Topic Modelling for Scientific Documents

Jethro Kuan
WING-NUS

Abstract

Topic models are statistical models, used to discover abstract topics that occur in a collection of documents. In this paper, we look at the application of topic models in the domain of scientific documents, and discuss the available shortcomings and solutions.

Introduction

The availability of knowledge through scientific documents is not met appropriately with the tools to navigate it. Tools like Semantic Scholar use artificial intelligence methods to digest scientific documents and present relevant results. An example of one such technique is topic modeling.

Topic models are probabilistic models for uncovering the underlying semantic structure of a document collection based on a hierarchical Bayesian analysis of the original texts. For more than a decade, Latent Dirichlet Allocation (LDA) and its variants have been the predominant technique for topic modeling.

LDA

Here we develop LDA from the principles of generative probabilistic models.

LDA models each document as a mixture of topics, where a topic β_k is a probability distribution over a fixed vocabulary of terms. The generative process is described as follows:

```
for each document w do
  Draw topic distribution θ ~ Dir(α);
  for each word at position n do
    Sample topic z_n ~ Multinomial(θ)
    Sample word w_n ~ Multinomial(β_{z_n})
```

then:

$$p(w|\alpha,\beta) = \int_\theta \left(\sum_{n=1}^{N} \sum_{z_n=1}^{k} p(w_n|z_n,\beta)p(z_n|\theta) \right) p(\theta|\alpha)d\theta \tag{1}$$

Posterior inference over the hidden variables θ and z is intractable due to the coupling between θ and β under the multinomial assumption. (Blei et al., 2003)

Statistical Assumptions

Despite its successes, LDA makes various statistical assumptions that make it less suitable for the task of modelling scientific documents.

1. The order of documents do not matter.

The meaning of keyphrases used in scientific literature change over time. For example, neural networks meant a different thing two decades ago.

2. Bag of Words

Inference Methods

Many learning algorithms have been developed, including collapsed Gibbs Sampling, Variational Inference, Collapsed Variational Inference, and MAP estimation.

A comprehensive look at the different inference approximation algorithms, shows that the performance differences can be explained away by setting certain smoothing hyperparameters (Asuncion et al., 2012). Nevertheless, we will take a closer look at them.

Variational Inference

Mean field variational inference (MFVI) breaks the coupling between θ and z by introducing free variational parameters γ over θ and ϕ over z and dropping the edges between them. This results in an approximate posterior $q(\theta, z|\gamma, \phi) = q_\gamma(\theta) \prod_n q_\phi(z_n)$.

To best approximate the true posterior, we frame it as an optimization problem, minimizing L where:

$$L(\gamma,\phi|\alpha,\beta) = D_{KL}\left[q(\theta,z|\gamma,\phi)||p(\theta,z|\alpha,\beta)\right] - \log p(w|\alpha,\beta) \tag{2}$$

This optimization has closed form coordinate descent equations, because the Dirichlet is conjugate to the Multinomial distribution. This computational convenience comes at the expense of robustness, making it difficult to apply to other more complicated topic models.

Extensions to LDA

Extension to Inference Methods

References

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. *CoRR*, 2012. URL `http://arxiv.org/abs/1205.2662v1`.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.