Complexity of Inference in Latent Dirichlet Allocation

David Sontag, Daniel Roy
(NYU, Cambridge)

W66

Topic models are powerful tools for exploring large data sets and for making inferences about the content of documents

Documents

Topics

politics .0100	religion .0500	sports .0105
president .0095	hindu .0092	baseball .0100
obama .0090	judiasm .0080	soccer .0055
washington .0085	ethics .0075	basketball .0050
religion .0060	buddhism .0016	football .0045
...	...	...

$$\beta_t = \{\, p(w \mid z = t) \,\}$$

Almost all uses of topic models (e.g., for unsupervised learning, information retrieval, classification) require **probabilistic inference**:

New document

Words w_1, …, w_N

What is this document about?

weather	.50
finance	.49
sports	.01

Distribution of topics θ

Complexity of Inference in Latent Dirichlet Allocation

David Sontag, Daniel Roy
(NYU, Cambridge)

W66

Main Results

Maximize $p(z_{1:N} \mid w_{1:N})$

For any α

# topics in MAP assignment	Complexity	Intuition
Small	Easy	First choose topic sizes, then match words to topics
Large	NP-hard	Reduction from set packing

Most common setting →

Maximize $p(\theta \mid w_{1:N})$

Dirichlet hyper-parameters	Complexity	Intuition
$\alpha_t \geq 1$	Easy	Maximizing concave function
$\alpha_t < 1$	NP-hard	Reduction from set cover

Most common setting →

Sample from $p(\theta \mid w_{1:N})$

Dirichlet hyper-parameters	Complexity	Intuition
$\alpha_t \geq 1$	Easy	Log-concave distribution
$\alpha_t \approx 0$	NP-hard	Reduction from set cover