A Natural Language Processing Analysis of General Conference Discourses Employing Time-Series and Classification Models

Clark Brown, Mikelle Rogers, Daniel Swingle,
Joseph Wilkes, Caleb Wilson

April 15, 2021

Abstract

Members of the Church of Jesus Christ of Latter-day Saints around the world study General Conference talks, thus motivating a system to enhance study sessions of people worldwide. Using several natural language processing techniques on General Conference talks from the Church, we seek to identify topics and speakers as part of a modular recommendation system. We utilize Latent Semantic Indexing (LSI) and Non-negative Matrix Factorization (NMF) to identify similar talks, Latent Dirichlet Collapsed Gibbs Sampling (LDACGS) to identify themes in the text, and Hidden Markov Models to identify topics and speakers. Overall, we achieved satisfactory results in addition to generating additional research questions.

1 Problem Statement and Motivation

Topic modeling, creating models to find patterns in a corpus, is an important field of study in natural language processing (NLP) and information retrieval (IR), which are both fields that analyze written language. We explore the corpus of discourses (ranging from 1942-2018) from the General Conference Sessions of the Church of Jesus Christ of Latter-day Saints using probabilistic methods for analyzing sequences. The various models included serve as the foundation of a recommendation system for a reader exploring these documents. Such a recommendation system would enhance study sessions of worldwide members of the Church who study General Conference talks. We provide models to identify themes, predict topics, find speakers' patterns, identify speakers, and retrieve similar talks.

NLP and IR have been applied to countless corpora. Work similar to our own, with Latent Semantic Indexing (LSI), was performed by [7] but applied to a different corpus. Several hobbyists and professionals have utilized the General Conference corpus of the Church. Their methods have included a Latent Dirichlet Allocation (LDA) analysis of topics [3], LDA with General Conference data in combination with other documents for scripture citation recommendation [5], and an analysis of topics over time presented in [6].

2 Data

The website, scriptures.byu.edu, containing over 8000 General Conference talks from 1942-2018, was scraped resulting in a corpus with meta-data including title, speaker, and date of each talk. Topic tags (about 300 distinct topics) were scraped from churchofjesuschrist.org, resulting in 3,000 talks from 1971-2018 with topic tags. We note that topics are subjectively assigned; however, it is not difficult for a member of the Church to read a talk and assign a justifiable topic. Thus, we assume these topics are valid. A link to our GitHub repository with the relevant scraping and model code is listed in the Appendix.

We acknowledge that the dataset is imbalanced in its distribution of topics and speakers. Some topics have as few as one talk, whereas the topic 'Jesus Christ' is assigned to over 800 talks and 'Faith' has around 500 talks. Similarly, the number of talks per speaker is also imbalanced with one-time speakers as the least frequent and Gordon B. Hinckley and Thomas S. Monson as the most frequent speakers both with 207 talks.

3 Methods

The Latent Dirichlet Collapsed Gibbs Sampling (LDACGS), Hidden Markov Models (HMM), and Naive Bayes are all generative models. These models compute the joint probability of words and the topic rather than the probability of the topic given the words. As a result, training a generative model is generally much faster than minimizing the loss function (like the cross-entropy), which occurs in discriminative models.

3.1 Gibbs Sampling & Latent Dirichlet Allocation (LDA)

We explored how to identify themes and topics incorporated into General Conference talks. Similar to work done previously, we started with a Latent Dirichlet Allocation (LDA) model. We use a Collapsed Gibbs Sampling (CGS) algorithm to learn the parameters of the LDA. We can combine LDA and CGS into one process to identify words that represent latent topics for the discourses.

As a result of using sampling, the LDACGS model is relatively fast, however, the stochastic nature of the process can lead to varying results based upon distinct initializations of LDACGS. To overcome this weakness, we ran LDACGS multiple times on each discourse and grouped the results. These results consisted of key words which are not associated with a specific topic. This introduces subjectivity as a result of manually associating key words with a collection of potential topics. LDACGS produces a log probability which represents the probability of a word given the learned parameters of topic likelihoods [4].

