Theological topics through time: An application of Gibbs sampling and other metrics to analyze topic venues in religious discourses

Michael Bean, M.A. Linguistics student
Dr. Eric Ringger, Computer Science

Brigham Young University

1 Introduction (and Inspiration/Motivation)

Much research has been performed concerning topic modeling. A subset of this research aims to analyze topical trends over time. Such work includes that of [9] where entropy, applied on chronological disjoint sets of texts, is used as a metric of showing broadening/narrowing of topics over time. Hall et al. also demonstrated that the Jensen Shannon divergence between venue[1] pairings could be used to measure their level of similarity. This work aims to perform a similar task, but on venues that are determined by meta-data other than just conference name. Meta-data used for groupings include gender of speaker and season of year (April/October). Since our dataset is different in nature than those used by Hall et al., it is hoped that their methodology will prove to be robust enough to allow the data to speak for itself, shedding light on the topics contained in religious discourses.

This study aims to prove the following hypotheses:

H_1 : The distribution of probability mass of topics fluctuates over time.

H_2 : Time of year will have an effect on the entropy of each venue.

H_3 : Gender will have an effect on the entropy of each venue.

H_4 : The Jensen-Shannon divergence between gender will be larger than the divergences found by Hall et al..

We prove these hypotheses applying metrics on each venu alone and within pairs of venues. Resulting values are graphed over time for easy interpretation.

2 Related/Prior Work

Prior text mining on LDS religious documents include [12], which aimed to identify intertextual similarity (plagiarism, paraphrase, etc.) between chapters of LDS-specific texts. In his work, Hilton focused on *The Book of Mormon*, although he has since demonstrated that similar results exist between *The Holy Bible* and *The Book of Mormon* [11]. Although topic models were not employed in this work, it probably could have benefitted from it. Notwithstanding, computational analysis was involved.

In concurrent work, I aim to use a combinational approach of gene sequence alignments and machine learning to automatically identify alignment positions of intertext. Although this is not directly related to this work, it has increased my familiarity with the LDS documents, dataset metadata, dataset format, and data warehouse–all of which have proven to be helpful in works such as this.

[1] or disjoint sets of documents

2.1 Topic Analysis Over Time

Hall et al. aptly demonstrated that topic entropy, when applied to topics on a per-year basis, could be used to describe the ebb and flow of each topic's popularity over time [9].

2.2 Topic Analysis on LDS Religious Documents

To the best of my knowledge nothing has been done in this area using the methodology we employ in this work.

2.3 Venue Comparison (in Light of Estimated LDA Parameters)

Again, this is where [9] shine. They showed the JS divergence was helpful in comparing/contrasting the distribution of topics and between datasets.

3 Methodology

First data is prepared by being canonicalized, stemmed, and stored. Then we use collapsed Gibbs sampling to infer parameters for our model, Latent Dirichlet Allocation. We divide our data into disjoint venues, then apply metrics on the estimated parameters of our model. We employ entropy to show how a venue narrows/broadens over time. This is graphed for easy interpretation. We normalize topic distributions for key time spans and graph the values of the resulting normalized topic probability mass over time. Lastly, we apply Jensen-Shannon divergence between pairs of venues to visualize how different/similar the venues are. This is also graphed over time for analysis.

3.1 Data

Discourses are taken from the LDS General Conference (GC)[2] from 1942 to 2013. There are traditionally general conferences per year: one in April and one in October. Most sessions are considered general and are for meant for all members of the church, while some are gender-specific. Interestingly, sessions that are for the women in the church [3] are generally held $\tilde{1}$ month prior to the general sessions. This means that the women's sessions are broadcasted in March and September. Women speakers are more prevalent, of course, at the women's sessions, so it is important for these discourses to be in the dataset. For purposes of venue grouping, they are treated as having occurred 1 month later; this means taht March women's sessions will be grouped with April general sessions, and likewise the September session will be grouped with the October general session.

3.2 Data Preparation

Our dataset is by no means free of noise. Our texts are derived from over 73 years of broadcasted religious discourse. Over this time, language is sure to have changed (except for the majority of specialized religious words). The speakers of the church have varied in type of discourse as well (extemporaneous vs. prepared). Gender is sure to play a role in some of the 'noisiness' of the data when viewed as a whole not only because of preferences in speech but because the probability that certain genders are given opportunity vary from year to year. By preparing the data for the algorithm, it is hoped that end result is a clearer picture of the noisy data.

[2] The LDS members convene during GC to hear from authorities and selected members of their church to be uplifted and edified together. Important announcements such as large-scale changes are diseminated in these conferences when appropriate. A recent change in minimum age for missionary service for the church is one example.

