How to infer topics from a text collection?

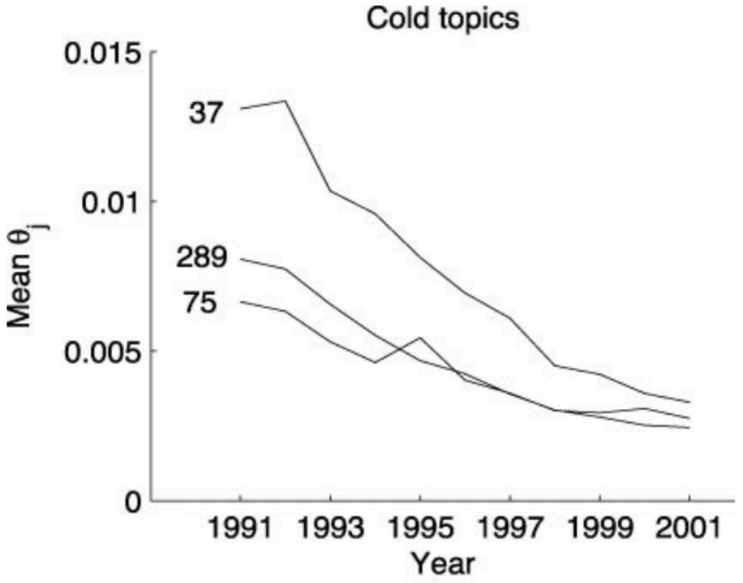
Sandeep Soni

09/24/2024

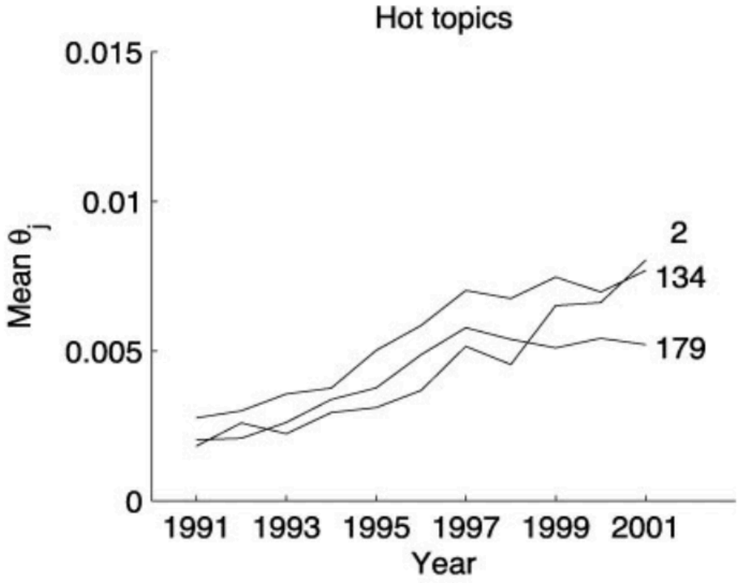
Some possible Research questions

- What are some themes within a given document collection?
- What themes are popular, trending, or out-of-fashion?
- What are similar themes? What are dissimilar themes?

- What are the different research areas?
- Which ones are in vogue and which have fell out of fashion?



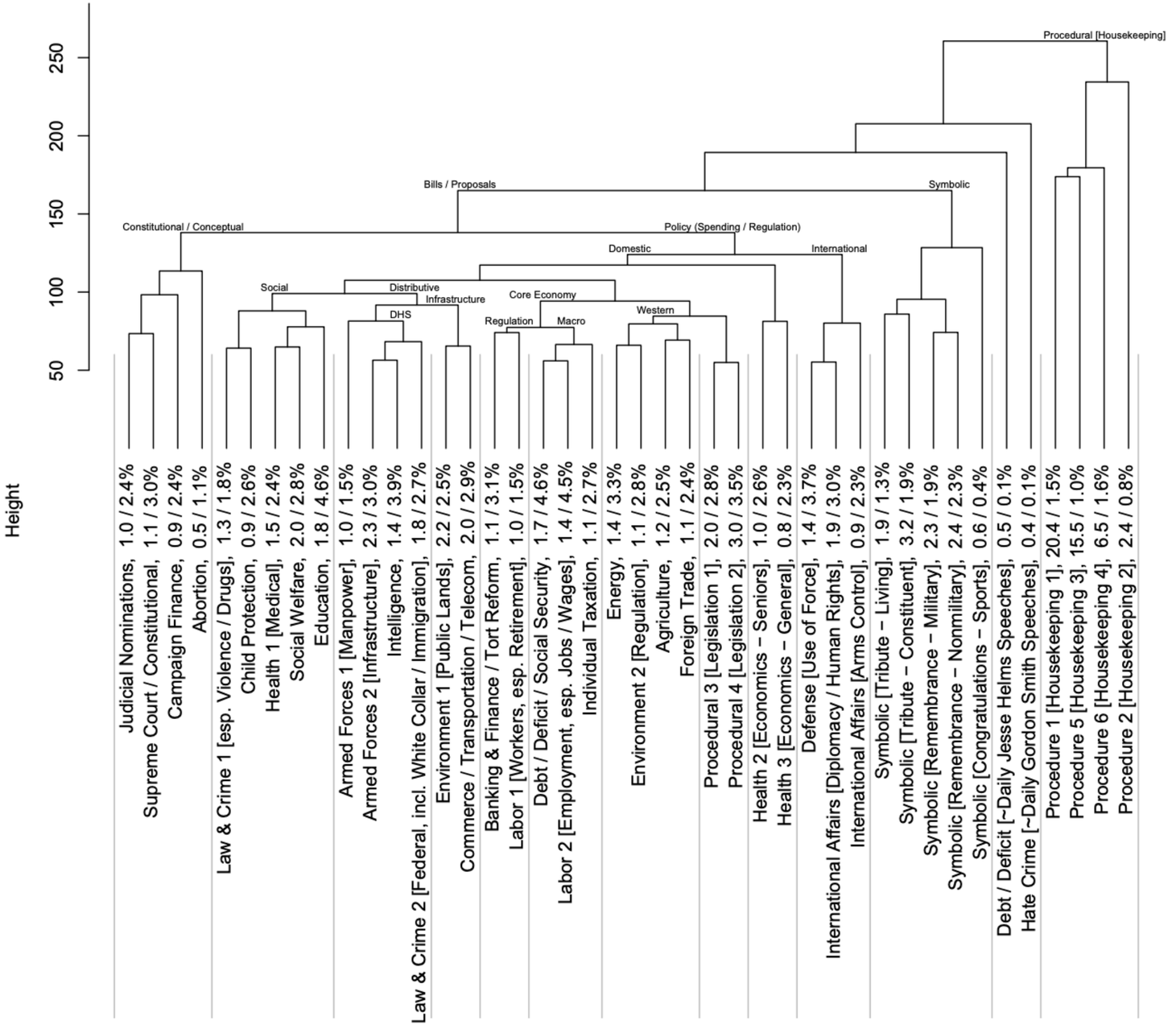
37	289	75
CDNA	KDA	ANTIBODY
AMINO	PROTEIN	ANTIBODIES
SEQUENCE	PURIFIED	MONOCLONAL
ACID	MOLECULAR	ANTIGEN
PROTEIN	MASS	IGG
ISOLATED	CHROMATOGRAPHY	MAB
ENCODING	POLYPEPTIDE	SPECIFIC
CLONED	GEL	EPITOPE
ACIDS	SDS	HUMAN
IDENTITY	BAND	MABS
CLONE	APPARENT	RECOGNIZED
EXPRESSED	LABELED	SERA



2	134	179
SPECIES	MICE	APOPTOSIS
GLOBAL	DEFICIENT	DEATH
CLIMATE	NORMAL	CELL
CO2	GENE	INDUCED
WATER	NULL	BCL
ENVIRONMENTAL	MOUSE	CELLS
YEARS	TYPE	APOPTOTIC
MARINE	HOMOZYGOUS	CASPASE
CARBON	ROLE	FAS
DIVERSITY	KNOCKOUT	SURVIVAL
OCEAN	DEVELOPMENT	PROGRAMMED
EXTINCTION	GENERATED	MEDIATED