This log probability, when maximized, can be used as a metric for model selection with the hyperparameters α and β. In order to create the best model, we did a minor grid search amongst the hyperparameter values of α and β as seen in Figure 7 in the Appendix. From the gridsearch we deduced that the smallest α value ($\alpha = 0.001$) and the largest β value ($\beta = 1$) were the hyperparameters that maximized the log-likelihood. A small α value means the prior probability distribution for the topics over the set of documents (paragraphs of a talk) is Dirichlet with a higher variance than the uniform distribution over $[0, 1]$. The prior reflects an assumption of very low confidence, where we prefer that the data immediately swamp the prior in our guess at the true probability distribution. The $\beta = 1$ value means that the model uses a uniform prior for the distribution of topics over the vocabulary, which is a good choice since it is the maximum entropy prior. We used these values to run the model on a larger scale of data.

We trained a Random Forest algorithm on a mapping between the LDACGS key words and the true topics given by the pre-determined tags.

3.2 Hidden Markov Models (HMM)

We used Hidden Markov Models for classification by speaker and by topic. HMMs can handle inputs of various lengths and can also work in log probability to handle very small probabilities. We assume that the text of a talk is an observed sequence of words, where the probability of observing each word depends on a hidden state of the word, and the probability of moving to a different hidden state depends only on the current hidden state. Thus, the HMM does not provide a long term memory. We trained a separate model on each subset of the data with a shared characteristic (speaker or topic). Each validation talk ran through each model, and the speaker or topic with the model resulting in the highest log likelihood was predicted. When calculating the highest log likelihood, we used Bayes' Rule to account for speakers or topics having different probabilities of occurring in the training set.

For each of the 20 most frequent speakers, an HMM with 10 hidden states was trained. Using 10 hidden states was a good balance between allowing the models to have sufficient predictive power and controlling the time spent in training. Although talk frequency varied among speakers, we used a training size of 20 talks for each speaker.

An HMM was trained on all talks for the 78 topics with at least 50 corresponding talks. We performed cross validation, via minimizing the AIC, on the number of hidden states (2-5) for each topic. A predicted topic was correct if it was in the topic list of the corresponding talk. Various additional classification combinations were run from two to six topics, with multiple trials.

3.3 Naive Bayes Classifier

As a result of talks being classified as more than one topic, we trained a Bernoulli Naive Bayes classifier per topic to classify a talk as either in a topic or not in a topic. Naive Bayes assumes that words in the talks appear independently but that their probability of appearing is conditional on the topic of the talk. While this assumption does not hold, Naive Bayes classifiers perform well on textual data and are extremely fast to train compared to other methods.

3.4 Latent Semantic Indexing (LSI)

An effective way to compare talks is to find which talks are most similar to each other based on words and their frequencies used by speakers. The vocabulary of our data set includes roughly 80,000 unique words which makes most word frequency analyses computationally difficult. We generalized the method presented in [1] for LSI with Principle Component Analysis of documents as vectors of word frequencies.

With the trained model we can use cosine similarity to find the most similar talks to a given talk in the corpus. This allows the recommender system to provide a similar talk based on talks already read.

3.5 Kaplan Meier

Kaplan Meier curves are used in survival analysis to predict the probability of survival given a non-parametric time-event dataset. We created a time-event dataset where the event variable was the speaker saying “Christ” and the time variable was the number of words used before “Christ” was said in the talk. Our motivation for this method was to predict whether a speaker is an apostle using the probability of saying “Christ” in their talk. We define apostle as a member of the Quorum of the 12 Apostles or in the First Presidency of the Church of Jesus Christ of Latter-Day Saints, a distinction made using scraped data in addition to datetime objects.

3.6 Nonnegative Matrix factorization (NMF)

We used Nonnegative Matrix Factorization (NMF) to obtain hints regarding additional talks readers would enjoy. As a result of the absence of reader habit data, topics were used instead, and talks were features of the NMF. This assumption limits the usefulness of our NMF as a talk recommender. Given a talk with a specific topic, the matrix value indexed at the row corresponding to the talk and the column corresponding to the topic was 1, otherwise it was 0. NMF differs from LSI in that it can only recommend talks with similar topics instead of talks that are the most ‘similar’.

