Latent Dirichlet Allocation

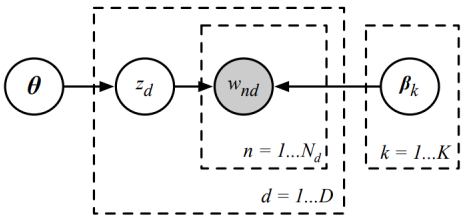
David Mareček

October 22, 2024


Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Many of the slides in this presentation were taken from the presentations of Carl Edward Rasmussen (University of Cambridge)

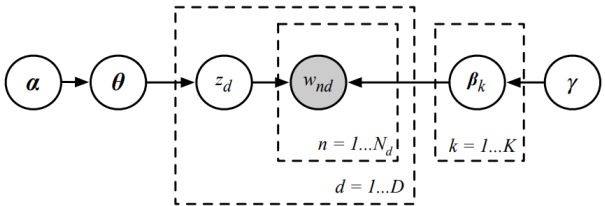
Mixture of Categoricals Model



$$z_d \sim Cat(\vec{\theta})$$
$$w_{nd}|z_d \sim Cat(\vec{\beta}_{z_d})$$

With the Expectation-Maximization algorithm we have essentially estimated $\vec{\theta}$ and $\vec{\beta}$ by maximum likelihood.

Bayesian Mixture of Categoricals Model



$$z_d \sim Cat(\vec{\theta})$$
$$\vec{\theta} \sim Dir(\vec{\alpha})$$
$$w_{nd}|z_d, \vec{\beta} \sim Cat(\vec{\beta}_{z_d})$$
$$\vec{\beta}_k \sim Dir(\vec{\gamma})$$

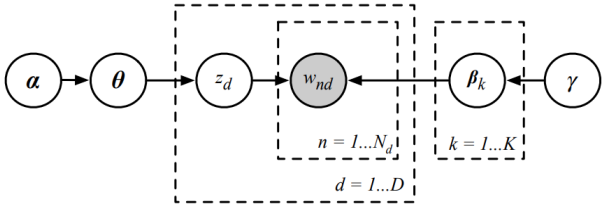
An alternative, Bayesian treatment infers these parameters starting from priors, e.g.:

- $\vec{\theta} \sim Dir(\vec{\alpha})$ is a symmetric Dirichlet over category probabilities,
- $\vec{\beta}_k \sim Dir(\vec{\gamma})$ are symmetric Dirichlets over vocabulary probabilities.

What is different?

- We no longer want to compute a point estimate of $\vec{\theta}$ and $\vec{\beta}$.
- We are now interested in computing posterior distributions.

Limitations of the mixture of categoricals model



$$z_d \sim Cat(\vec{\theta})$$
$$\vec{\theta} \sim Dir(\vec{\alpha})$$
$$w_{nd}|z_d, \vec{\beta} \sim Cat(\vec{\beta}_{z_d})$$
$$\vec{\beta}_k \sim Dir(\vec{\gamma})$$

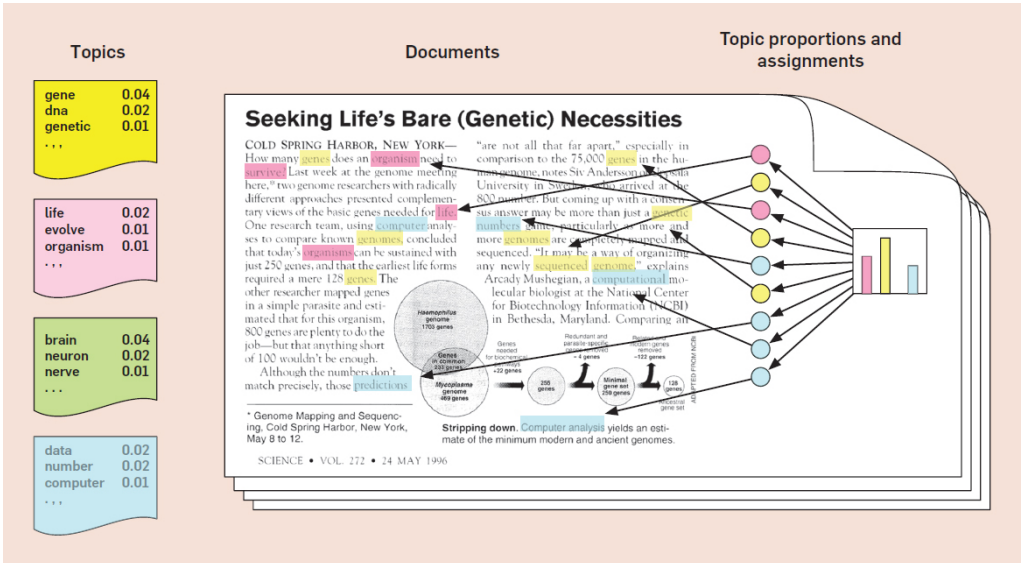
A generative view of the mixture of categoricals model:

1. Draw a distribution $\vec{\theta}$ over K topics from a $Dir(\vec{\alpha})$.
2. For each topic k, draw a distribution $\vec{\beta}_k$ over words from a $Dir(\vec{\gamma})$.
3. For each document d, draw a topic z_d from a $Cat(\vec{\theta})$
4. For each document d, draw N_d words w_{nd} from a $Cat(\vec{\beta}_{z_d})$

