ILLINOIS INSTITUTE OF TECHNOLOGY

Unsupervised Methods for NLP

CS-585

Natural Language Processing

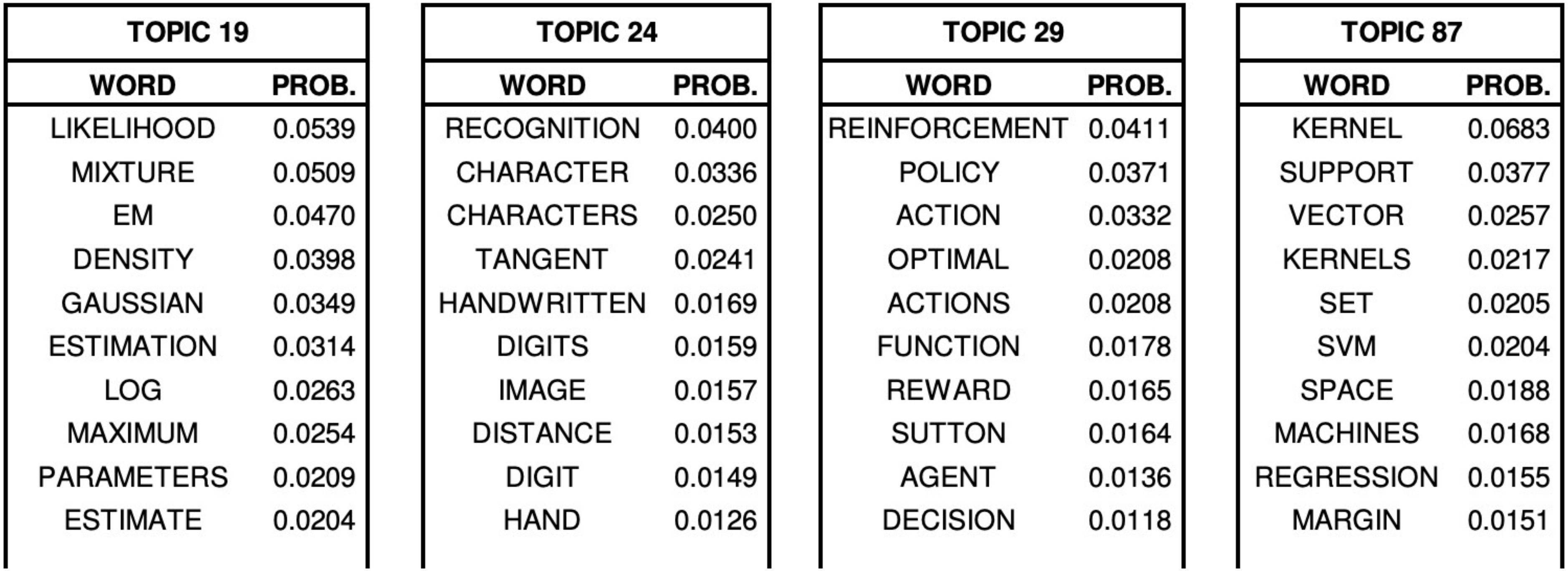
Derrick Higgins

Unsupervised topic modeling

- One common NLP task is clustering documents into "topics" – automatically inferred themes or categories that characterize important aspects of the document collection

- We have seen how to do this already with Naïve Bayes
- *Generative* models like Naïve Bayes allow us to learn parameters by marginalizing (summing over all possible values) of latent variables (like topics)
- Expectation-maximization is one algorithm for this

Goal



TOPIC 19	
WORD	PROB.
LIKELIHOOD	0.0539
MIXTURE	0.0509
EM	0.0470
DENSITY	0.0398
GAUSSIAN	0.0349
ESTIMATION	0.0314
LOG	0.0263
MAXIMUM	0.0254
PARAMETERS	0.0209
ESTIMATE	0.0204

TOPIC 24	
WORD	PROB.
RECOGNITION	0.0400
CHARACTER	0.0336
CHARACTERS	0.0250
TANGENT	0.0241
HANDWRITTEN	0.0169
DIGITS	0.0159
IMAGE	0.0157
DISTANCE	0.0153
DIGIT	0.0149
HAND	0.0126

TOPIC 29	
WORD	PROB.
REINFORCEMENT	0.0411
POLICY	0.0371
ACTION	0.0332
OPTIMAL	0.0208
ACTIONS	0.0208
FUNCTION	0.0178
REWARD	0.0165
SUTTON	0.0164
AGENT	0.0136
DECISION	0.0118

TOPIC 87	
WORD	PROB.
KERNEL	0.0683
SUPPORT	0.0377
VECTOR	0.0257
KERNELS	0.0217
SET	0.0205
SVM	0.0204
SPACE	0.0188
MACHINES	0.0168
REGRESSION	0.0155
MARGIN	0.0151

Rosen-Zvi, Michal, et al. 2004. The Author-Topic Model for Authors and Documents. UAI2004.

