg-GjlbfuAhVGJhoKHWz_DE0QFjADegQIAxAC&url=http%3A%2F%2Fasv.informatik.uni-leipzig.de%2Fuploads%2Fdocument%2Ffile_link%2F321%2FLI08_Dokumenten_und_topicmodelle.pdf&usg=AOvVaw36WG7oUIBQvD8l1pfbpXn5
- *Automatisierte Inhaltsanalyse mit R*, zuletzt aufgerufen: 25.01.21
- http://inhaltsanalyse-mit-r.de/themenmodelle.html