THE UNIVERSITY OF CHICAGO

Department of Statistics

MASTER'S THESIS PRESENTATION

LYNNE BUTLER

Department of Mathematics and Statistics
Haverford College

Topics and Themes in a Corpus of Prayers: An Application of Latent Dirichlet Allocation

WEDNESDAY, December 19, 2012, at 10:00 AM

110 Eckhart Hall, 5734 S. University Avenue

ABSTRACT

What are Baha'i prayers about? A statistical model called Latent Dirichlet Allocation helps us answer this question for a collection of 346 prayers. In the generative model, the K topics are distributions over words; and each document in the corpus has a unique distribution over those K topics, selected using a Dirichlet prior. Words in the document are generated one by one, by first choosing a topic from that document's distribution over topics, and then choosing a word from that topic's distribution over words.

Model parameters, $\alpha > 0$ for the symmetric Dirichlet prior and word distributions β_k for each of the topics, are found using variational EM. With a stemmed vocabulary, a model with 6 topics seems to find broad themes (spiritual growth and advancement, God's protection and forgiveness, His guidance, humility before Him, detachment from all but Him, His covenant with man).

What is the right number of topics in an LDA model for this small corpus? The log-likelihood of the corpus increases with the number of topics, but the likelihood ratio statistic does not have a chi-squared distribution. We report the results of simulations used to decide whether $K + 1$ topics is more suitable than K topics for our corpus, where $4 < K < 10$.