NAÏVE BAYES / MIXTURE OF UNIGRAMS

Generative story for Naïve Bayes

For a document:

1. Select a topic from prior distribution $P(T)$
2. For each word to be generated within the document:
    1. Select a word according to $P(W|T)$

Overall probability of corpus is

$$\prod_{d \in D} \sum_{t \in T} P(t) \prod_{w \in d} P(w|t)$$

Note:

1. Each document has a unique topic
2. Each word's probability depends only on the topic

Plate diagrams

- Probabilistic model can also be shown as a *plate diagram*
- Open circles represent latent variables; filled circles represent observed variables
- Arrows represent dependency relationships
- Boxes illustrate variables/components that are duplicated multiple times

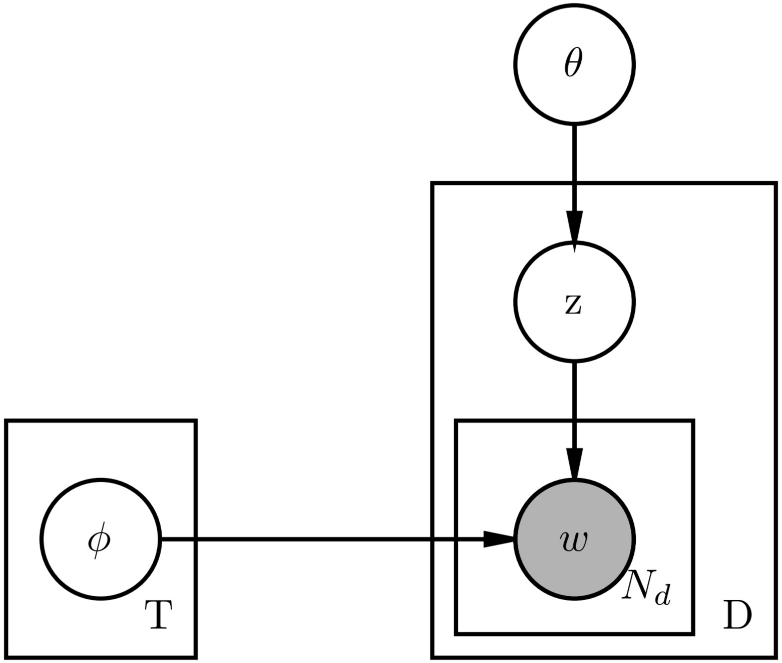
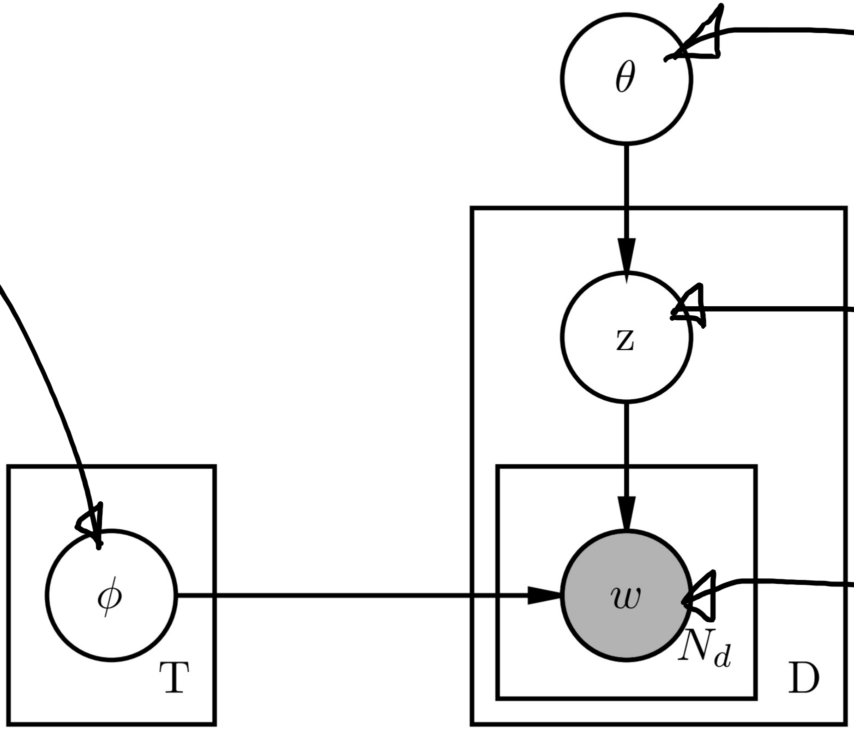


Plate diagrams



One probability distribution $P(W|t)$ for each of K topics

Single probability distribution $P(T)$ across topics

Single topic chosen per document

Each word in document chosen according to $P(W|t)$

LATENT DIRICHLET ALLOCATION (LDA)

Generative story for LDA

For a document:

1. Select a distribution θ over topics
2. For each word to be generated within the document:
   1. Select a topic z associated with the word according to θ
   2. Generate the word according to a distribution ϕ_z for that topic