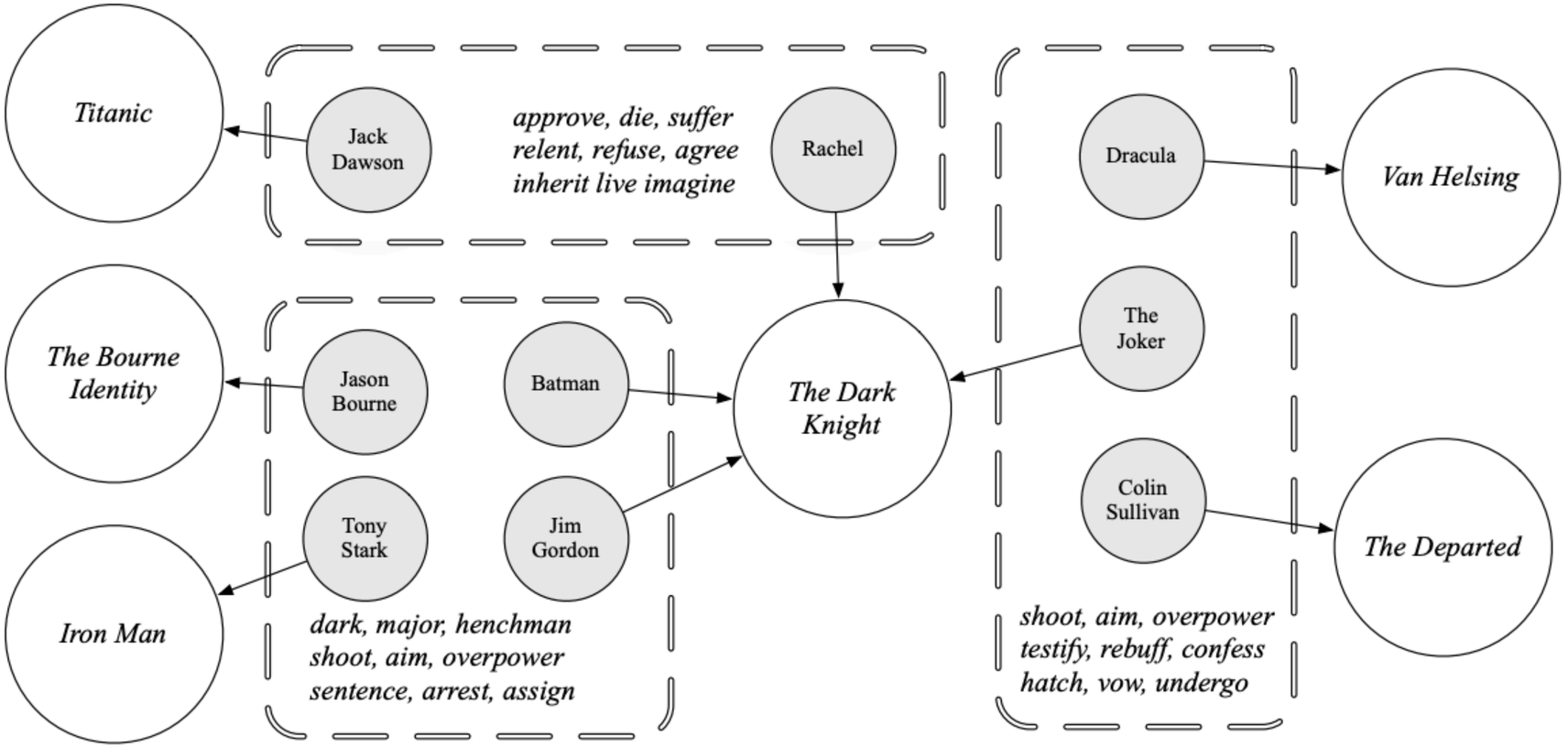
Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." *Proceedings of the National academy of Sciences* 101.suppl_1 (2004): 5228-5235.

- Which topics are politicians paying attention?



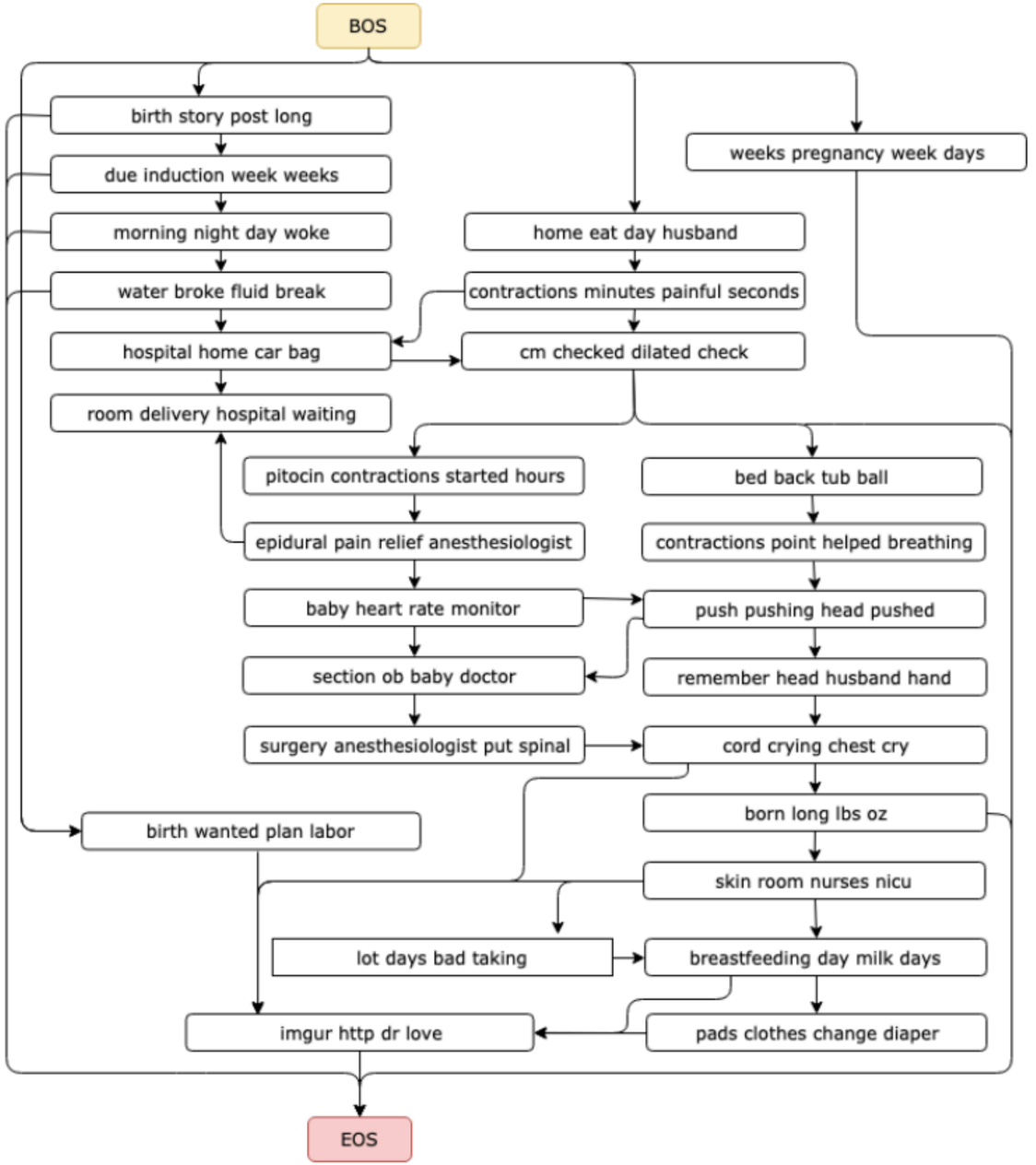
Quinn, Kevin M., et al. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54.1 (2010): 209-228.

