Related Papers

Motivation

- Searching and exploring documents based on the themes that run through them
- Rather than finding documents through keyword search alone, we might first find the theme that we are interested in, and then examine the documents related to that theme.
- **Probabilistic topic modeling**: a suite of algorithms that aim to discover and annotate large archives of documents with thematic information.
- **Note**: these algorithms, sometimes in different names, are used for other data types (audio, image, video...)}
**Topic Models (1)**

- **Unigram model**: the words of every document are drawn independently from a single multinomial distribution

\[ p(w) = \prod_n p(w_n) \]  

\[ (1) \]
Mixture of unigrams: each document is generated by first choosing a topic $z$ and then generating $N$ words independently from the conditional multinomial $p(w|z)$

$$p(w) = \sum_z p(z) \prod_n p(w_n|z)$$

(2)
**Probabilistic latent semantic indexing (PLSI):** it captures the possibility that a document \( d \) may contain multiple topics

\[
p(d, w_n) = p(d) \sum_z p(w_n|z)p(z|d)
\]  

(3)
Latent Dirichlet Allocation (LDA) expands PLSI by introducing priors on probability distributions.

Better generalisability on new documents.
Latent Dirichlet Allocation (2)

Generative Process

1. Randomly choose a distribution over topics
2. For each word in the document
   - Randomly choose a topic from the distribution over topics in step 1
   - Randomly choose a word from the corresponding distribution over the vocabulary

![Diagram showing probability distribution of topics]
Graphical Model
Latent Dirichlet Allocation (4)

Posterior Computation

- Posterior is intractable - need to approximate it
- Variational inference
- MCMC - Gibbs sampling
Extensions

LDA can be extended by relaxing some of the original assumptions

- **Bag-of-words**: Not suitable for language generation
  - Solution: Integrating syntax

- **Bag-of-documents**: Not suitable for chronologically ordered documents
  - Solution: Dynamic topic models

- **Number of topics**: Assumed to be known and fixed
  - Solution: Bayesian nonparametric topic models
Open Issues

- Evaluation and model checking: Interpretability over goodness of fit.
- Visualization and user interfaces: Intuitive ways to visualise topics.
- Data discovery: Seeking help of domain experts.