Cluster-Based Topic Modeling

Hanna M. Wallach (joint work with David Mimno, Mark Dredze, Andrew McCallum)

University of Massachusetts Amherst

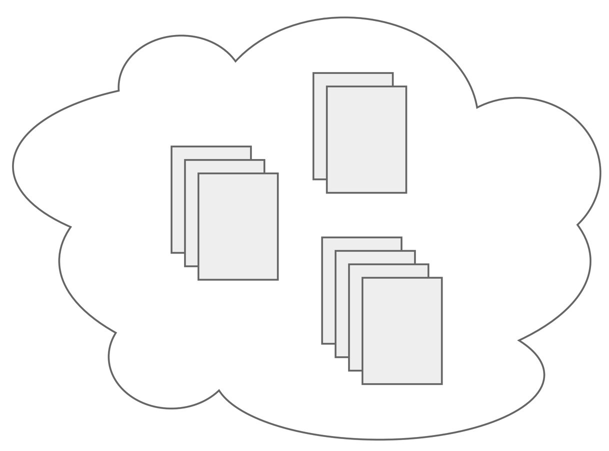
wallach@cs.umass.edu

Abstract

A nonparametric Bayesian model that clusters documents by topic:

- Robust to variations in terminology
- Automatically infers the number of clusters
- Cluster and topic inference are performed simultaneously

Structured Document Collections



Many document collections exhibit document groupings:

- e.g., papers from a single conference on closely related topics

Document Groupings

Information about these groupings is useful for:

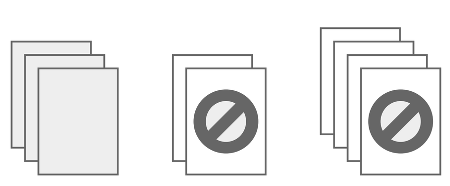
- Navigating and visualizing large corpora
- Learning about relationships between topics
- Learning about relationships between authors and topics
- Performing coarse-grained corpus-based analyses
- Detecting granularity of topics

but...

▹ Document groupings are often unobserved

Applications

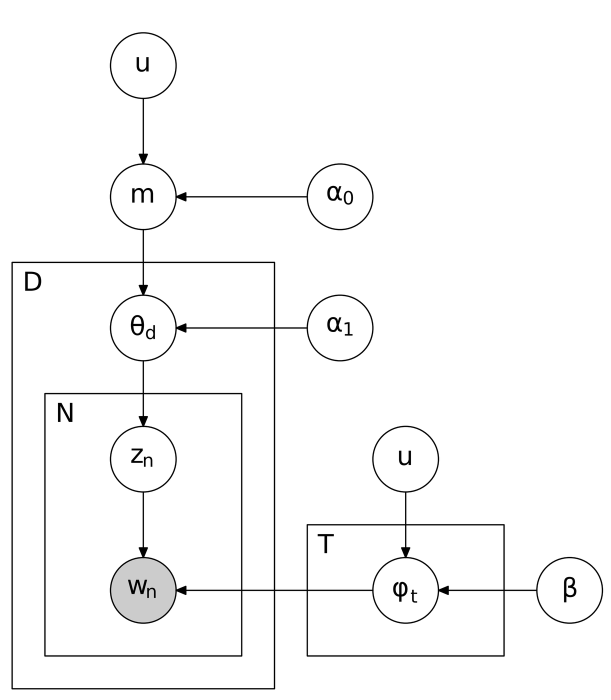
Document groups (clusters) can be used to guide navigation of corpora and to select relevant subsets of documents



The set of topics associated with each cluster can be used for:

- Topic-based navigation, e.g., which topics co-occur with this one?
- Identification of more and less specific topics, e.g., if a topic occurs in all clusters it is probably a very general topic

Background: LDA (Blei et al., 2003)



Topics and words are drawn from multinomial distributions:

$$z_n \sim \mathrm{Mult}(\boldsymbol{\theta}_{d_n})$$
$$w_n \sim \mathrm{Mult}(\boldsymbol{\phi}_{z_n})$$

Asymmetric hierarchical Dirichlet prior over $\boldsymbol{\theta}_d$:

$$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha_1, \boldsymbol{m})$$
$$\boldsymbol{m} \sim \mathrm{Dir}(\alpha_0, \boldsymbol{u})$$

Symmetric Dirichlet prior over $\boldsymbol{\phi}_t$:

$$\boldsymbol{\phi}_t \sim \mathrm{Dir}(\beta, \boldsymbol{u})$$

Given observed documents (i.e., words), latent topic assignments can be inferred using Gibbs sampling or variational inference.