- What are the different types of characters in movies?



Bamman, David, Brendan O'Connor, and Noah A. Smith. "Learning latent personas of film characters." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013.

- What are narrative sequences in birth stories?
- When do expecting mothers feel most vulnerable?



Antoniak, Maria, David Mimno, and Karen Levy. "Narrative paths and negotiation of power in birth stories." *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019): 1-27.

Question for the day

"How to find topics given a document collection?"

Class activity I

We live in a time of extraordinary change -- change that's reshaping the way we live, the way we work, our planet, our place in the world. It's change that promises amazing medical breakthroughs, but also economic disruptions that strain working families. It promises education for girls in the most remote villages, but also connects terrorists plotting an ocean away. It's change that can broaden opportunity, or widen inequality. And whether we like it or not, the pace of this change will only accelerate.

Class activity I

We live in a time of extraordinary change -- change that's reshaping the way we live, the way we work, our planet, our place in the world. It's change that promises amazing medical breakthroughs, but also economic disruptions that strain working families. It promises education for girls in the most remote villages, but also connects terrorists plotting an ocean away. It's change that can broaden opportunity, or widen inequality. And whether we like it or not, the pace of this change will only accelerate.

Class activity I

We live in a time of extraordinary change -- change that's reshaping the way we live, the way we work, our planet, our place in the world. It's change that promises amazing medical breakthroughs, but also economic disruptions that strain working families. It promises education for girls in the most remote villages, but also connects terrorists plotting an ocean away. It's change that can broaden opportunity, or widen inequality. And whether we like it or not, the pace of this change will only accelerate.

Agenda

- Intro to topic models
- LDA
- Variations of LDA
- Evaluation

Document-Term Matrix

- If this matrix is called D, then D_{ij} is the number of times the w_j appeared in d_i
- Every document is a distribution over word counts

	w_1	w_2	…	w_m
d_1	0	3	…	1
d_2	2	1	…	0
…	…	…	…	
d_n	1	0	…	0

Matrix decomposition

D

	w_1	w_2	…	w_m
d_1	0	3	…	1
d_2	2	1	…	0
…	…	…	…	
d_n	1	0	…	0

$\approx$

U

	t_1	t_2	…	t_k
d_1	0.3	0.15	…	0.1
d_2	0.45	0.05	…	0.2
…	…	…	…	…
d_n	0.1	0.25	…	0

$\times$

V

	w_1	w_2	…	w_m
t_1	0.1	0.2	…	0.25
t_2	0.15	0.2	…	0.15
…	…	…	…	…
t_k	0.05	0.1	…	0.1

Matrix decomposition

D

	w_1	w_2	…	w_m
d_1	0	3	…	1
d_2	2	1	…	0
…	…	…	…	
d_n	1	0	…	0

$\approx$

U

	t_1	t_2	…	t_k
d_1	0.3	0.15	…	0.1
d_2	0.45	0.05	…	0.2
…	…	…	…	…
d_n	0.1	0.25	…	0

$\times$

V

	w_1	w_2	…	w_m
t_1	0.1	0.2	…	0.25
t_2	0.15	0.2	…	0.15
…	…	…	…	…
t_k	0.05	0.1	…	0.1

Matrix decomposition

D

	w_1	w_2	…	w_m
d_1	0	3	…	1
d_2	2	1	…	0
…	…	…	…	
d_n	1	0	…	0

$\approx$

U

	t_1	t_2	…	t_k
d_1	0.3	0.15	…	0.1
d_2	0.45	0.05	…	0.2
…	…	…	…	…
d_n	0.1	0.25	…	0

$\times$

V

	w_1	w_2	…	w_m
t_1	0.1	0.2	…	0.25
t_2	0.15	0.2	…	0.15
…	…	…	…	…
t_k	0.05	0.1	…	0.1

- $D_{i,j} \approx \sum_{r=1}^{k} U_{i,k} \times V_{k,j}$
- Each document can be thought of as a mixture of K “topics”
- Each topic is a distribution over words

Latent semantic analysis

	w_1	w_2	…	w_m
d_1	0	3	…	1
d_2	2	1	…	0
…	…	…	…	
d_n	1	0	…	0

$\approx$

	t_1	t_2	…	t_k
d_1	0.3	0.15	…	0.1
d_2	0.45	0.05	…	0.2
…	…	…	…	…
d_n	0.1	0.25	…	0

$\times$

	w_1	w_2	…	w_m
t_1	0.1	0.2	…	0.25
t_2	0.15	0.2	…	0.15
…	…	…	…	…
t_k	0.05	0.1	…	0.1

Latent semantic analysis

	w_1	w_2	…	w_m
d_1	0	3	…	1
d_2	2	1	…	0
…	…	…	…	
d_n	1	0	…	0

$\approx$

	t_1	t_2	…	t_k
d_1	0.3	0.15	…	0.1
d_2	0.45	0.05	…	0.2
…	…	…	…	…
d_n	0.1	0.25	…	0

$\times$

	w_1	w_2	…	w_m
t_1	0.1	0.2	…	0.25
t_2	0.15	0.2	…	0.15
…	…	…	…	…
t_k	0.05	0.1	…	0.1

Can we do this probabilistically?

Probability basics

- Let's assume that we have a vocabulary V = {"the", "dog", "runs"}

X	Y	Count(X,Y)	P(X,Y)
"the"	"dog"	45	
"the"	"runs"	15	
"dog"	"runs"	40	
"dog"	"the"	25	
"runs"	"the"	30	
"runs"	"dog"	45	

Probability basics

- $P(X,Y) = \frac{Count(X,Y)}{N}$
- N is the number of pairwise cooccurrences
- This is called the joint probability

X	Y	Count(X,Y)	P(X,Y)
"the"	"dog"	45	0.225
"the"	"runs"	15	0.075
"dog"	"runs"	40	0.2
"dog"	"the"	25	0.125
"runs"	"the"	30	0.15
"runs"	"dog"	45	0.225

N=200

Marginalization

- What is P(X="dog")?

X	Y	Count(X,Y)	P(X,Y)
"the"	"dog"	45	0.225
"the"	"runs"	15	0.075
"dog"	"runs"	40	0.2
"dog"	"the"	25	0.125
"runs"	"the"	30	0.15
"runs"	"dog"	45	0.225

Marginalization

- $$P(X) = \sum_{y} P(X, Y = y)$$

- P(X="dog") =
  P(X="dog",Y="runs") +
  P(X="dog",Y="the")

- P(X="dog") = 0.2 + 0.125
  = 0.325

X	Y	Count(X,Y)	P(X,Y)
"dog"	"runs"	40	0.2
"dog"	"the"	25	0.125

Conditional probability

- P(X|Y="dog")

X	Y	Count(X,Y)	P(X,Y)
"the"	"dog"	45	0.225
"the"	"runs"	15	0.075
"dog"	"runs"	40	0.2
"dog"	"the"	25	0.125
"runs"	"the"	30	0.15
"runs"	"dog"	45	0.225

Conditional probability

- P(X|Y="dog")
- This is a distribution
- Since it should be a probability distribution, we'll normalize it to sum up to 1

X	Y	Count(X,Y)	P(X,Y)
"the"	"dog"	45	0.225
"runs"	"dog"	45	0.225

X	Y	Count(X,Y)	P(X|Y="dog")
"the"	"dog"	45	0.5
"runs"	"dog"	45	0.5

Discrete probability distributions

X=1	Pr(X=1)
X=2	Pr(X=2)
X=3	Pr(X=3)
…	…
X=n	Pr(X=n)

P(X)