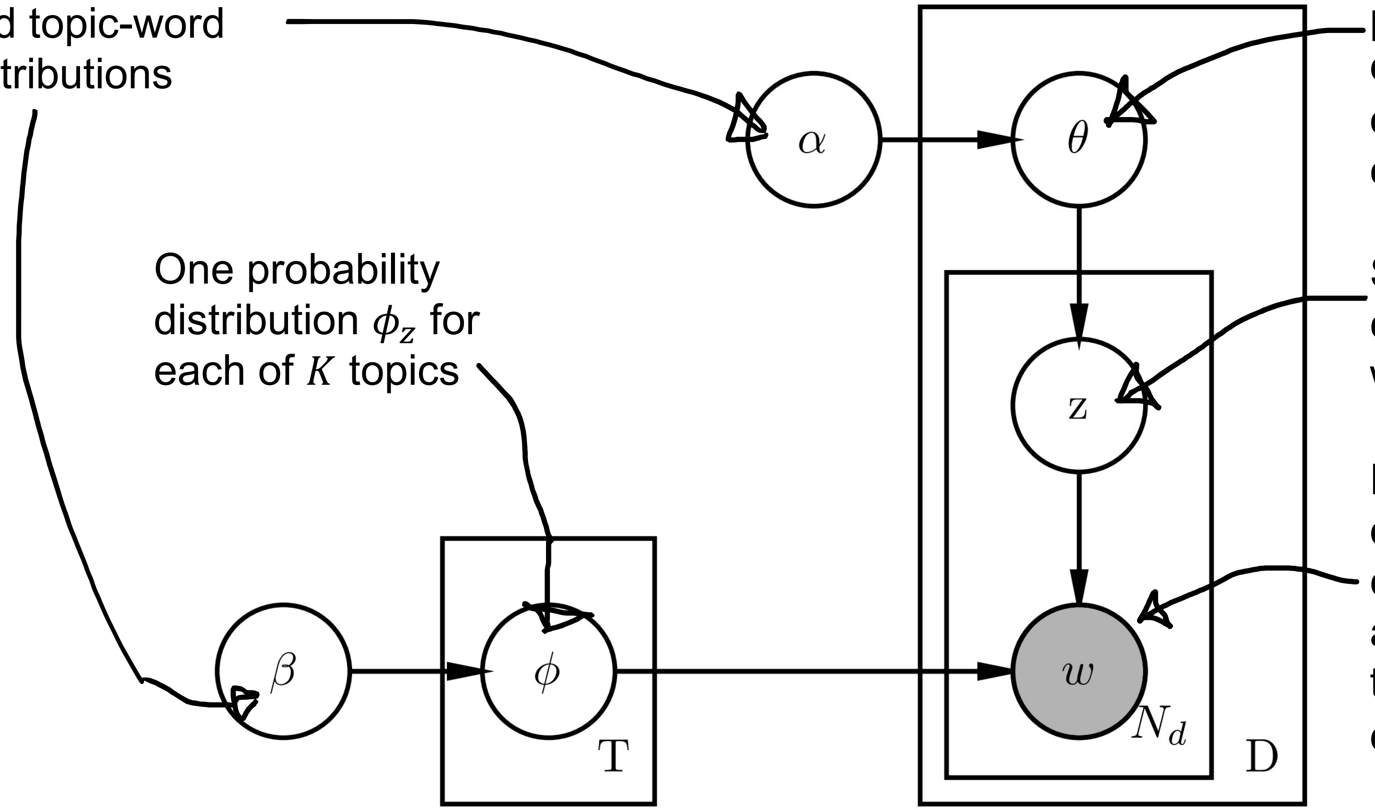
From where?

Note:

1. Each document is associated with a distribution over topics
2. Each word within a document is associated with a single topic
3. Each word's probability depends only on the topic

Plate diagram for LDA

Parameters of Dirichlet priors on document-topic and topic-word distributions



Select a probability distribution θ_d over topics for each document

One probability distribution ϕ_z for each of K topics

Single topic z chosen for each word

Each word in document chosen according to topic-word distribution ϕ_z

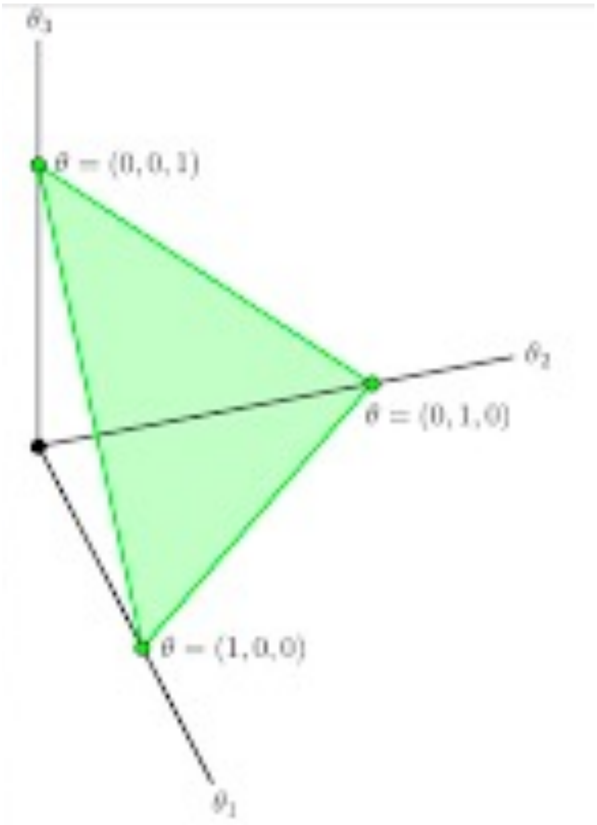
DIRICHLET DISTRIBUTIONS

Dirichlet distributions

- For LDA, we need to choose a topic distribution for each document
- …so we need to be able to draw from a *distribution over distributions*
  - Specifically, a distribution over distributions on the k-simplex
  - (Categorical distributions with k categories)
- The Dirichlet family of distributions is a flexible choice

Dirichlet distributions

- Since a Dirichlet distribution assigns probabilities to distributions of k categories, it is only defined when $\sum_{i=1}^{k} P(c_i) = 1$
- Therefore, its domain is a plane in k-space



http://ml4dummies.blogspot.com/2018/01/the-dirichlet-distribution-is-conjugate.html