4 Results

4.1 LDACGS Topic Classification

We provide an example of the results of topic classification using the features from the LDACGS by way of a Random Forest classification model. More in

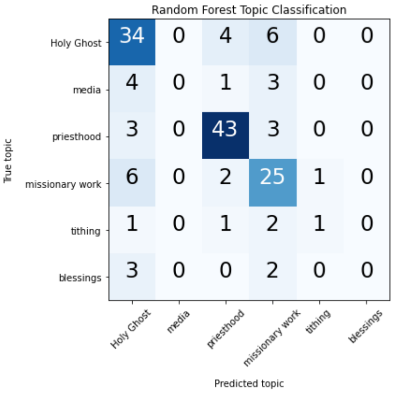


Figure 1: This confusion matrix represents the predicted and true values of the 6 topics based upon the LDACGS Random Forest. A disparity of number of samples between topics in the test set is apparent as noted by horizontal sums compared between rows.

depth analyses are described in section 5. The results of a separate example are depicted in Figure 1.

We trained a random forest classifier for 3 topics: hope, home, and missionary work using the output of the LDA as features. The words associated with each talk in the topic were one-hot encoded, totalling 3402 features. Any talk that was labeled with both topics was dropped from the training and testing data. The random forest got 84.6% accuracy on the test set. This shows that the output of the LDA meaningfully identifies words that indicate the topic of the talk. The most important words for the 3 topics hope, home, and missionary work are visualized in Figure 8 in the Appendix.

4.2 HMM Topic Classification

In Figure 2, we summarize some outcomes of training HMMs to predict topics of discourses. We expected accuracy to decrease as number of topics increased, however, there was a large range of accuracy among the same number of topics that we found surprising. This variation could result from talks having the potential to be in multiple classes. Potentially, topics with more overlapping talks produced more accurate results. Training on the 78 topics with more than 50 corresponding talks resulted in an accuracy of 31.81%.

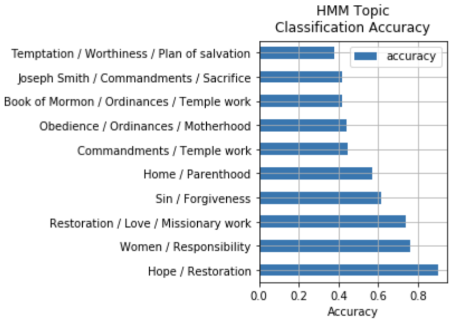


Figure 2: A depiction of multiple instances of HMM topic classification and the resulting accuracy.

4.3 HMM Speaker Classification

Running the HMM speaker classification model on talks by Thomas S. Monson and Gordon B. Hinckley resulted in models identifying the speaker with an accuracy of 81%. Some other speaker pairs we compared were James E. Faust and Boyd K. Packer, and Henry B. Eyring and L. Tom Perry. We observed an average accuracy of roughly 75% on speaker pairs and roughly 50% on speaker triples. Speaker identification for the 20 most common speakers achieved an overall accuracy of 28%. Increasing the size of the training set improved the accuracy for Thomas S. Monson and Gordon B. Hinckley, each of whom gave over 200 talks: with a training set of 100 talks each, the model successfully classified talks between Thomas S. Monson and Gordon B. Hinckley 88% of the time.

4.4 Naive Bayes Topic Classification

Our ensemble of Bernoulli Naive Bayes classifiers performed very well in the topic classification task. A sample of the accuracy of these topic classifiers is given in Table 1. Greatest accuracy was seen for specific topics like 'Tithing' or 'Pornography' that have a unique vocabulary associated with them. These topics had accuracy above 95%. We saw for more general topics like 'Jesus Christ' or 'Faith', the classifiers were still accurate, but on the order of 75-85%.

4.5 HMM Text Generation

Hidden Markov Models can be sampled from to generate text. As a result, we attempted to generate text using our trained speaker classification

Topic	Test Accuracy	Train Proportion	Test Proportion
Jesus Christ	76.92%	22.56%	21.83%
Faith	86.54%	13.69%	12.69%
Family	87.60%	10.97%	11.44%
Peace	96.92%	3.22%	3.17%
Tithing	98.65%	1.53%	1.44%

Table 1: Accuracy of Bernoulli Naive Bayes classifiers for select topics. The 'Train Proportion' and 'Test Proportion' are the proportion of talks with the topic label in each data set.

