Topic Dynamics in US Court Decisions on School Desegregation

Michael Ge Alexander Goldberg Karl Otness

Abstract

There have been numerous court cases in the United States on the subject of school desegregation, but little quantitative work has been done to understand how the arguments made and the reasoning followed in these court cases has developed over time. In response to similar problems, probabilistic topic models have been exploited to uncover underlying classes of arguments, but there has been less investigation into accounting for the change in arguments over a temporal axis. In this paper, we attempt to model both topics and their temporal dynamics over a corpus of unlabeled court decisions on school desegregation since the "separate but equal" ruling in Brown v. Board of Education. We compare models capturing and ignoring time as a feature and explore the benefits of incorporating time dynamics into topic modeling of legal documents. Our findings suggest that arguments on school desegregation have been surprisingly stable over the last five decades.

1. Introduction

The issue of school segregation has sparked controversy even in modern times, constantly defining and redefining the ways academic institutions have tried to balance equal opportunity in an unequal world. Yet, despite the importance of understanding the underlying issues behind school segregation and the large corpus of court documents available on the subject, there has been little quantitative work on how these issues have changed in priority over time.

One potential solution to the problem is to use a probabilistic approach in which documents are modeled as bags of words drawn from an underlying topic and word-topic distribution. However, this basic "topic model" assumes that the topics do not change over time. A paper by Blei and Lafferty [2] proposes a modification of standard topic models that incorporates time dynamics, though due to the complexity of the model, most of the details of dynamic topic models have not been worked through in detail.

In this project, we attempt to implement Blei and Lafferty's model to capture both the distribution of topics in school segregation-related court opinions as well as their evolution over a span of time. In working on the project, we provide a detailed Python implementation of a static topic model, expand this model under a dynamic topic context, and explore large-scale findings using a library dedicated to dynamic topic models. Our findings show that the most relevant topics pertaining to school segregation are surprisingly stable and remain fairly consistent from year to year, suggesting that the arguments, reasoning and law applied in court opinions issued in school segregation cases have remained relatively static.

2. Background

This project was suggested by a JD-PhD candidate in the Harvard Sociology department, Jimmy Biblarz, who studies (using qualitative methods) the development of social thought on school segregation. In 1954, the US Supreme Court declared as a result of Brown v. Board of Education that states could not legally segregate their schools. Since then, numerous attempts have been made to overturn or modify the ruling based on a variety of arguments, many of which are captured in a series of publicly available court opinions. The corpus is fairly large, motivating the application of quantitative modeling of the court cases to try to uncover underlying patterns.

3. Related Work

Similar questions of how arguments in political and legal texts have changed over time have been asked on other datasets in the quantitative sociology literature. For instance, applications have included studying changes in State of the Union discourse [10], development in German party platforms over time based on the text of party manifestos, [11], and shifts in American nationalism [4] based on textual survey responses. However, most of these text sources lack the structured text of a court case, which enables richer models in describing our dataset. Further, many of the approaches taken are fairly ad-hoc – for instance, Rule et. al.[10], find topics by clustering over the network implied by taking words as nodes and their co-occurences as weighted edges – rather than being based in an underlying probabilistic model.

To our knowledge, Blei and Lafferty's dynamic topic model

[2] has not before been used to study changes in a corpus of legal documents over time. In their initial paper, Blei and Lafferty apply DTMs to study changes in scientific papers over time, and the model has been used to study changes in online discourse, such as topic dynamics on Twitter [1]. Thus, the application of a dynamic topic model to summarize and interpret a large body of legal documents is a fairly novel approach.

4. Data Sources

Our work is based on a dataset of court opinions collected by CourtListener[5]. CourtListener has collected the raw text of opinions from court websites and provides some related data on the cases including the court level, citations used in the opinion and the judge presiding over the case.

We retrieved the bulk opinions dataset from CourtListener which contains 30GB worth of court opinions. From this we produced a smaller set of school segregation-related cases by taking only opinion texts containing *both* “school” and “segregation” as substrings. This produced a set of about 8 497 opinions.

4.1. Data Preparation

The raw dataset required significant processing, cleaning and normalization. Some of the opinions are in HTML while others are in plain text. Further, many cases contain artifacts of automated processing such as control characters or unexpected symbols. To correct this we began with a normalization pass. We replaced typographical quotes with their plain counterparts and replaced a commonly-occurring Unicode control character with a plain space. After this, if the text was in HTML format, we converted it to plain text using BeautifulSoup4.