Dirichlet distributions

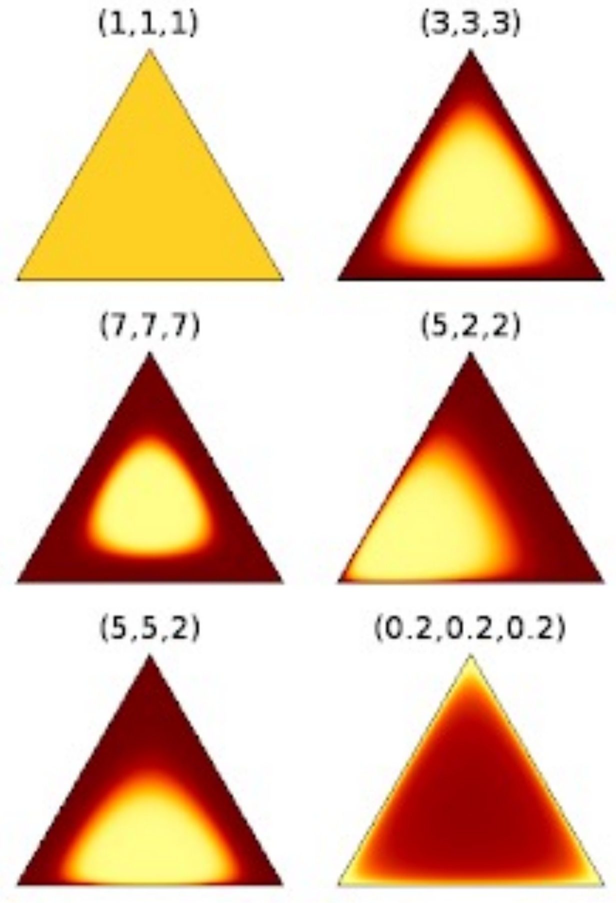
- The probability density function of a Dirichlet distribution is defined as

$$\frac{1}{\mathrm{B}(\alpha)}\prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

- This is related to the Beta distribution (and includes a reference to it)
- But the functional form is not important for us
- We just need to know that it is parameterized by a vector alpha that influences each category's affinity toward the average or extremes of probabilities across topics

Dirichlet distributions

- When $\alpha_i > 1$, $P(c_i)$ will tend toward values closer to $\frac{1}{k}$
- When $\alpha_i < 1$, $P(c_i)$ will tend toward values further from $\frac{1}{k}$ (closer to 1 or 0)
- We can set α to get topic distributions for documents that are more "mixed" or "pure" in terms of topics



http://ml4dummies.blogspot.com/2018/01/the-dirichlet-distribution-is-conjugate.html

LDA TRAINING

Iterative topic model training

- If we knew the topic proportions per document (latent variable), we could estimate parameters of our model (α,β/ϕ)
- If we knew the optimal parameters, we could calculate topic proportions per document
- Candidate for EM!
  - E: Calculate posterior probabilities of z and θ – the topic distributions associated with each document and the topics from which each word in a document was generated
  - M: Estimate α,β/ϕ based on current estimates of z/θ

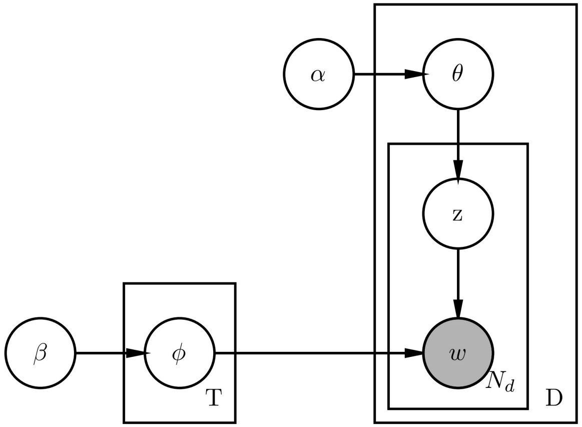
Unfortunately, there is no closed-form solution for this

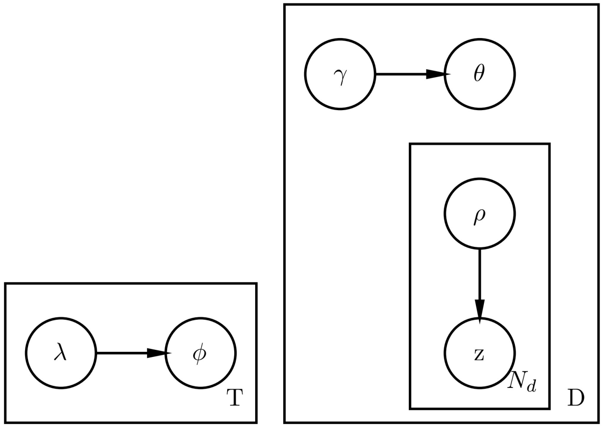
Alternative: Variational Inference

- An alternative to EM is Variational Inference

- Consider alternative (more tractable) probabilistic model q with parameters γ,λ,ρ
- Instead of trying to calculate actual posteriors of hidden variables p, find variational parameters γ,λ,ρ that minimize KL divergence $D_{KL}(p||q(\gamma,\lambda,\rho))$

- There are also Monte Carlo sampling approaches to inference for LDA, but this is generally less efficient