HMM models. Neither character-level nor word-level text generation models were effective; the respective models learned some spelling and grammatical patterns but did not produce intelligible results. This is likely due to the little amount of training data available to the models and the incorrect assumptions Markov chains make about the sequence of text.

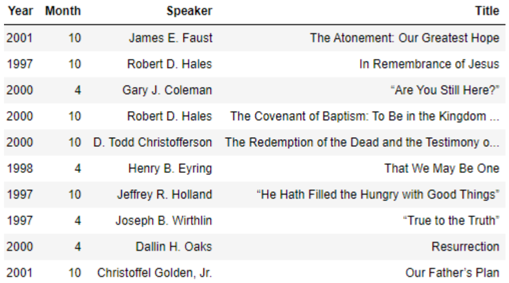
4.6 Kaplan Meier

As a result of the survival curves in Figure 10 (in the Appendix) crossing, according to [2], the logrank test will give an inaccurate assessment of differences. We observe that the curves are nearly on top of each other, so we conclude that the occurrence of the word 'Christ' in a General Conference talk does not predict whether the speaker is an apostle.

4.7 Latent Semantic Indexing

Using 7 principle components explains only about one tenth of the variance, but greatly reduces the dimension of the problem by more than one ten-thousandth. Using a larger number of principle components led to memory problems with the computing machines available to us. An even greater reduction to 2 principle components to try and visualize the data was not insightful (see Figure 9 in the Appendix).

We also built an undirected graph where the vertices of the graph were talks, and an edge was added between each talk and the talk most similar to it. We then analyzed the connected components of this graph to discover communities by topic or by speaker. However, it seems that 7 principle components are too few for this to be meaningful because the components included a more diverse distribution of topics than we had hoped, and none



Year	Month	Speaker	Title
2001	10	James E. Faust	The Atonement: Our Greatest Hope
1997	10	Robert D. Hales	In Remembrance of Jesus
2000	4	Gary J. Coleman	"Are You Still Here?"
2000	10	Robert D. Hales	The Covenant of Baptism: To Be in the Kingdom ...
2000	10	D. Todd Christofferson	The Redemption of the Dead and the Testimony o...
1998	4	Henry B. Eyring	That We May Be One
1997	10	Jeffrey R. Holland	"He Hath Filled the Hungry with Good Things"
1997	4	Joseph B. Wirthlin	"True to the Truth"
2000	4	Dallin H. Oaks	Resurrection
2001	10	Christoffel Golden, Jr.	Our Father's Plan

Figure 3: Our NMF recommended the above ten talks related to "Church doctrine". From inspection of the talk titles we can see this is quite a reasonable recommendation.

of the larger components consisted of unique speakers.

4.8 NMF Talk Recommendation

Our NMF factorization converged with relative tolerance 0.001 for a rank 3 factorization. The success of the NMF can only be evaluated qualitatively. The recommendations are better for some topics than others. Despite being completely unsupervised however, in our observations, the NMF produces relevant recommendations. We can also see which topics belong most to each of the 3 component dimensions involved in the rank 3 factorization, which could be used to decide which topics are similar to each other. Unfortunately, the results of this topic grouping were not enlightening.

5 Analysis

LDACGS and HMMs were successful at differentiating and properly predicting the topic of a talk in many cases. We compare the accuracy of these two models for various topics in Figure 4. We will describe separately the specifics of the LDACGS and the HMM models at classifying topics.