[3] "The Relief Society is the oldest and largest women's organization in the world. Relief Society was established in 1842 for women 18 years of age and older. Its purpose is to build faith and personal righteousness, strengthen families and homes, and help those in need." [15]

3.2.1 Canonicalization

Since many of our documents are dated before the 19th century, punctuation is often inconsistent; we therefore opt to ignore it except to aid in parsing. We do this for all documents. Furthermore, we lower-case all text.

3.2.2 Stemming

We stem words using the Java 7 port of Java 6 Porter Stemmer which was available online [18]; the Java 6 implementation was buggy in Java 7. In English, this stemmer allows words of different tenses to be simplified to a shorter, simpler form [19]. The porter stemmer algorithm is not perfect, but it is a straight-forward. A lemmatizer might perform better than a stemmer on this task, but we leave exploration in this space to future work. Beyond stemming and canonicalization, words are unchanged.

3.2.3 Stopwording

As is common in LDA models that whose parameters are estimated using gibbs sampling, we remove/ignore stopwords. Besides typical English stopwords such as 'the' and 'it', we ignore the following words while applying gibbs sampling: *god, jesus, christ, father.* These words have term frequency–inverse document frequency (TF-IDF) [21] scores that approach 1. These words were discovered by using Luke to perform queries over a Lucene index [1].

3.2.4 Format

Discourses are stored as comma-delimited words in JSON format in a MySQL database.

3.3 Model

We used a generative topic model known as Latent Dirichlet Allocation (LDA) [5]. It allows for multiple topics to be contained within a document by allowing each word instance (excluding stopwords) to be assigned 1 of K topics. This means that there is a topic distribution for each document. LDA explicitly models a topical distribution over all words as well.

3.4 Algorithm

3.4.1 Collapsed Gibbs Sampling

We use collapsed Gibbs sampling to infer the parameters of LDA [5]. Given hyperparameters and some parameters, the sampling algorithm outputs estimates for the unknown parameters (theta, phi). Collapsed Gibbs sampling provides only point-estimates of the parameters; this is sufficient for this work.

3.4.2 Hyperparameters

For the number of topics, K, we start with 50, but then settle to a lower number after some trial and error. We use uninformed hyperparameters:each value in the alpha vector is 0.2; values of the beta matrix are uniformly set to $1/K$.

3.5 Data and Meta-data

We view the data as being divisible into 2 types of meta-data before application of sampling, and as 1 type of meta-data after application. Although these are somewhat arbitrary/contrived categories, they are interesting ways to think of the data and are meant to help intuit future work.

3.5.1 Intrinsic Meta-Data

For each document, we have or can derive the following meta-data:

- *date discourse was given*;
- *time of year: April/October*;
- *length of document*; and
- *the text itself.*

Although one can imagine building a model which conditions variables on such intrinsic data, similar to how one might condition a model on time or dialect—which is the case for [20] and [6], respectively—we do not condition variables in our model directly on any of this data. Instead, we use the same post-hoc method of Hall et al. in dividing the dataset after applying gibbs sampling. We will only divide data into venues based on gender and time of year. The possibility for future work which directly models more intrinsic data is viable.

3.5.2 Extrinsic Meta-Data

Although we do not directly leverage the use of any extrinsic **meta-data** in our research here, with the exception of speaker's gender, we believe it is important to note that it exists. Intuition hints that accounting for these variables in the model could be of benefit, just as accounting for dialect would be helpful. Extrinsic meta-data which we have available include:

- *Name, age, gender of speaker*; and
- *Speaker's Name*
- *Speaker's Age*
- *Speaker's Gender*
- *Speaker's Church Position/Assignment*
- *Date*

3.5.3 Postrinsic Data

Upon applying gibbs sampling, we can obtain and/or estimate:

- *topic assignments for each non-stopword of each document*;
- *topic distribution within each document*;
- *topic distribution within each venue*;
- *entropy of each venue for each session of GC* ***overall and for each year***; and
- *Jensen Shannon divergence between each pair of venues, described in section 3.6.1* ***for each year***.

3.6 Metrics: Post-hoc Analysis

Similar to [9], we use the metrics described in Table 1 with the notations described in Table 2. These metrics allow us to show how a topic moves over time and how a venue is different/similar to another.

Metric	**Purpose**
Entropy	show how topics trend from year to year; show how some venues broaden/shrink over time
Jensen-Shannon Divergence	compare venues overall compare venues on a per-session basis

Tab. 1: Metrics and Purposes

Metric	**Notation**
Entropy	$H(C)$
Jensen-Shannon Divergence	$JS(C, C')$

Tab. 2: Metrics and notations, where C is some venue, and C' is another venue.