Iterative training using variational inference

Variational EM

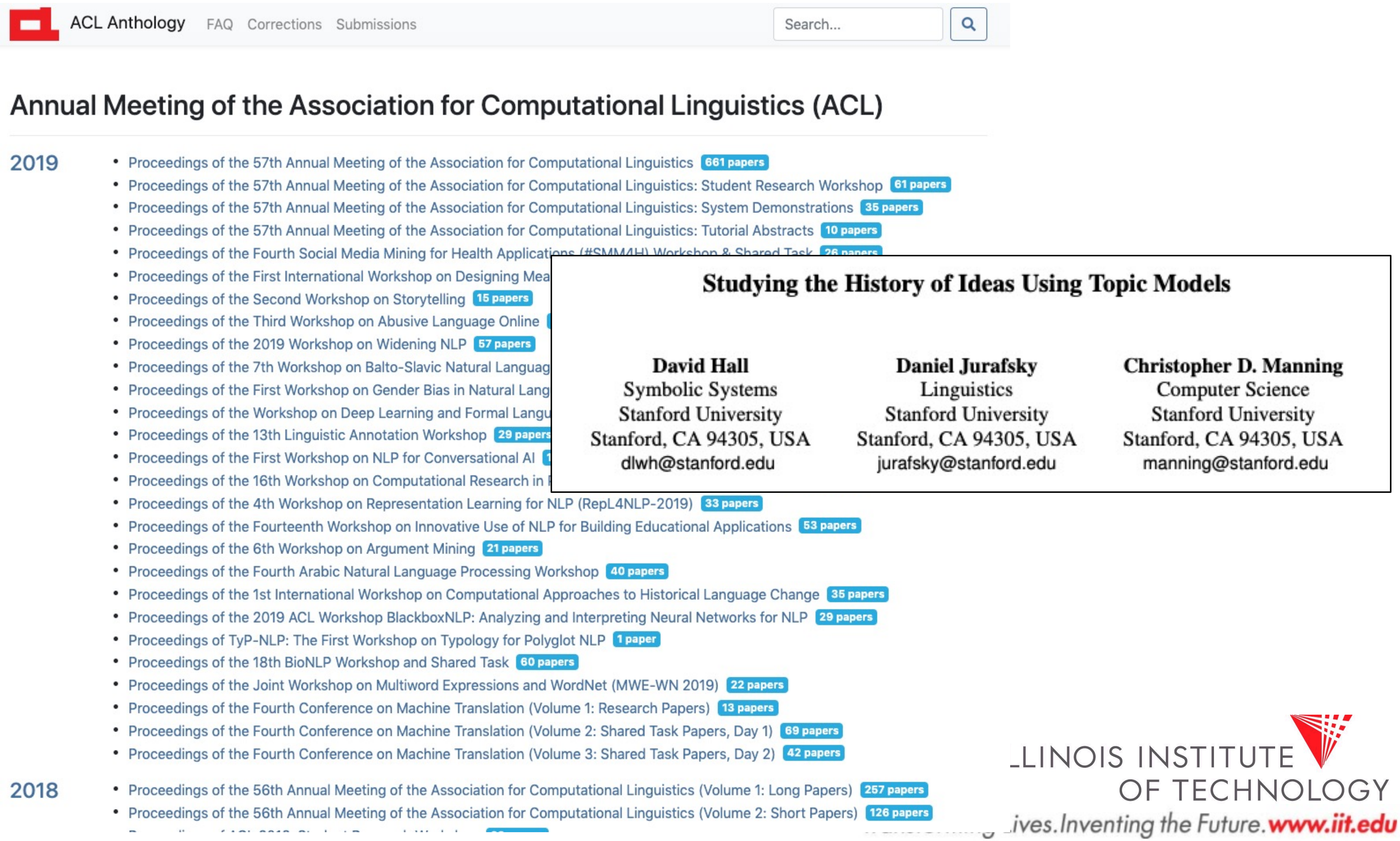
You don’t have to know the details

Just that there is an iterative procedure for inferring parameters based on the data

Maybe take CS583?

- E: Es KL Dive ch docu
- M: S wer bour d using q in place of p (with γ, λ and ρ fixed)

Example: topic models on ACL Anthology



Annual Meeting of the Association for Computational Linguistics (ACL)

2019

- Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics 661 papers
- Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop 61 papers
- Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations 35 papers
- Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts 10 papers
- Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task 26 papers
- Proceedings of the First International Workshop on Designing Mea
- Proceedings of the Second Workshop on Storytelling 15 papers
- Proceedings of the Third Workshop on Abusive Language Online
- Proceedings of the 2019 Workshop on Widening NLP 57 papers
- Proceedings of the 7th Workshop on Balto-Slavic Natural Languag
- Proceedings of the First Workshop on Gender Bias in Natural Lang
- Proceedings of the Workshop on Deep Learning and Formal Langu
- Proceedings of the 13th Linguistic Annotation Workshop 29 papers
- Proceedings of the First Workshop on NLP for Conversational AI
- Proceedings of the 16th Workshop on Computational Research in I
- Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019) 33 papers
- Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications 53 papers
- Proceedings of the 6th Workshop on Argument Mining 21 papers
- Proceedings of the Fourth Arabic Natural Language Processing Workshop 40 papers
- Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change 35 papers
- Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP 29 papers
- Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP 1 paper
- Proceedings of the 18th BioNLP Workshop and Shared Task 60 papers
- Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019) 22 papers
- Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers) 13 papers
- Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1) 69 papers
- Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2) 42 papers

2018

- Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 257 papers
- Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 126 papers