LDACGS was fast at generating key words and using LDACGS made it easy to control the feature space for relating words to topics because we got 50 words each time we ran LDACGS on a talk. It would become less useful to rerun LDACGS on the same talk over and over because it would just increase the number of times certain words showed up as features. We observed that having to connect the output of LDACGS to a separate classification algorithm introduced more variability into the process of associating key words with their topic tags. This variability lead to a high variance of

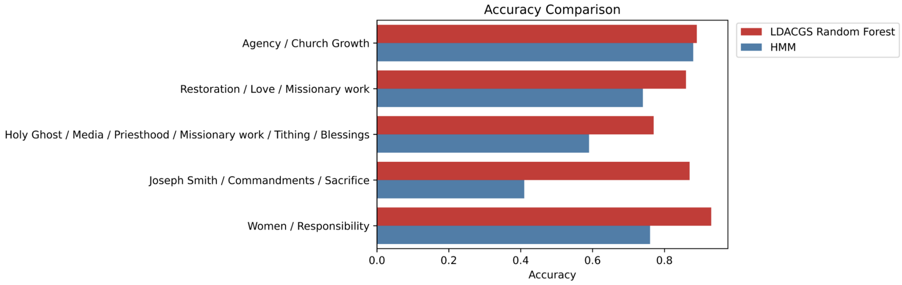


Figure 4: When comparing topic classification accuracy of the LDACGS Random Forest to the HMMs we can see that in some cases the Random Forest is superior whereas in other cases they both give comparable results.

accuracies of the Random Forest. This was minimized when running a grid search with the Random Forest and yielded great intuition when we were able to quantitatively look at feature importances. The feature importances gave a weight associated to how important the words were for predicting the topic.

The HMM approach of topic classification also had a large variance in the accuracy. Various classification trials were run on multiple randomly chosen groups (of size 2, 3, 4, 5, 6) of topics. As a result of the over 300 topics there were many groups of talks to be classified. Depending on the pair of topics chosen the accuracies ranged from 30% to 90%. Complete understanding of this extensive range would require further analysis, however, we give the following conjectures. The differing results could stem from talks having multiple topics rather than one definitive label. Additionally, although HMMs utilized Bayes' Rule to scale the probability, a negative influence resulting from imbalanced data is possible.

Our simple Naive Bayes approach to topic classification yielded better results than anticipated. For niche topics, we found accuracy over 95%. Even the general topics were more accurate than randomly guessing (50%). We note that the test accuracy was inversely correlated with the proportion of talks with that topic label in the test and train sets as seen in Table 1. Hence, a classifier that always makes the classification of *not in topic* would have achieved almost approximately the same accuracy as the trained classifiers for each of the topics in Table 1.

Training an HMM for each speaker allowed us to achieve almost 90% accuracy classifying between two speakers with a larger training set of 100 talks each. Our conclusion is that provided sufficient training data, HMMs can be an effective way to predict the author of a text sample; however, they are less effective when distinguishing between more than three speakers. This would be alleviated by having access to a larger training set, as relatively few speakers spoke frequently in conference: only two speakers in our dataset gave more than 100 talks.

6 Ethical Considerations

The corpus of text used for this research does not contain private information. The data are publicly available, but are protected by copyright law as intellectual property of the Church. In order to respect the Church's right to the talks we cannot make our scraped dataset publicly available; however, we can provide the code to allow others access to the same information.

Analyses of the topics of these discourses reflect on the positions and values of the Church of Jesus Christ of Latter-day Saints. The authors of this paper made all analyses independent of their personal religious beliefs and did not censor or exaggerate any results.

7 Conclusion

When ample data is present, Hidden Markov Models achieve quality speaker classification results. Relatively few speakers frequently give talks in conference, so our dataset was unbalanced. If the dataset were expanded to include more speeches, we would likely achieve higher accuracy when classifying orators; however, this is outside of the scope of our project.

We achieved acceptable topic classification results across our models. The sheer number of topics proved challenging. Understanding the high variance present in results produced by HMMs is a subject of further research. LDACGS produced more consistent results when combined with the random forest classifier. The approach taken with Naive Bayes where a talk was classified as either in or out of a category could generalize well and allow us to assign multiple subjects to a single talk.

Using the topic classifiers we trained we could label the topics of the talks that did not have assigned topics and retrain our NMF to make a better recommendation system. If we built an app with the recommender system we could collect user data to validate our recommender and incorporate user feedback to further improve the recommender.

References

[1] Finding patterns in data: Lsi and more about scikit-learn, Nov 2020.

[2] Cam Davidson-Pilon. Statistics, 2021. Revision 96b7bf81.