Marginal distribution

X=1	Y=1	Pr(X=1,Y=1)
X=1	Y=2	Pr(X=1,Y=2)
…	…	…
X=1	Y=k	Pr(X=1,Y=k)
X=2	Y=1	Pr(X=2,Y=1)
X=2	Y=2	Pr(X=2,Y=1)
…	…	…
X=n	Y=k	Pr(X=n,Y=k)

P(X,Y)

Joint distribution

Y=1	Pr(Y=1 | X=1)	
…	…	X=1
Y=k	Pr(Y=k | X=1)	
Y=1	Pr(Y=1 | X=2)	
…	…	X=2
Y=k	Pr(Y=k | X=2)	
…		…
Y=1	Pr(Y=1 | X=n)	
…	…	X=n
Y=k	Pr(Y=k | X=n)	

P(Y|X)

Conditional distribution

Discrete probability distributions

X=1	Pr(X=1)
X=2	Pr(X=2)
X=3	Pr(X=3)
…	…
X=n	Pr(X=n)

P(X)

Marginal distribution

X=1	Y=1	Pr(X=1,Y=1)
X=1	Y=2	Pr(X=1,Y=2)
…	…	…
X=1	Y=k	Pr(X=1,Y=k)
X=2	Y=1	Pr(X=2,Y=1)
X=2	Y=2	Pr(X=2,Y=1)
…	…	…
X=n	Y=k	Pr(X=n,Y=k)

P(X,Y)

Joint distribution

Y=1	Pr(Y=1 | X=1)	
…	…	X=1
Y=k	Pr(Y=k | X=1)	
Y=1	Pr(Y=1 | X=2)	
…	…	X=2
Y=k	Pr(Y=k | X=2)	
…		…
Y=1	Pr(Y=1 | X=n)	
…	…	X=n
Y=k	Pr(Y=k | X=n)	

P(Y|X)

Conditional distribution

How many numbers are needed to parameterize these probability distributions?

Discrete probability distributions

X=1	Pr(X=1)
X=2	Pr(X=2)
X=3	Pr(X=3)
…	…
X=n	Pr(X=n)

P(X)

X=1	Y=1	Pr(X=1,Y=1)
X=1	Y=2	Pr(X=1,Y=2)
…	…	…
X=1	Y=k	Pr(X=1,Y=k)
X=2	Y=1	Pr(X=2,Y=1)
X=2	Y=2	Pr(X=2,Y=1)
…	…	…
X=n	Y=k	Pr(X=n,Y=k)

P(X,Y)

Y=1	Pr(Y=1 | X=1)	
…	…	X=1
Y=k	Pr(Y=k | X=1)	
Y=1	Pr(Y=1 | X=2)	
…	…	X=2
Y=k	Pr(Y=k | X=2)	
…		…
Y=1	Pr(Y=1 | X=n)	
…	…	X=n
Y=k	Pr(Y=k | X=n)	

P(Y|X)

Marginal distribution

Joint distribution

Conditional distribution

N-1

N*K - 1

N*K - N

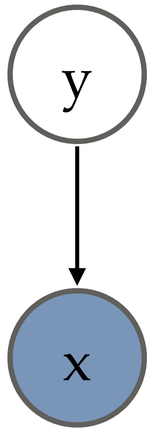
Factorization

- $P(X,Y) = P(X)P(Y|X)$
- $P(X,Y) = P(Y)P(X|Y)$

X	Y	Count(X,Y)	P(X,Y)
“the”	“dog”	45	0.225
“the”	“runs”	15	0.075
“dog”	“runs”	40	0.2
“dog”	“the”	25	0.125
“runs”	“the”	30	0.15
“runs”	“dog”	45	0.225

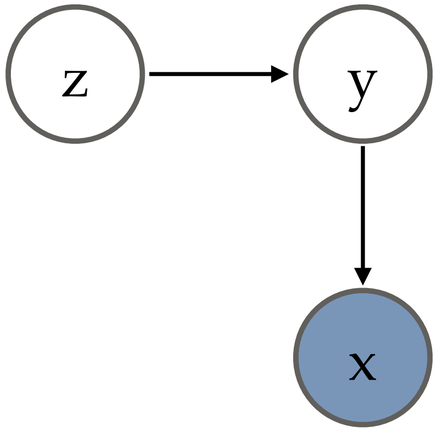
Probabilistic Graphical models

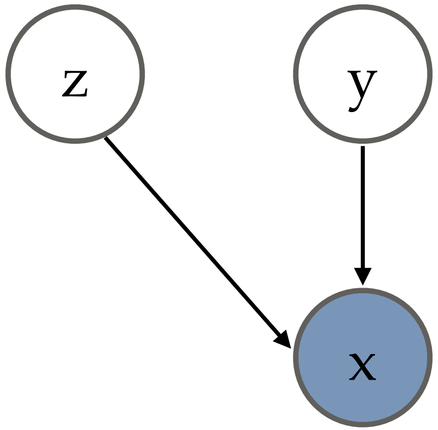
- All nodes are random variables; dark nodes are observed and the others are latent
- Arrows indicate conditional relationships
- Visual specification of how the joint distribution factors