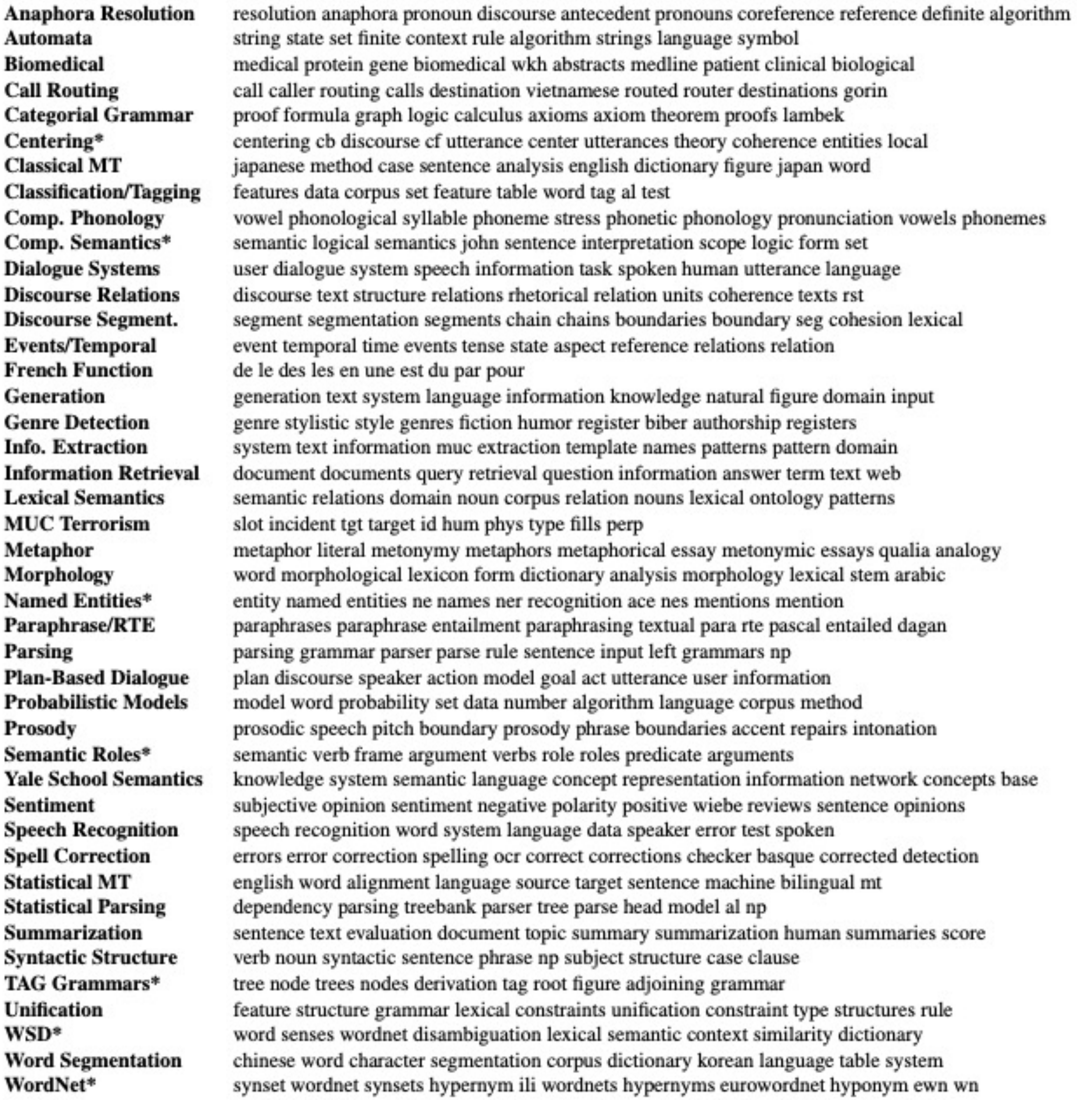
Studying the History of Ideas Using Topic Models

David Hall
Symbolic Systems
Stanford University
Stanford, CA 94305, USA
dlwh@stanford.edu

Daniel Jurafsky
Linguistics
Stanford University
Stanford, CA 94305, USA
jurafsky@stanford.edu

Christopher D. Manning
Computer Science
Stanford University
Stanford, CA 94305, USA
manning@stanford.edu