[3] Peter Empey. Using lda topic modeling on lds general conference, Dec 2015.

[4] Shusei Eshima. Lda log likelihood, 2021. See the Journal Article on the website as well.

[5] Joshua Mathias. Contextual scripture recommendation for writers. Master's thesis, University of Washington, 2019.

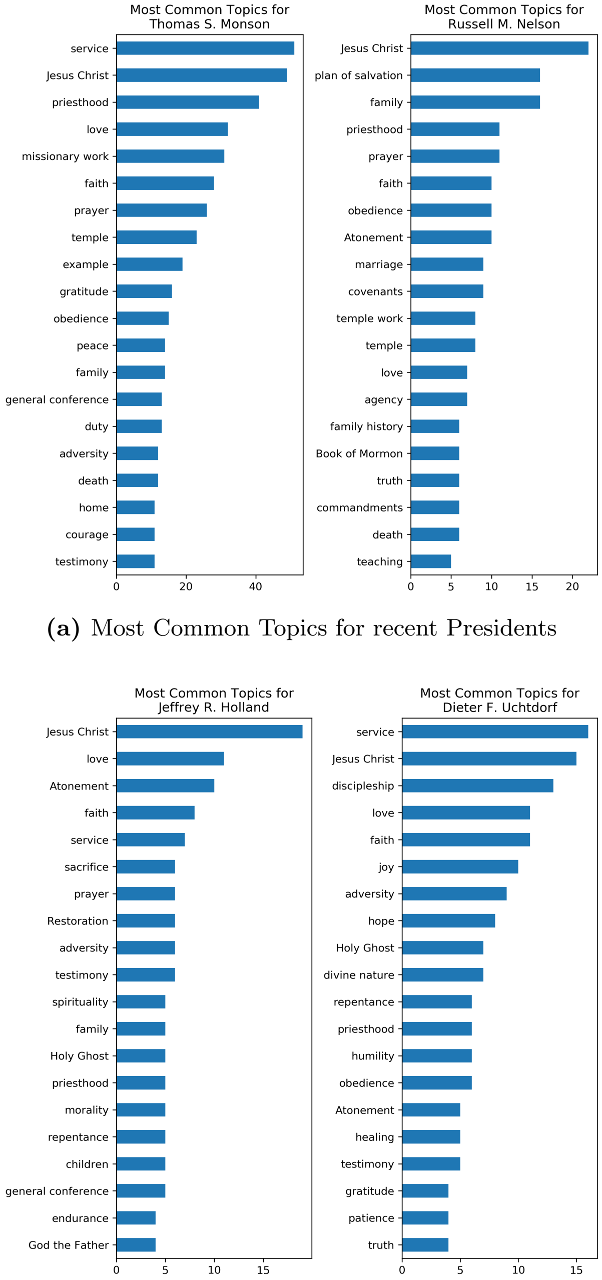
[6] Quentin Spencer. What don't the prophets say anymore? a text analysis of general conference 1942-2020, 2020.

[7] Quan Wang, Jun Xu, Hang Li, and Nick Craswell. Regularized latent semantic indexing: A new approach to large-scale topic modeling. *ACM Trans. Inf. Syst.*, 31(1), January 2013.

Appendix

Code for the models presented in this paper can found at `https://github.com/lodeous/gencon-nlp`.

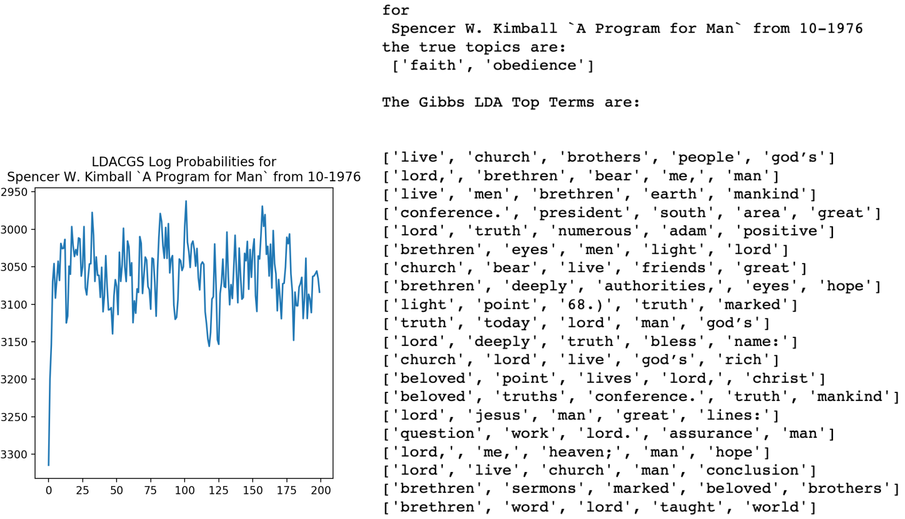
See more Figures on the following pages.