LDA: Predictive Distributions

Integrate over probability vectors to obtain predictive distributions, e.g., for the predictive probability of topic t in document d:

$$P(t \mid d, \boldsymbol{z}, \alpha_1, \boldsymbol{m}) = \int \theta_{t|d} P(\boldsymbol{\theta}_d \mid \boldsymbol{z}, \alpha_1, \boldsymbol{m}) \, \mathrm{d}^T \boldsymbol{\theta}_d = \frac{N_{t|d} + \alpha_1 m_t}{\sum_t N_{t|d} + \alpha_1}$$

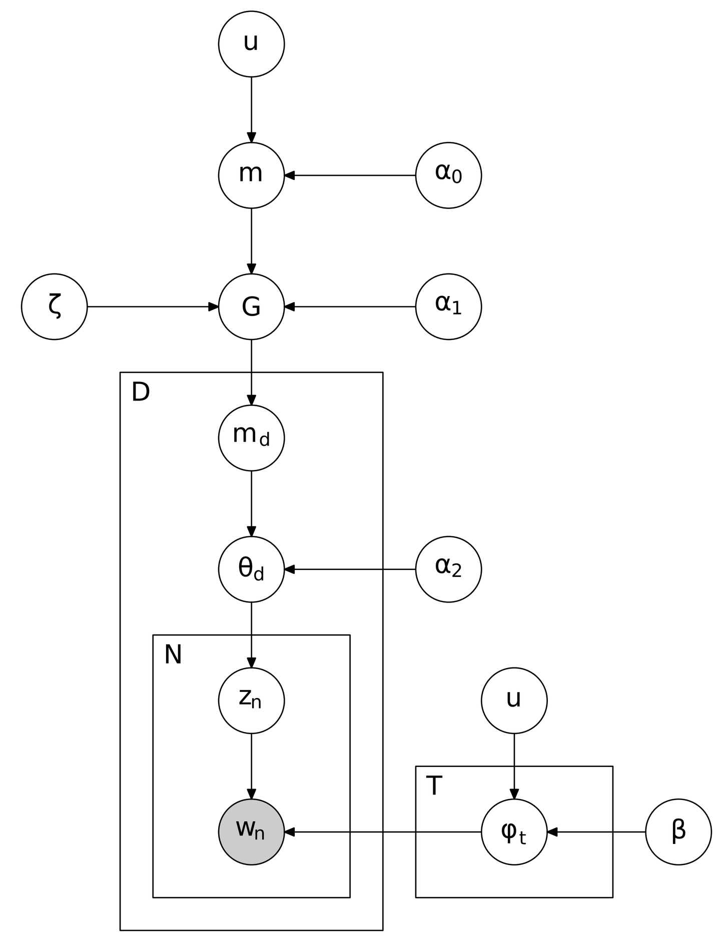
and so

$$P(t \mid d, \boldsymbol{z}, \alpha_1, \alpha_0, \boldsymbol{u}) = \int P(t \mid d, \boldsymbol{z}, \alpha_1, \boldsymbol{m}) P(\boldsymbol{m} \mid \boldsymbol{z}, \alpha_0, \boldsymbol{u}) \, \mathrm{d}^T \boldsymbol{m} = \frac{N_{t|d} + \alpha_1 \frac{\hat{N}_t + \alpha_0 u_t}{\sum_t \hat{N}_t + \alpha_0}}{\sum_t N_{t|d} + \alpha_1}$$

Count $N_{t|d}$ is always equal to the number of times topic t has been used in document d. However, count $\hat{N}_t$ can either be

- the total number of times topic t has been used in the corpus,
- the number of documents in which t has been used,
- or somewhere between the two.