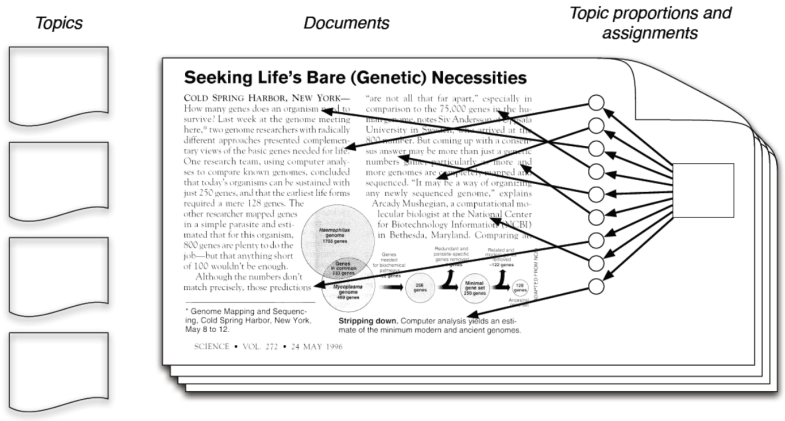
Limitations:

- All words in each document are drawn from one specific topic distribution.
- This works if each document is exclusively about one topic, but if some documents span more than one topic, then "blurred" topics must be learnt.

Latent Dirichlet Allocation



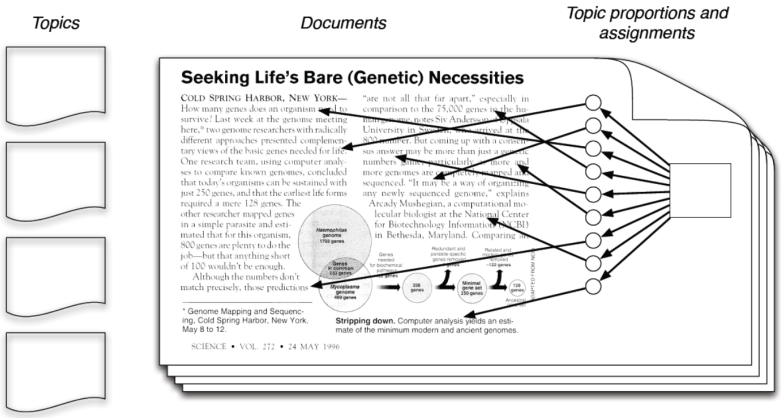
Latent Dirichlet Allocation: what we observe



In reality, we only observe the documents.

The other structure are hidden variables.

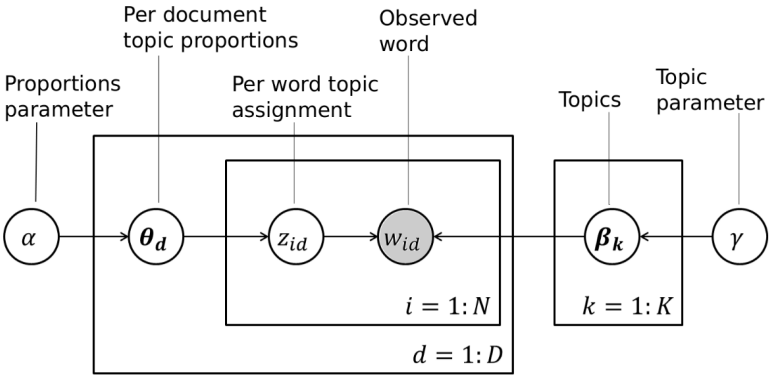
Latent Dirichlet Allocation: what we observe



Our goal is to infer the hidden variables.

This means computing their distribution conditioned on the documents $p(topics, proportions, assignments|documents)$

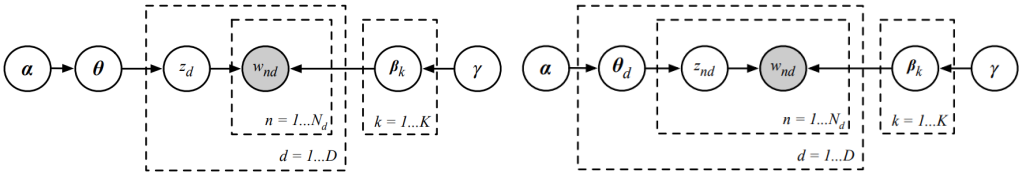
Latent Dirichlet Allocation: graphical model



Nodes are random variables; edges indicate dependence.

Shaded nodes indicate observed variables.

Mixture of Categoricals vs. LDA



A generative view of LDA:

1. For each document d draw a distribution $\vec{\theta}_d$ over topics from a $Dir(\vec{\alpha})$.
2. For each topic k draw a distribution $\vec{\beta}_k$ over words from a $Dir(\vec{\gamma})$.
3. Draw a topic z_{nd} for the n-th word in document d from a $Cat(\vec{\theta}_d)$.
4. Draw word w_{nd} from a $Cat(\vec{\beta}_{z_{nd}})$.

Differences with the mixture of categoricals model:

- In LDA, every word in a document can be drawn from a different topic,
- and every document has its own distribution over topics $\vec{\theta}_d$.

Gibbs sampling algorithm

- Initialize z_{nd} randomly for all words in all documents
- Choose random word and sample a new category based on all other words in all other documents.
- The distribution over categories is the predictive distribution of the posterior Dirichlet distribution (integration across all possible $\vec{\theta}$ and $\vec{\beta}$).
- Perform these small changes in many iterations over the data. The algorithm will converge to good solutions.