(a) Most Common Topics for recent Presidents

(b) Most Common Topics for some Apostles

Figure 5: For four prominent individuals, we can show what the most common topics are of their talks. It is interesting to compare the most common topics for the two most recent Presidents of the Church. The x-axis is the number of talks that have the topic included.



```
for
 Spencer W. Kimball `A Program for Man` from 10-1976
the true topics are:
 ['faith', 'obedience']

The Gibbs LDA Top Terms are:


['live', 'church', 'brothers', 'people', 'god’s']
['lord,', 'brethren', 'bear', 'me,', 'man']
['live', 'men', 'brethren', 'earth', 'mankind']
['conference.', 'president', 'south', 'area', 'great']
['lord', 'truth', 'numerous', 'adam', 'positive']
['brethren', 'eyes', 'men', 'light', 'lord']
['church', 'bear', 'live', 'friends', 'great']
['brethren', 'deeply', 'authorities,', 'eyes', 'hope']
['light', 'point', '68.)', 'truth', 'marked']
['truth', 'today', 'lord', 'man', 'god’s']
['lord', 'deeply', 'truth', 'bless', 'name:']
['church', 'lord', 'live', 'god’s', 'rich']
['beloved', 'point', 'lives', 'lord,', 'christ']
['beloved', 'truths', 'conference.', 'truth', 'mankind']
['lord', 'jesus', 'man', 'great', 'lines:']
['question', 'work', 'lord.', 'assurance', 'man']
['lord,', 'me,', 'heaven;', 'man', 'hope']
['lord', 'live', 'church', 'man', 'conclusion']
['brethren', 'sermons', 'marked', 'beloved', 'brothers']
['brethren', 'word', 'lord', 'taught', 'world']
```

(a) Log Probabilities from LDA

(b) Top Terms from LDA

Figure 6: Applying the LDACGS to an individual discourse by Spencer W. Kimball.