A Cluster-Based Topic Model



Model differs from LDA only in the prior over $\boldsymbol{\theta}_d$:

$$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha_2, \boldsymbol{m}_d)$$

where

$$\boldsymbol{m}_d \sim G$$
$$G \sim \mathrm{DP}(\zeta, G_0)$$

The distribution over topics for each document $\boldsymbol{\theta}_d$ is drawn from a *document-specific* Dirichlet distribution with base measure $\boldsymbol{m}_d$. This base measure is itself drawn from G, which is a draw from a Dirichlet Process with base distribution G_0 and concentration parameter ζ. Using the stick-breaking construction, this choice of prior means that

$$G(\boldsymbol{m}_d) = \sum_{c=1}^{\infty} \pi_c \, \delta_{\boldsymbol{m}_c}(\boldsymbol{m}_d)$$

where

$$\boldsymbol{m}_c \sim G_0$$

Here, G_0 is an asymmetric hierarchical Dirichlet distribution. This choice of G_0 ensures that the only effect of the Dirichlet process on the prior over $\boldsymbol{\theta}_d$ is to allow a variable number of document clusters.

Given observed documents (i.e., words) latent topic and cluster assignments can be inferred using Gibbs sampling by alternating between sampling topics given clusters and clusters given topics.

Predictive Distributions

The predictive probability of selecting cluster c is:

$$P(c \mid \boldsymbol{c}, \zeta) \propto \begin{cases} N_c & \text{c is an existing cluster} \\ \zeta & \text{c is a new cluster} \end{cases}$$

The predictive probability of selecting topic t in document d is:

$$P(t \mid d, c_d{=}c, \boldsymbol{z}, \boldsymbol{c}, \alpha_2, \alpha_1, \alpha_0, \boldsymbol{u}) = \frac{N_{t|d} + \alpha_2 \frac{\hat{N}_{t|c} + \alpha_1 \frac{\hat{N}_t + \alpha_0 u_t}{\sum_t \hat{N}_t + \alpha_0}}{\sum_t \hat{N}_{t|c} + \alpha_1}}{\sum_t N_{t|d} + \alpha_2}$$

Count $N_{t|d}$ is always the number of times topic t has been used in document d (regardless of cluster). However, count $\hat{N}_{t|c}$ can be

- the total number of times topic t has been used in cluster c,
- the number of documents in cluster c in which t has been used,
- or somewhere between the two.

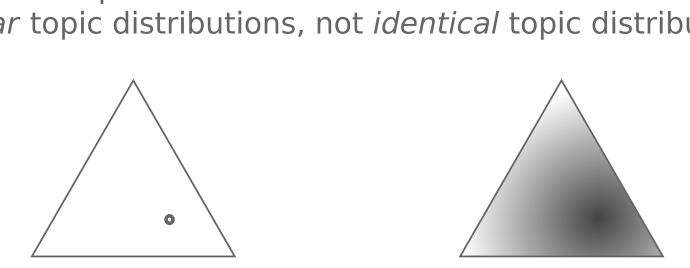
Similarly, count $\hat{N}_t$ can be

- the total number of times topic t has been used in the corpus,
- the number of documents in which t has been used,
- or somewhere between the two.

As $\alpha_1 \to \infty$, this predictive probability tends towards that of LDA.

Related: Dirichlet-Enhanced LSA (Yu et al., 2005)

In Dirichlet-enhanced LSA, cluster-specific topic distributions are used *without modification* as document-specific topic distributions. Here, document-specific topic distributions are allowed to vary around the cluster-specific topic distribution: documents in the same cluster have *similar* topic distributions, not *identical* topic distributions.



Dirichlet-enhanced LSA — new cluster-based model

Dirichlet-enhanced LSA effectively ignores document-specific counts and relies only on cluster- and corpus-specific counts. The cluster-based topic model in this poster can infer the extent to which document-specific counts influence the selection of future topics.