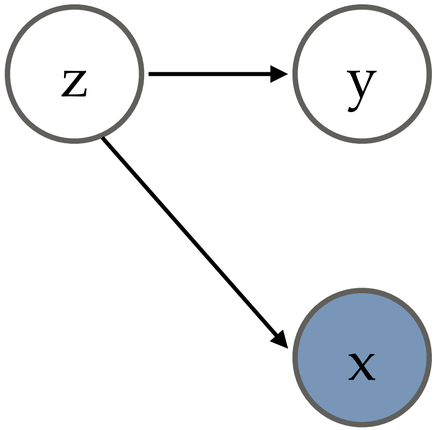


$$P(x, y) = P(y)P(x \mid y)$$

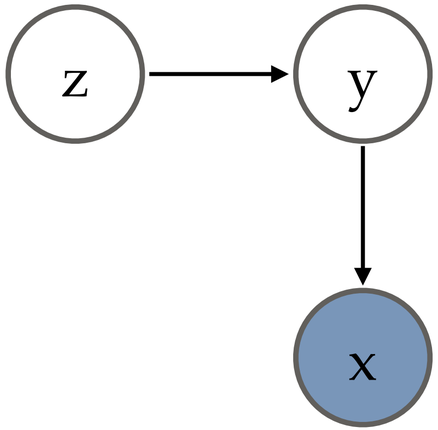
Class Activity II



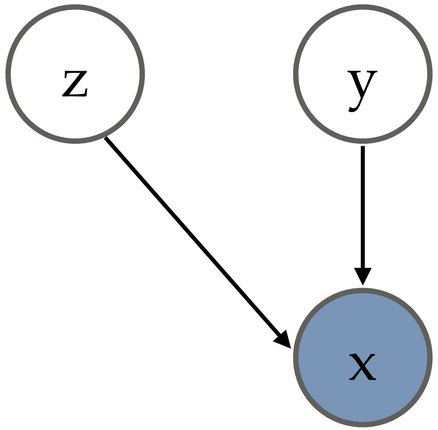




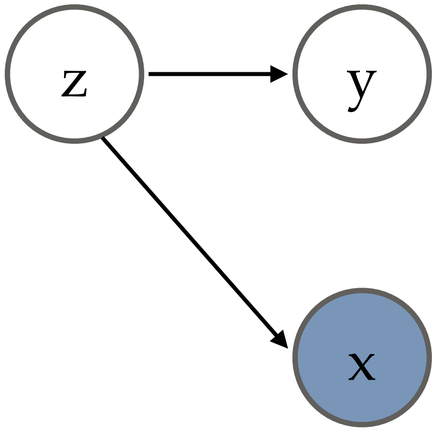
Class Activity II



$$P(x, y, z) = P(z)P(y \mid z)P(x \mid y)$$



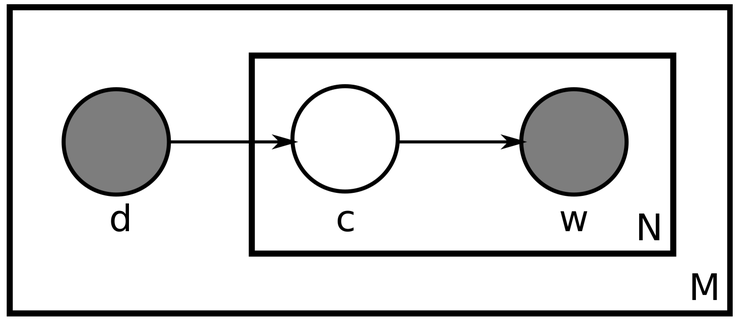
$$P(x, y, z) = P(z)P(y)P(x \mid y, z)$$



$$P(x, y, z) = P(z)P(y \mid z)P(x \mid z)$$

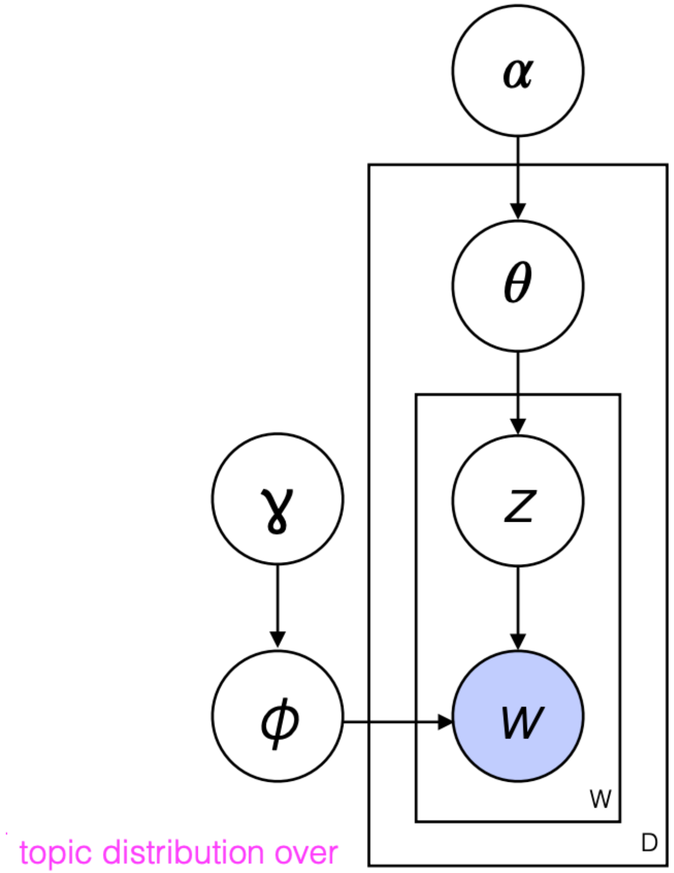
Probabilistic Latent semantic Analysis

- There are M documents and N words in the vocabulary.
- Blocks surrounding the circle means repeating structure
- We observe document indices and the words within each document
- c is the topic assigned to every word



- Generative story:
  - Generate doc indices
  - For every doc index
    - generate the topic
    - Generate word for the topic

Latent Dirichlet Allocation



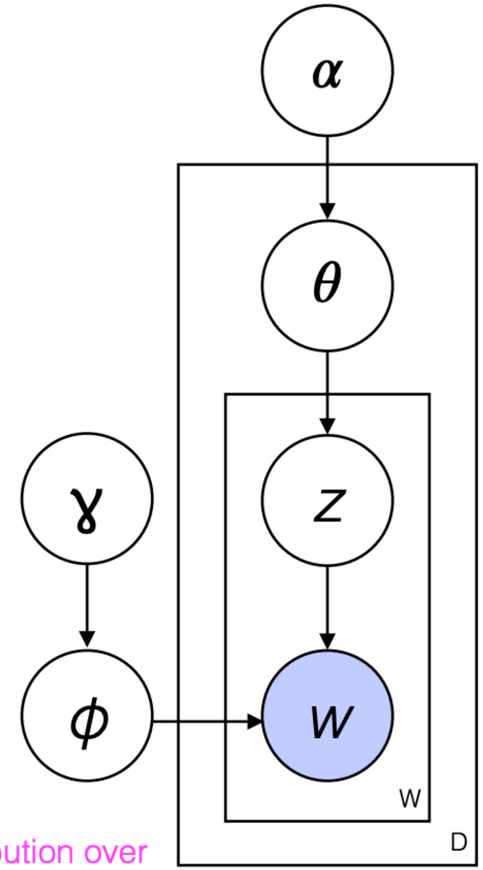
document distribution over topics

topic indicators for words

words

topic distribution over words

Topic models



document distribution over topics

topic indicators for words

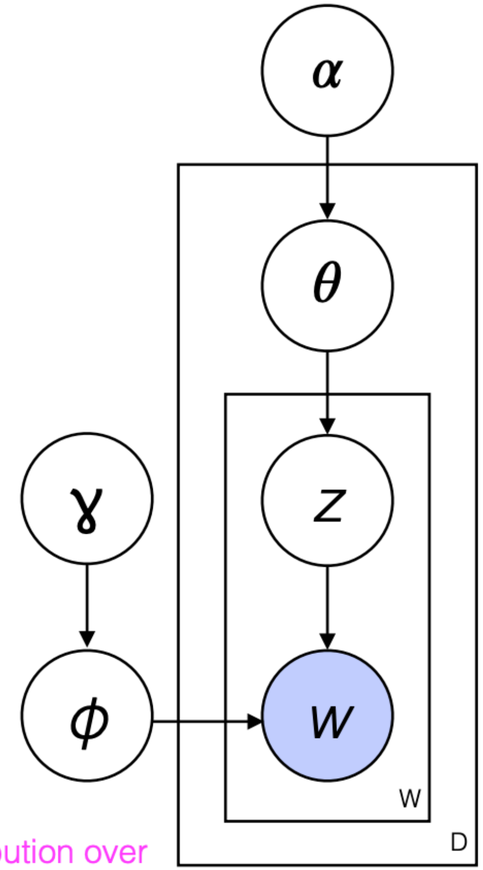
words

topic distribution over words

Topic Models

- A document is conceptualized as a distribution over topics
- θ gives this distribution

Topic models



document distribution over topics

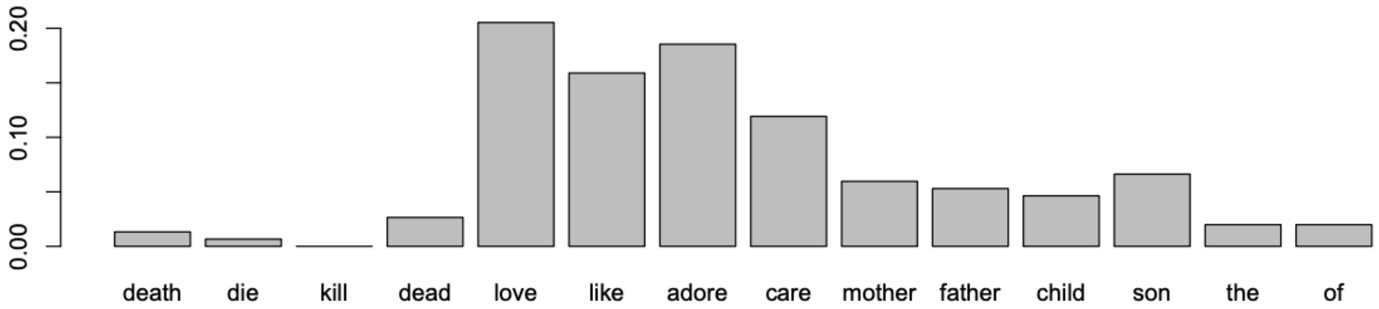
topic indicators for words

words

topic distribution over words

Topic Models

- A topic is a distribution over words
- ϕ gives this distribution
- P("like"|topic=love)=0.15