3.6.1 Venue Grouping

Following the methodology of Hall et al. we group our data according to venue. Following are the disjoint venues:

gender venues female/male

semi-annual venues April/October

time-gender venues female April, female October, male April, male October

3.6.2 Entropy Over Time

For sake of simplicity, the venues are all graphed according to the the largest time span in common: semi-annual.

By graphing entropy over time, we get a sense of the general trends for each venue with respect to topic diversity. Lower entropy indicates a venue is less diverse than another. Following is the formula for entropy:

$$H(C) = -\sum_{c \in C} p(c) \log p(c)$$

3.6.3 Topic Probability Mass Over Time

In each venue there is a distribution over topics. Like entropy, this can be graphed. This allows us to visualize how each topic varies in terms of popularity over time.

3.6.4 Jensen-Shannon Divergence Over Time

Since we have the probability distribution for topics in each venue, we can also compute a divergence between pairs of venues. This can also be graphed over time, allowing us to visualize the divergence. We use Jensen Shannon divergence (JS)[4]. Divergence values are proportional to the difference of the probability distributions. Thus a high divergence would indicate that two venues are different. Below is the formula for JS divergence, given two probability distributions, P, Q where M is the average of the two.

$$JS(P||Q) = 0.5 * KL(P||M) + KL(Q||M)$$
$$\text{where } KL(P||Q) = \sum_{i=1}^{n} \ln(\frac{P(i)}{Q(i)})$$

We define $KL(P||Q)$ to be zero when $Q(i) = 0$.

[4] JS divergence is the symmetrical corrollary of Kullback-Leibler divergence

Acknowledgements: A big thank-you is in order for Dr. Ringger for having taught me what I was lacking for this project. I thank my wife for being my best friend. She has proven to be true help-meet for me, especially when times get tough, such as when your database gets corrupted or locks up or when you think you are a victim of an amazing virus and is willing to help you change all your passwords at a moment's notice.

References

[1] Luke - lucene index toolbox, November 2013. [Online; Accessed on 2013-11-14].

[2] Ehsaneddin Asgari and Jean-Cedric Chappelier. Linguistic resources & topic models for the analysis of persian poems. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 23–31, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[3] David M. Blei. *Introduction to Probabilistic Topic Models.* 2007.

[4] David M. Blei and John D. Lafferty. Dynamic topic models, 2006.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[6] Steven P Crain, Shuang-Hong Yang, Hongyuan Zha, and Yu Jiao. Dialect topic modeling for improved consumer medical search. In *AMIA Annual Symposium Proceedings*, volume 2010, page 132. American Medical Informatics Association, 2010.

[7] John D. Lafferty David M. Blei. Correlated topic models, 2006.

[8] Christiane Fellbaum. *WordNet: An Electronic Lexical Database.* Bradford Books, 1998.

[9] David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the history of ideas using topic models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 363–371, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[10] S. C. Herring and A. J. Kurtz. Visualizing dynamic topic analysis. ACM Press, 2006.

[11] John Hilton III. Old testament psalms in the book of mormon. 2008.

[12] John Hilton III. Textual similarities in the words of abinadi and almas counsel to corianton. 2008.

[13] Abram Hindle, Michael W. Godfrey, and Richard C. Holt. What's hot and what's not: Windowed developer topic analysis. pages 339–348. International Conference on Software Maintenance - ICSM, 2009.

[14] Smith Krstovski and Wallach McGregor. Efficient nearest-neighbor search in the probability simplex, 2013.

[15] Mormon.org. What is the relief society?, October 2013. [Online; Accessed on 2013-10-24].

[16] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *In NAACL-HLT*, 2010.

[17] Justin Snyder, Rebecca Knowles, Mark Dredze, Matthew Gormley, and Travis Wolfe. Topic models and metadata for visualizing text corpora. In *Proceedings of the 2013 NAACL HLT Demonstration Session*, pages 5–9, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[18] C.J. van Rijsbergen, S.E. Robertson, and M.F. Porter. Java 7 port of porter stemmer, October 2013. [Online; Accessed on 2013-10-24; Ported by Michael Bean].

[19] Cornelis J Van Rijsbergen, Stephen Edward Robertson, and Martin F Porter. *New models in probabilistic information retrieval.* Computer Laboratory, University of Cambridge, 1980.

[20] Xuerui Wang and Andrew McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 424–433, New York, NY, USA, 2006. ACM.

[21] Wikipedia. Lucene — wikipedia, the free encyclopedia, 2013. [Online; accessed 14-November-2013].