LDA	DELSA
no cluster-specific counts	no document-specific counts

new model
← interpolates between the two →

Experimental Setup

20 years of NIPS proceedings:

- Training data: papers from 1997–2003 (2,325 papers)
- Test data: papers from 2004–2006 (614 papers)

Three baseline models: latent Dirichlet allocation, Dirichlet-enhanced LSA, a simple word-based Dirichlet process mixture model.

Results: Perplexity

Perplexity of test data:

$$\text{Perp.} = \exp\left(-\frac{\log_2 P(\boldsymbol{w}^{\text{test}} \mid \boldsymbol{w}^{\text{train}})}{N^{\text{test}}}\right),$$

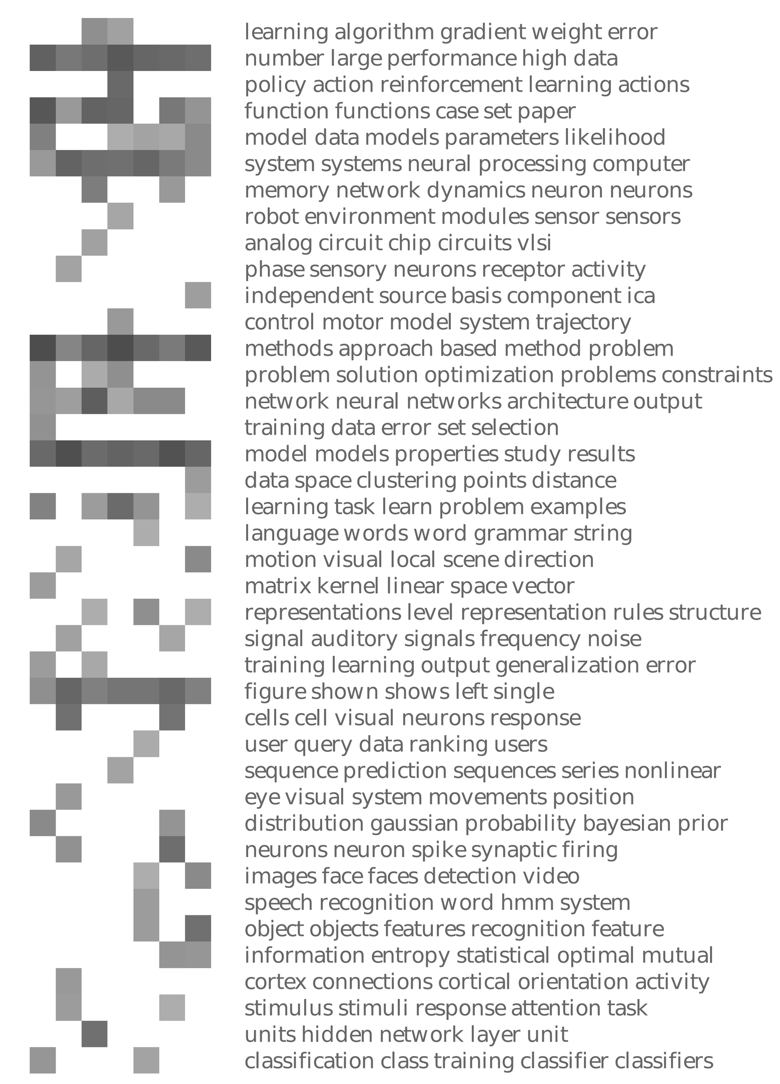
where N^{test} is the number of tokens in the test data. Lower perplexity = better model. $P(\boldsymbol{w}^{\text{test}} \mid \boldsymbol{w}^{\text{train}})$ can be approximated using a variant of the "harmonic mean" method of Griffiths and Steyvers (2004), which simulates marginalization over topic/cluster assignments.

model	perplexity
word-based DPMM	1489
latent Dirichlet allocation	333
new model	**321**

Inferred Topics and Clusters

The clusters and topics inferred by Dirichlet-enhanced LSA were extremely hard-to-interpret and did not obviously correspond to coherent groups. They are therefore not discussed further. The word-based DPMM inferred 12 clusters. Four clusters assign high probability to a few specialized words, making them relatively easy to interpret. The top words for the other clusters are quite general.