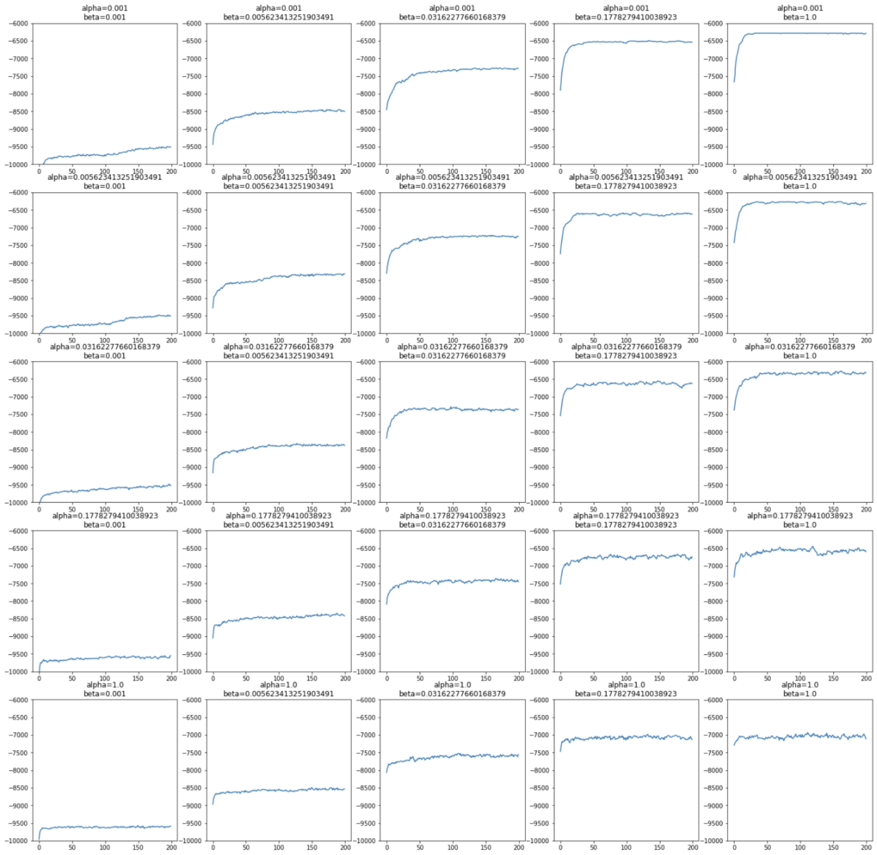


Figure 7: Hyper Parameter search for the LDACGS model with 5 values for both α and β.

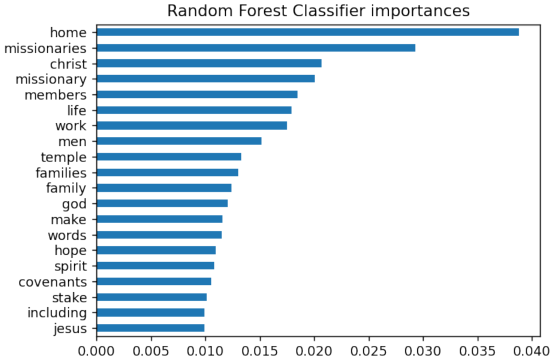


Figure 8: This visual shows the words most associated with identifying talks labelled as one of the following topics: home, hope, and missionary work as determined by the Random Forest model.

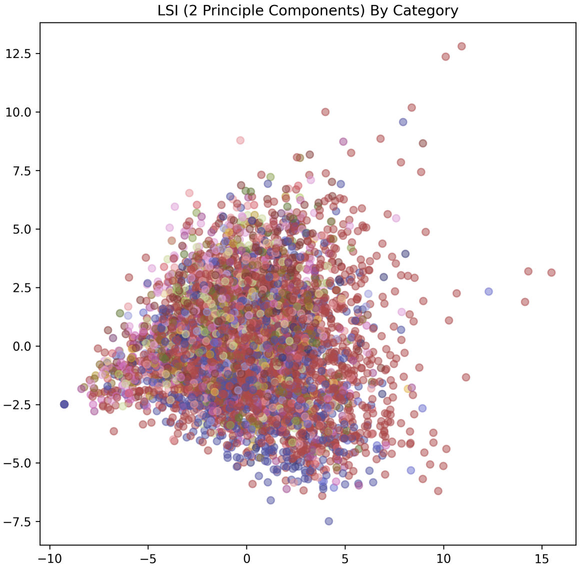


Figure 9: Visualizing the entire body of discourses with LSI/PCA with 2 components, colored by topic

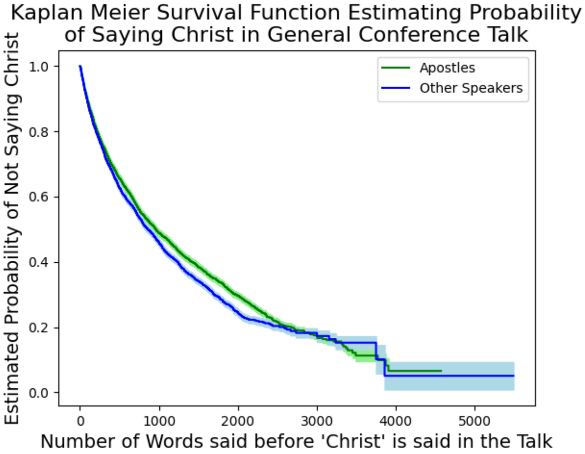


Figure 10: Separation of Apostles from other speakers at General Conference reveals that there is not significant separation between either the time at which Christ is said nor the probability of not saying Christ.