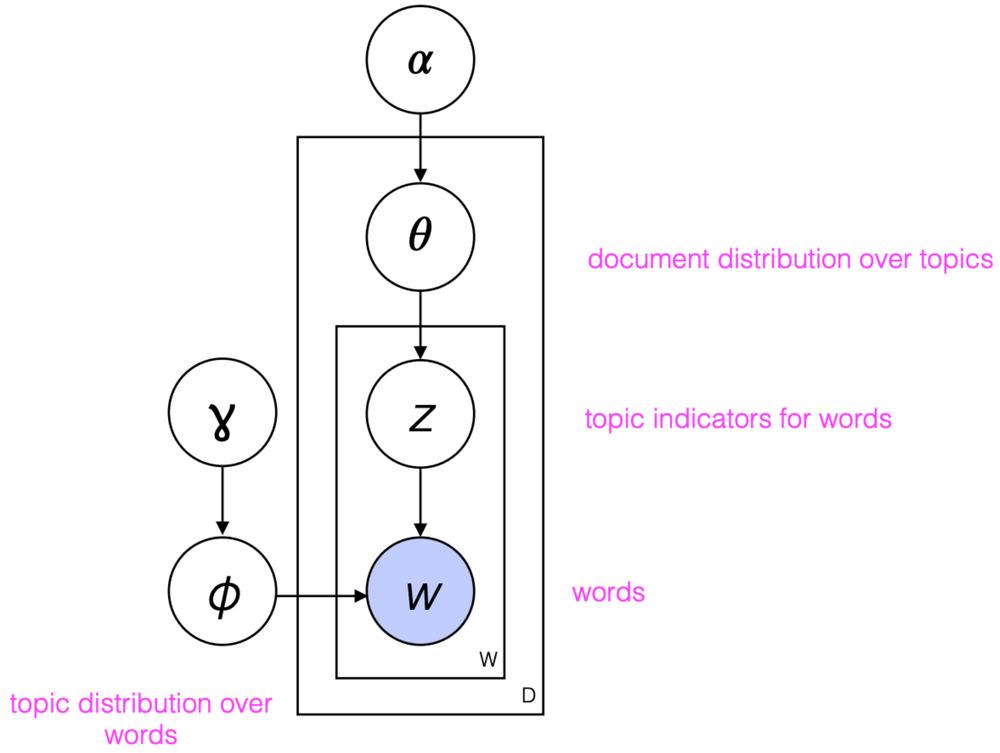


Topic Models

- For every document, you will have a different topic distribution
- For every topic, you will have a different word distribution

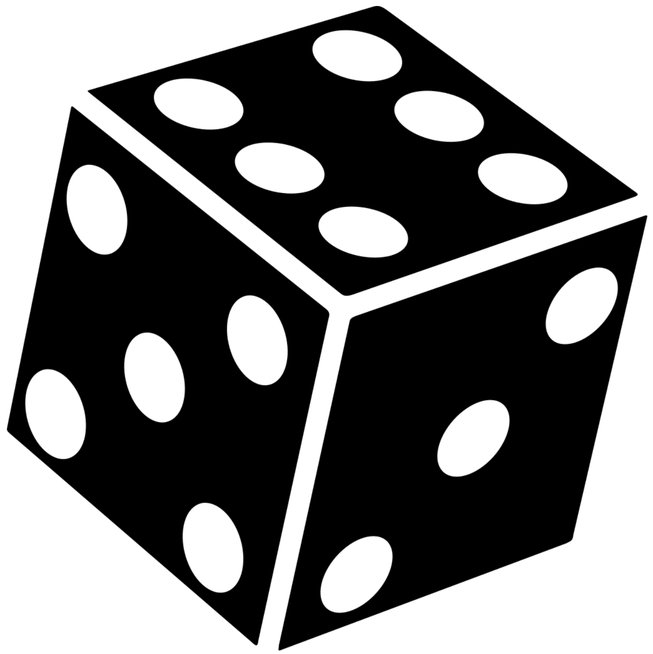
Generative Story

- For every document d:
  - Generate topic distribution θ_d
  - For every word position in d:
    - Draw topic $z_{w,d}$
    - Draw word w from $\phi_{z_{w,d}}$



Generative Story

- For every document d:
  - Generate topic distribution θ_d
  - For every word position in d:
    - Draw topic $z_{w,d}$
    - Draw word w from $\phi_{z_{w,d}}$



Generate first word

θ_d = {“love”=0.5,

“war”=0.2,

“aliens” = 0.3}

like

- $z_{w,d}$ = “love”
- w = like

ϕ_{love}

Generate Next word

θ_d = {“love”=0.5,

“war”=0.2,

“aliens” = 0.3}

like peace

- $z_{w,d}$ = “war”
- w = peace

ϕ_{war}

Generate Next word

θ_d = {“love”=0.5,

“war”=0.2,

“aliens” = 0.3}

like peace love

- $z_{w,d}$ = “love”
- w = love

ϕ_{love}

Generate Next word

θ_d = {“love”=0.5,

“war”=0.2,

“aliens” = 0.3}

like peace love people

- $z_{w,d}$ = “aliens”
- w = people

ϕ_{aliens}

And so on

like peace love people …

Software

Software

- Mallet (https://mimno.github.io/Mallet/)

Software

- Mallet (https://mimno.github.io/Mallet/)
- Gensim (https://radimrehurek.com/gensim/)

Software

- Mallet (https://mimno.github.io/Mallet/)
- Gensim (https://radimrehurek.com/gensim/)
- BertTopic (https://maartengr.github.io/BERTopic/index.html)

Inference

Inference

- How do we estimate the latent variables and the probability distributions?

Inference

Inference

- Randomly initialize document and topic distributions; randomly assign topics to each word

Inference

- Randomly initialize document and topic distributions; randomly assign topics to each word
- Repeat until convergence:

Inference

- Randomly initialize document and topic distributions; randomly assign topics to each word
- Repeat until convergence:
  - E-Step: Update topic assignments based on current estimated distributions

Inference

- Randomly initialize document and topic distributions; randomly assign topics to each word
- Repeat until convergence:
  - E-Step: Update topic assignments based on current estimated distributions
  - M-Step: Update distribution based on current topic assignments

Inference

Inference

- How do we estimate the distributions?

Inference

- How do we estimate the distributions?
  - Markov chain Monte Carlo (Metropolis Hastings, Gibbs sampling)

Inference

- How do we estimate the distributions?
  - Markov chain Monte Carlo (Metropolis Hastings, Gibbs sampling)
  - Variational methods

Inference

- How do we estimate the distributions?
  - Markov chain Monte Carlo (Metropolis Hastings, Gibbs sampling)
  - Variational methods
  - Hybrid methods

Things to remember

Things to remember

- Priors [Wallach et. al. 2009]

Things to remember

- Priors [Wallach et. al. 2009]
- Preprocessing [Schofield et. al. 2016, 2017]

Things to remember

- Priors [Wallach et. al. 2009]
- Preprocessing [Schofield et. al. 2016, 2017]
- Inference techniques

LDA Assumptions

- A word gets a single topic assigned
- A document gets a single distribution of topics
- No sequential info. used
- No other structure assumed except word and document identities

Some popular topic models

- Probabilistic graphical models offer much flexibility in modeling the generation of documents