The new cluster-based model inferred 7 clusters. Although a small number of topics appear in every cluster, all but one of the clusters assign high probabilities to at least two specialized topics:



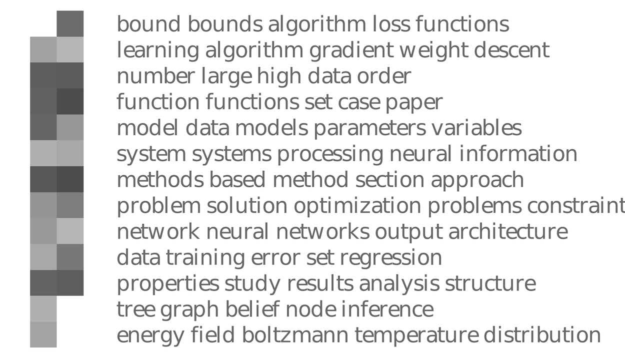
Interpreting the Clusters

The five most frequently used topics for the top four clusters:

function	model	learning	distribution	**training**
functions	data	task	gaussian	**data**
case	models	learn	probability	**error**
set	parameters	problem	bayesian	**set**
paper	likelihood	examples	prior	**selection**
section	mixture	algorithm	noise	**risk**
defined	variables	set	posterior	**regression**
assume	density	learned	random	**regularisation**
vector	probability	training	density	**generalisation**
general	estimation	tasks	estimate	**parameters**
cells	neurons	**eye**	**cortex**	function
cell	neuron	**visual**	**connections**	functions
visual	spike	**system**	**cortical**	case
neurons	synaptic	**movements**	**orientation**	set
response	firing	**position**	**activity**	paper
stimulus	spikes	**velocity**	**layer**	section
receptive	membrane	**vor**	**lateral**	defined
field	potential	**model**	**development**	assume
responses	model	**target**	**dominance**	vector
cortex	neuronal	**retina**	**patterns**	general
network	function	**units**	memory	learning
neural	functions	**hidden**	network	algorithm
networks	case	**network**	dynamics	gradient
architecture	set	**layer**	neuron	weight
output	paper	**unit**	neurons	error
weights	section	**output**	networks	descent
feedforward	defined	**weights**	associative	function
trained	assume	**activation**	model	convergence
recurrent	vector	**networks**	hopfield	algorithms
training	general	**net**	patterns	stochastic
function	**policy**	learning	problem	**control**
functions	**action**	task	solution	**motor**
case	**reinforcement**	learn	optimisation	**model**
set	**learning**	problem	problems	**system**
paper	**actions**	examples	constraints	**trajectory**
section	**optimal**	algorithm	function	**controller**
defined	**agent**	set	point	**feedback**
assume	**states**	learned	solution	**movement**
vector	**reward**	training	constraint	**arm**
general	**decision**	tasks	objective	**dynamics**

Clustering by Topic and Author

Can instead cluster documents by author and topic. Many more clusters are inferred: papers that use similar topics but are by different groups of people are unlikely to be clustered together.



Cluster 1 (219 Documents)	Cluster 2 (183 Documents)
G. Hinton	J. Shawe-Taylor
M. Jordan	A. Smola
C. Williams	B. Scholkopf
Z. Ghahramani	P. Bartlett
M. Opper	R. Williamson
T. Jaakkola	V. Vapnik
D. Saad	Y. Bengio
P. Sollich	R. Herbrich
C. Bishop	T. Graepel
D. Barber	J. Weston

Future Directions

Other priors:

- e.g., avoid "rich-get-richer" cluster usage by using a uniform process prior (Dicker and Jensen, 2008) instead of a Dirichlet process prior. Advantage: documents are assigned to clusters solely on the basis of "goodness-of-fit". Disadvantage: non-exchangeable, so inference of cluster assignments is slower.