Example: topic models on ACL Anthology



Anaphora Resolution	resolution anaphora pronoun discourse antecedent pronouns coreference reference definite algorithm
Automata	string state set finite context rule algorithm strings language symbol
Biomedical	medical protein gene biomedical wkh abstracts medline patient clinical biological
Call Routing	call caller routing calls destination vietnamese routed router destinations gorin
Categorial Grammar	proof formula graph logic calculus axioms axiom theorem proofs lambek
Centering*	centering cb discourse cf utterance center utterances theory coherence entities local
Classical MT	japanese method case sentence analysis english dictionary figure japan word
Classification/Tagging	features data corpus set feature table word tag al test
Comp. Phonology	vowel phonological syllable phoneme stress phonetic phonology pronunciation vowels phonemes
Comp. Semantics*	semantic logical semantics john sentence interpretation scope logic form set
Dialogue Systems	user dialogue system speech information task spoken human utterance language
Discourse Relations	discourse text structure relations rhetorical relation units coherence texts rst
Discourse Segment.	segment segmentation segments chain chains boundaries boundary seg cohesion lexical
Events/Temporal	event temporal time events tense state aspect reference relations relation
French Function	de le des les en une est du par pour
Generation	generation text system language information knowledge natural figure domain input
Genre Detection	genre stylistic style genres fiction humor register biber authorship registers
Info. Extraction	system text information muc extraction template names patterns pattern domain
Information Retrieval	document documents query retrieval question information answer term text web
Lexical Semantics	semantic relations domain noun corpus relation nouns lexical ontology patterns
MUC Terrorism	slot incident tgt target id hum phys type fills perp
Metaphor	metaphor literal metonymy metaphors metaphorical essay metonymic essays qualia analogy
Morphology	word morphological lexicon form dictionary analysis morphology lexical stem arabic
Named Entities*	entity named entities ne names ner recognition ace nes mentions mention
Paraphrase/RTE	paraphrases paraphrase entailment paraphrasing textual para rte pascal entailed dagan
Parsing	parsing grammar parser parse rule sentence input left grammars np
Plan-Based Dialogue	plan discourse speaker action model goal act utterance user information
Probabilistic Models	model word probability set data number algorithm language corpus method
Prosody	prosodic speech pitch boundary prosody phrase boundaries accent repairs intonation
Semantic Roles*	semantic verb frame argument verbs role roles predicate arguments
Yale School Semantics	knowledge system semantic language concept representation information network concepts base
Sentiment	subjective opinion sentiment negative polarity positive wiebe reviews sentence opinions
Speech Recognition	speech recognition word system language data speaker error test spoken
Spell Correction	errors error correction spelling ocr correct corrections checker basque corrected detection
Statistical MT	english word alignment language source target sentence machine bilingual mt
Statistical Parsing	dependency parsing treebank parser tree parse head model al np
Summarization	sentence text evaluation document topic summary summarization human summaries score
Syntactic Structure	verb noun syntactic sentence phrase np subject structure case clause
TAG Grammars*	tree node trees nodes derivation tag root figure adjoining grammar
Unification	feature structure grammar lexical constraints unification constraint type structures rule
WSD*	word senses wordnet disambiguation lexical semantic context similarity dictionary
Word Segmentation	chinese word character segmentation corpus dictionary korean language table system
WordNet*	synset wordnet synsets hypernym ili wordnets hypernyms eurowordnet hyponym ewn wn

TOPIC VISUALIZATION

Which words are important in the topic model?

Q1: Which words are most important in determining the topic distribution for a document?

- *Saliency*

Q2: Which words are most representative of a given topic?

- *Relevance*

Saliency

- Salient words (Chuang, 2012) are *frequent* words that contribute a great deal of *information* regarding the topic of a document

- Word frequency: $P(w)$
- Information on topic of a document (distinctiveness): $\sum_{t \in T} P(t|w) \log \frac{P(t|w)}{P(t)}$
  - K-L divergence between $P(t|w)$ – probability that a document containing w has topic t – and $P(t)$ – probability that any word will be generated by topic t

- Saliency: $P(w) \times \sum_{t \in T} P(t|w) \log \frac{P(t|w)}{P(t)}$

Chuang, J. et al. 2012. Termite: Visualization Techniques for Assessing Textual Topic Models. AVI 2012.

Relevance

- What words are most representative of a topic (how can I inspect and "understand" it)?
- We could use $P(w_i|t_k)$ from the LDA model directly (ϕ_{ik})
  - But some frequent words may have high $P(w_i|t_k)$ just because of their high general frequency
- We could use $\frac{P(wi|t_k)}{P(wi)}$ to capture for word-topic association while controlling for frequency
  - But rare words may swamp the results
- Solution: introduce a parameter λ to govern tradeoff between frequency and association with the topic

$$Relevance(w_i|t_k) = \lambda \log P(w_i|t_k) + (1-\lambda) \log \frac{P(wi|t_k)}{P(wi)}$$

TOPIC EVALUATION

Evaluating topic models

What makes a topic “good”?

Its words are strongly associated with one another

I.e., the topic is “coherent”

What makes a topic model “good”?

Its topics are coherent

Topic coherence

Coherence: different ways of modeling it, but generally based on pairwise relationships between words within a topic

- See Röder et al. (2015). *Exploring the Space of Topic Coherence Measures*.

One method: C_{UCI} (Newman, 2010)

- Word pairs within topics should have high pointwise mutual information (PMI)
- I.e., they should show up more frequently together in a document than expected by chance

UCI Coherence

[Stands for UC-Irvine]

Pointwise mutual information (PMI)

- Let $P(w_i)$ be the probability of w$_i$ occurring in a document, and $P(w_i, w_j)$ be the probability of w_i and w_j occurring together in a document.
- $PMI(w_i, w_j) \overset{\text{def}}{=} \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$
- If w_i and w_j occur independently at random, then $P(w_i, w_j) = P(w_i)P(w_j)$ and $PMI(w_i, w_j) = 0$
- If w_i and w_j are associated, then $P(w_i, w_j) > P(w_i)P(w_j)$ and $PMI(w_i, w_j) > 0$

