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 set
- Identification of more and less specific topics, e.g., a topic occurs in all clusters if it is a very general topic.

Background: LDA (Blei et al., 2003)

Topics and words are drawn from multinomial distributions:
- $z_n \sim \text{Mult}(\theta_n)$
- $w_n \sim \text{Mult}(\phi_n)

Asymmetric hierarchical Dirichlet prior over $\theta$:
- $\theta_n \sim \text{Dit}(\alpha, \beta)$

Symmetric Dirichlet prior over $\phi$:
- $\phi \sim \text{Dir}(\alpha_n)$

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 $k$ in document $d$ at

$$P(k|d) = \sum_z P(k|z)P(z|d)$$

and so

$$P(k|d, z) = \frac{N_{dk} + \alpha_k}{\sum_{k'} N_{dk'} + \alpha_k}$$

Count $N_{dk}$ is always equal to the number of times topic $k$ has been used in document $d$. However, count $N_{dk}$ can either be:
- the total number of times topic $k$ has been used in the corpus,
- the number of documents in which it has been used,
- somewhere between the two.

As $\alpha \to 0$, this predictive probability tends towards that of LDA.

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

In Dirichlet-enhanced LSA, document-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.

A Cluster-Based Topic Model

Model differs from LDA only in the prior over $\theta$

$$\theta_n \sim \text{Dir}(\alpha_n, \beta_n)$$

where

$$\text{Dir}(\alpha_n, \beta_n) \propto \beta_n^{|\alpha_n|}$$

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 DUMM inferred 32 clusters, four of which assign high probability to a few specialized topics.

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

Interpreting the Clusters

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

Clustering by Topic and Author

Can cluster documents by author and topic. Many more clusters are inferred if topics are used in clusters that are more specific to different groups of people or likely to be clustered together:

Future Directions

Other priors:
- $\alpha \sim \text{Dir}_{\epsilon}$ or $\text{Log-Gamma}$ cluster usage by using a uniform process prior (Dyk and Jensen, 2008) instead of a Dirichlet process prior, which further document assignments solely on the basis of goodness-of-fit. Disadvantage: non-exchangeable, so inference of cluster assignments is hard.