Hierarchical topic modeling

- Model the data generating process as a nested Chinese restaurant process

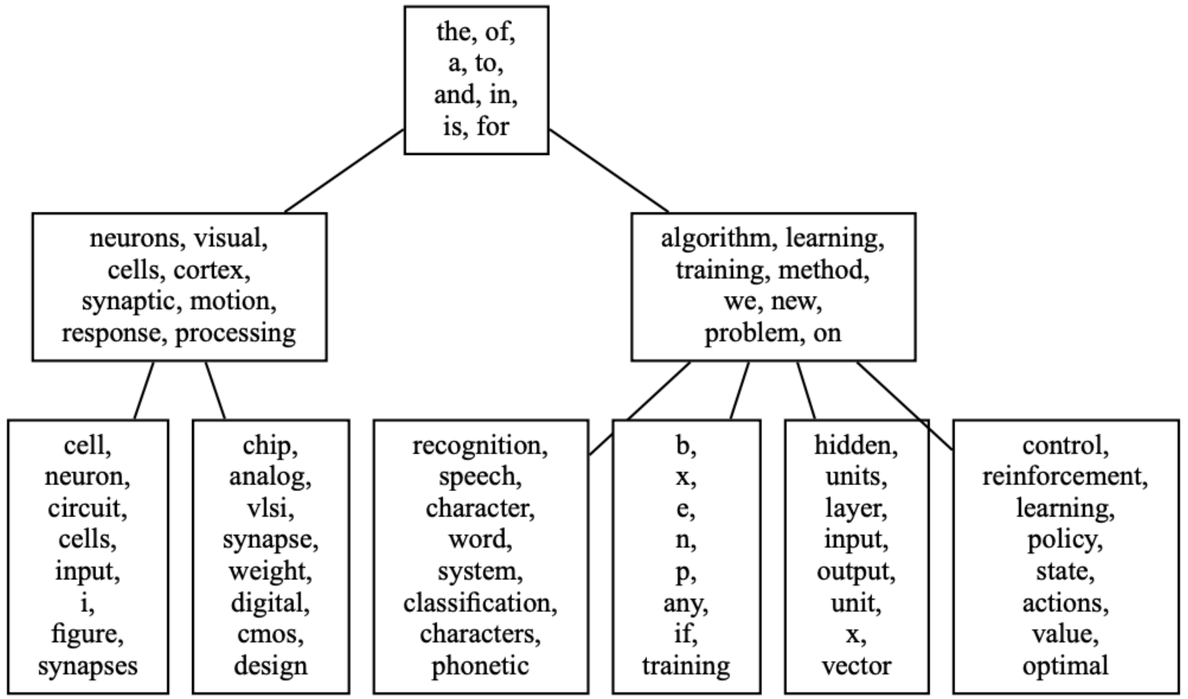


Figure 5: A topic hierarchy estimated from 1717 abstracts from NIPS01 through NIPS12. Each node contains the top eight words from its corresponding topic distribution.

Griffiths, Thomas, et al. "Hierarchical topic models and the nested Chinese restaurant process." *Advances in neural information processing systems* 16 (2003).

Geographical topic model

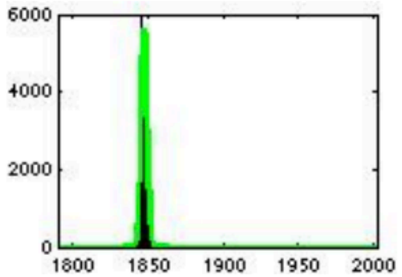
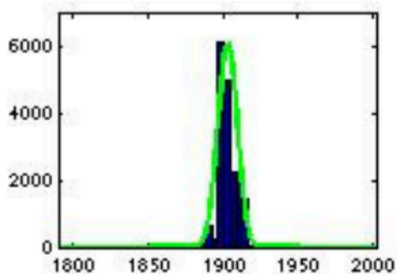
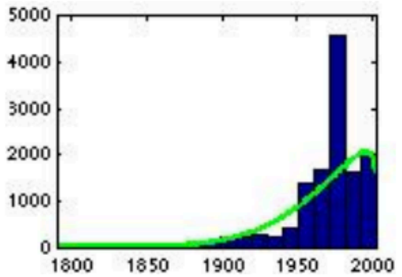
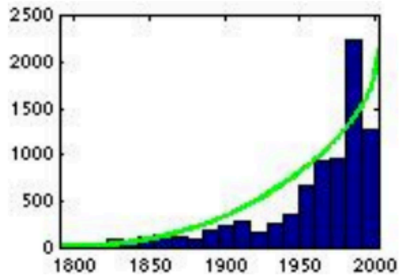
- Model text and author geolocation

	"basketball"	"popular music"	"daily life"	"emoticons"	"chit chat"
	PISTONS KOBE LAKERS game DUKE NBA CAVS STUCKEY JETS KNICKS	album music beats artist video #LAKERS ITUNES tour produced vol	tonight shop weekend getting going chilling ready discount waiting iam	:) haha :d :(;) :p xd :/ hahaha hahah	lol smh jk yea wyd coo ima wassup somethin jp
 Boston	CELTICS victory BOSTON CHARLOTTE	playing daughter PEARL alive war comp	BOSTON	;p gna loveee	*ese* exam suttin sippin
 N. California	THUNDER KINGS GIANTS pimp trees clap	SIMON dl mountain seee	6am OAKLAND	*pues* hella koo SAN fckn	hella flirt hut iono OAKLAND
 New York	NETS KNICKS	BRONX	iam cab	oww	wasssup nm
 Los Angeles	#KOBE #LAKERS AUSTIN	#LAKERS load HOLLYWOOD imm MICKEY TUPAC	omw tacos hr HOLLYWOOD	af *papi* raining th bomb coo HOLLYWOOD	wyd coo af *nada* tacos messin fasho bomb
 Lake Erie	CAVS CLEVELAND OHIO BUCKS od COLUMBUS	premiere prod joint TORONTO onto designer CANADA village burr	stink CHIPOTLE tipsy	;d blvd BIEBER hve OHIO	foul WIZ salty excuses lames officer lastnight

Eisenstein, Jacob, et al. "A latent variable model for geographic lexical variation." *Proceedings of the 2010 conference on empirical methods in natural language processing*. 2010.

Topics over time

- Text and time
- Time is discrete

Mexican War		Panama Canal		Cold War		Modern Tech	
							
states	0.02032	government	0.02928	world	0.01875	energy	0.03902
mexico	0.01832	united	0.02132	states	0.01717	national	0.01534
government	0.01670	states	0.02067	security	0.01710	development	0.01448
united	0.01521	islands	0.01167	soviet	0.01664	space	0.01436
war	0.01059	canal	0.01014	united	0.01491	science	0.01227
congress	0.00951	american	0.00872	nuclear	0.01454	technology	0.01227
country	0.00906	cuba	0.00834	peace	0.01408	oil	0.01178
texas	0.00852	made	0.00747	nations	0.01069	make	0.00994
made	0.00727	general	0.00731	international	0.01024	effort	0.00969
great	0.00611	war	0.00660	america	0.00987	administration	0.00957

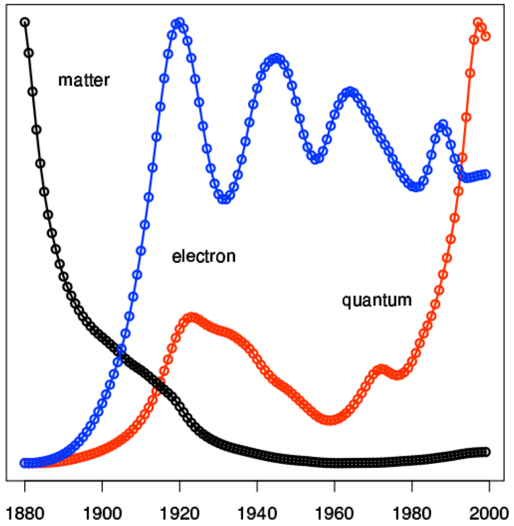
Wang, Xuerui, and Andrew McCallum. "Topics over time: a non-markov continuous-time model of topical trends." *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006.