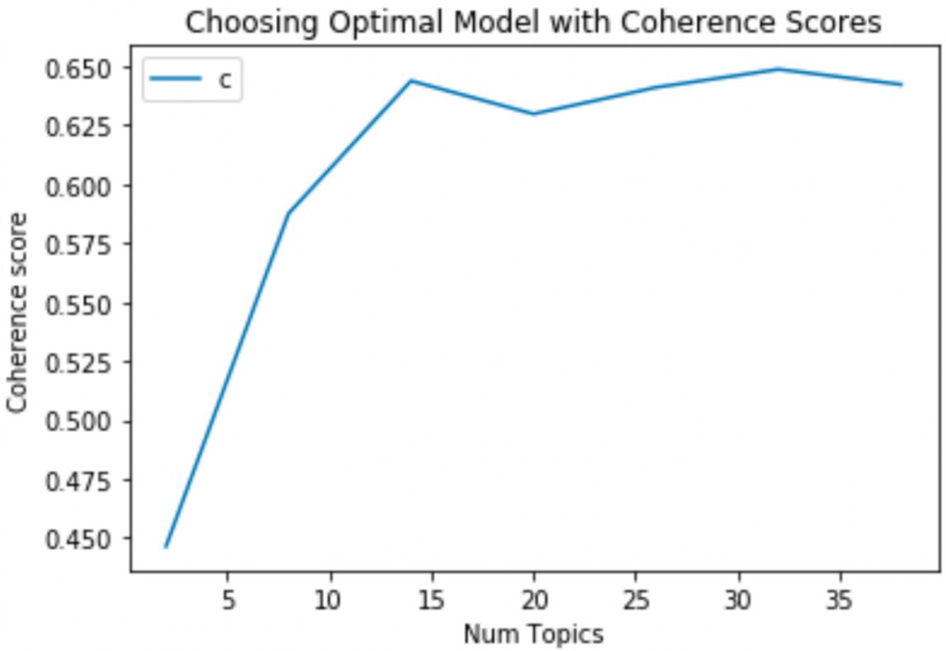
UCI Coherence

- UCI Coherence is the average PMI of all word pairs in the top N words of a topic

$$C_{UCI} = \frac{2}{N(N-1)} \sum_{i \in [1..N], j \in [1..N], i \neq j} PMI(w_i, w_j)$$

Using coherence to determine number of topics

- Coherence of topic model as a whole is average coherence of its topics
- Select number of topics *k* with maximum coherence value (or where coherence plateaus)



https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#17howtofindtheoptimalnumberoftopicsforlda

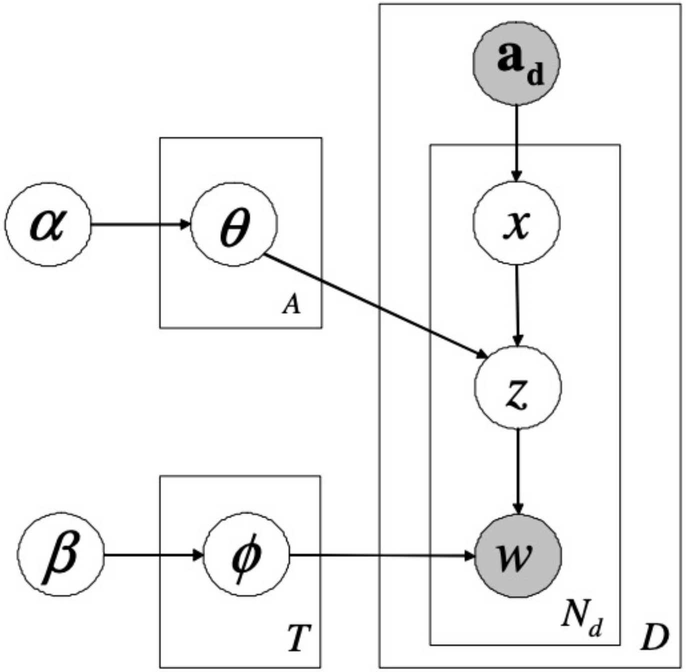
LDA EXTENSIONS

LDA Extensions

- LDA is an extension of the word unigram model that incorporates a more complex generative story
  - Distribution over topics for each document
  - Dirichlet prior over topic distributions
- There are extended versions of the LDA framework that involve even more complex generative stories
  - Instead of topic distributions generated independently for each document, these distributions may depend on attributes of the document

Author-topic model

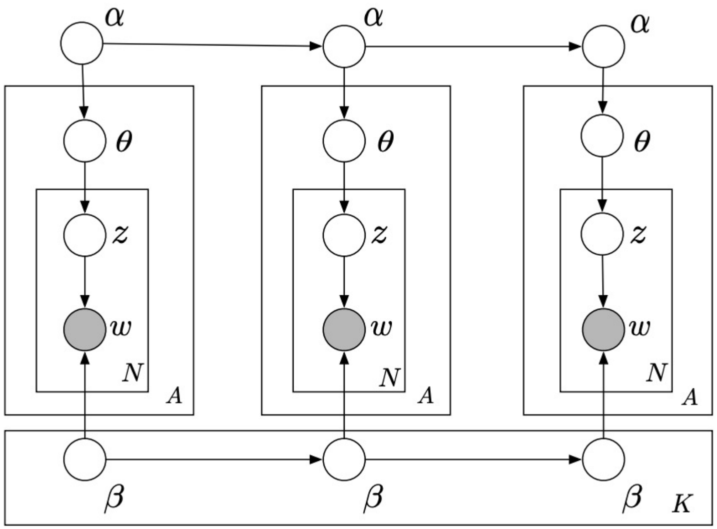
- Author(s) of a document influence topic distribution
- For a given document, select an author set $\mathbf{a}_d$
- For each word to be generated
  - Select an author **x**
  - Select a topic **z** depending on **x**'s topic distribution $\boldsymbol{\theta}$
  - Select a word from the distribution for the chosen topic



Dynamic topic model

- Timestamp of a document influences topic distribution

- Similar generative story to LDA, but
  - Document-topic parameters **α** and topic-word parameters **β** change over time
  - Have to model dynamics of this system via distributional constraints on temporally adjacent values


α
α
α
θ
θ
θ
z
z
z
w
w
w
N
N
N
A
A
A
β
β
β
K