Dynamic topic models

- Text and time
- Time is continuous

1881	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
force	motion	magnet	force	atom	ray	energy	energy	radiat	electron	electron	electron	state
energy	force	electric	magnet	theory	measure	measure	radiat	energy	energy	energy	atom	energy
motion	magnet	measure	theory	electron	energy	electron	ray	electron	atom	particle	energy	electron
differ	energy	force	electric	energy	theory	light	electron	measure	measure	field	structur	magnet
light	measure	theory	atom	measure	light	atom	measure	ray	radiat	radiat	field	field
measure	differ	system	system	ray	wave	particle	atom	atom	field	model	model	atom
magnet	direct	motion	measure	electr	radiat	ray	particle	field	ray	atom	state	system
direct	line	line	line	line	atom	radiat	two	two	model	two	two	two
matter	result	point	energy	force	electric	point	light	particle	particle	ray	magnet	quantum
result	light	differ	body	value	value	theory	absorpt	observe	magnet	measure	ray	physic

"Atomic Physics"

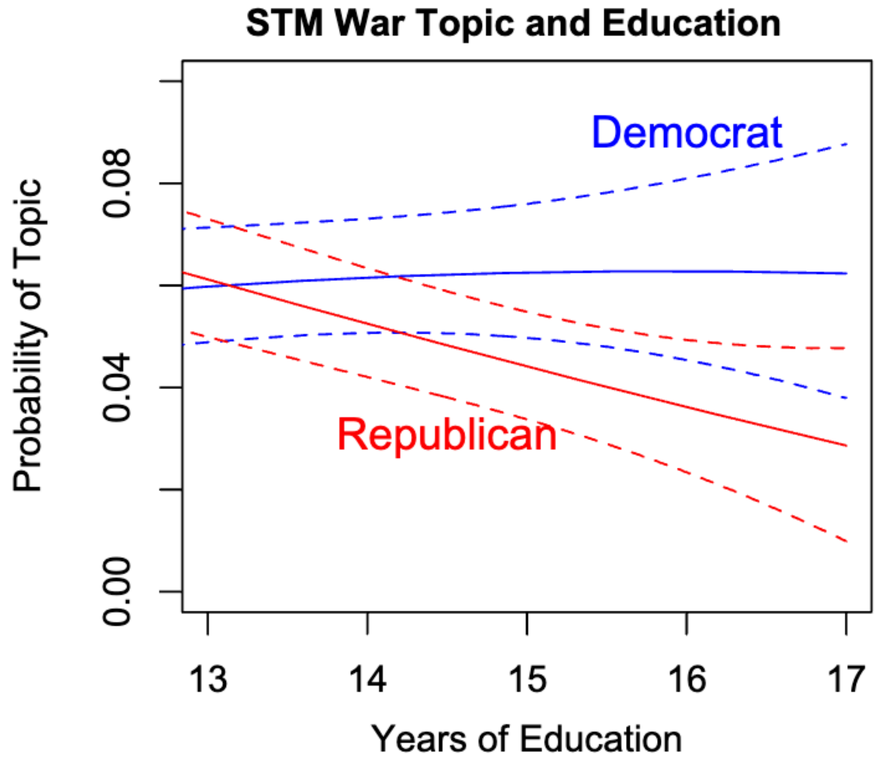


1881 On Matter as a form of Energy
1892 Non-Euclidean Geometry
1900 On Kathode Rays and Some Related Phenomena
1917 ``Keep Your Eye on the Ball''
1920 The Arrangement of Atoms in Some Common Metals
1933 Studies in Nuclear Physics
1943 Aristotle, Newton, Einstein. II
1950 Instrumentation for Radioactivity
1965 Lasers
1975 Particle Physics: Evidence for Magnetic Monopole Obtained
1985 Fermilab Tests its Antiproton Factory
1999 Quantum Computing with Electrons Floating on Liquid Helium

Blei, David M., and John D. Lafferty. "Dynamic topic models." Proceedings of the 23rd international conference on Machine learning. 2006.

Structural topic models

- Model the text and all the associated metadata together
- Useful to do causal analysis

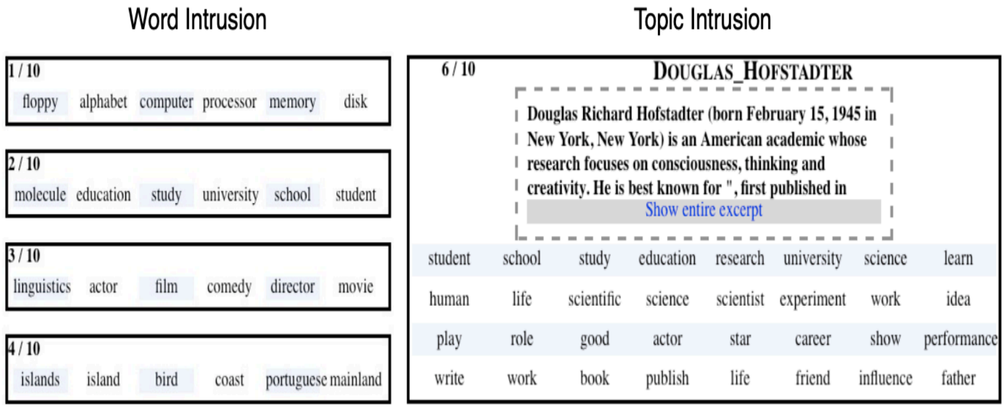


Roberts, Margaret E., et al. "Structural topic models for open-ended survey responses." *American journal of political science* 58.4 (2014): 1064-1082.

Evaluation

Evaluation

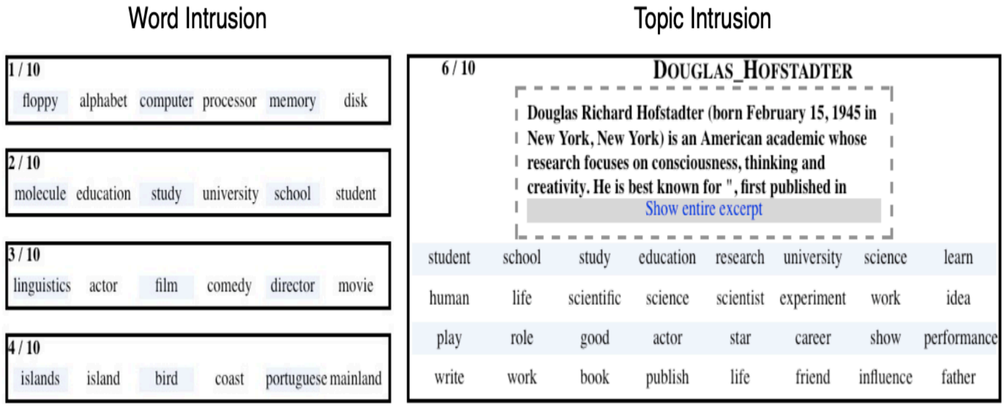
Human evaluation



Chang et. al. 2009

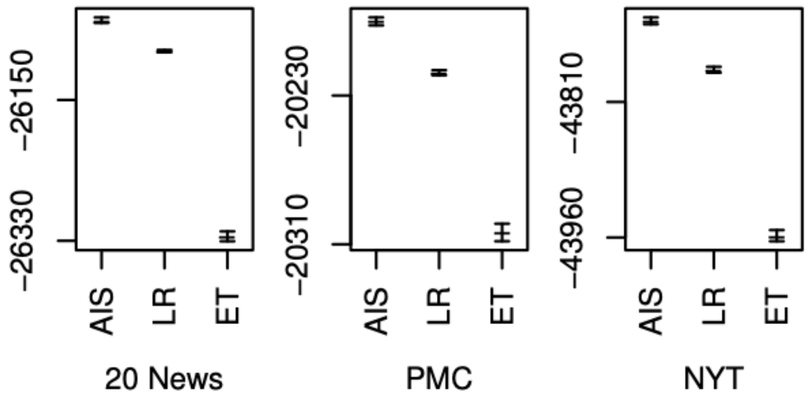
Evaluation

Human evaluation



Chang et. al. 2009

Evaluating generation ability



Wallach et. al. 2009

Topic Model abstraction

- Input: Document collection
- Output: Topics (document-topic and topic-word distributions)

In class

- Topic modeling demo