CourtListener does not provide publication dates for the opinions; however, this information is available from the text of the opinions. Therefore we parsed out dates from the opinions handling the variety of date formats found. We produced regular expressions to match several date formats with preference toward those which more confidently match dates. We match formats listing month, day and year then fall back to formats listing only month and year and finally only year. The distribution of date formats identified in the texts is listed in Table 1. This year distribution is plotted in Figure 1.

As illustrated in Figure 1, most of the documents were dated within the expected time frame. Due to our heuristic method for identifying publication dates, however, there are some incorrect values. In particular, very early dates for court cases are incorrect. In our time-dynamic analysis described below, we will ignore these cases and focus only on a range of years which have both a large number of cases and a reasonable likelihood of having correctly identified dates.

Format	Number Found
Month Day, Year	8 195
Month/Day/Year	208
Day Month Year	14
Month, Year	8
Year	72
No Date	1

Table 1. Date formats in text corpus

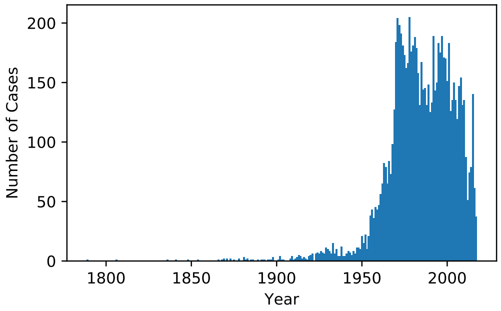


Figure 1. Year distribution in dataset

4.2. Vocabulary Generation

Our topic models are trained on word count vectors. To produce our training data we further normalized words by removing repeated punctuation such as periods or commas. Next, we produced vocabularies of words which appeared at least a certain number of times. For our time-independent LDA we set our cutoff at 500 occurrences. For the time-dynamic topic model we set the threshold to 19 800 to produce a smaller vocabulary. We excluded very common stop words from the vocabulary as these appear in every document and do not contribute to understanding the data. These steps produced a large vocabulary of 9 471 words and a smaller vocabulary of 472 words.

5. Model

5.1. Latent Dirichlet Allocation

One of the more commonly used topic models is the Latent Dirichlet Allocation (LDA) model proposed by Blei, Ng, and Jordan[3]. LDA generates documents from two sets of distributions: per-document topic distributions and global per-topic word distributions. Each word in a given document is chosen by first drawing a topic from the document’s topic distribution, then the word is chosen from that topic’s distribution over the vocabulary.

Figure 2 illustrates the structure of this model on K topics,

M documents and N_m words in document m. The latent variables z mediate the choices of topic (from the distribution θ) for each word and vocabulary words are then drawn from the appropriate per-topic distribution β.

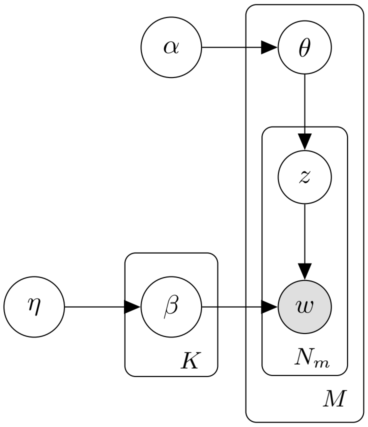


Figure 2. Generative model for LDA

LDA posits the following conditional distributions on the documents [7, Sec 27.3.1][3]:

$$\theta_m \mid \alpha \sim \mathrm{Dir}(\alpha) \tag{1}$$
$$z_{mn} \mid \theta_m \sim \mathrm{Cat}(\theta_m) \tag{2}$$
$$\beta_k \mid \eta \sim \mathrm{Dir}(\eta) \tag{3}$$
$$w_{mn} \mid z_{mn} = k,\ \beta_* \sim \mathrm{Cat}(\beta_k). \tag{4}$$

We will discuss inference for this model in Section 6.1.

5.2. Dynamic Topic Model

The LDA model makes several independence assumptions about the document corpus. First, it assumes that the words are independent of each other order-wise—bag of words—and second, it assumes that the documents are independent of each other. The former assumption is a relatively standard assumption and we believe it to be acceptable for our dataset. The latter assumption is undesirable in that it fails to capture how discussion evolves over time. The Dynamic Topic Model (DTM) by Blei and Lafferty[2] makes it possible to track this evolution.

This model is composed of several pillars each of which is an LDA model which is fitted to documents in a given time range. To require the topics to align across the time ranges, the topic distributions are linked. These distributions are allowed to smoothly vary over time. The graphical model structure of the DTM with three time steps is illustrated in Figure 3.

For our purposes, we take the prior hyperparameters α to be constant and uniform: $\alpha_{t-1} = \alpha_t = \alpha_{t+1}$. Therefore, the most salient difference between DTM and LDA is the

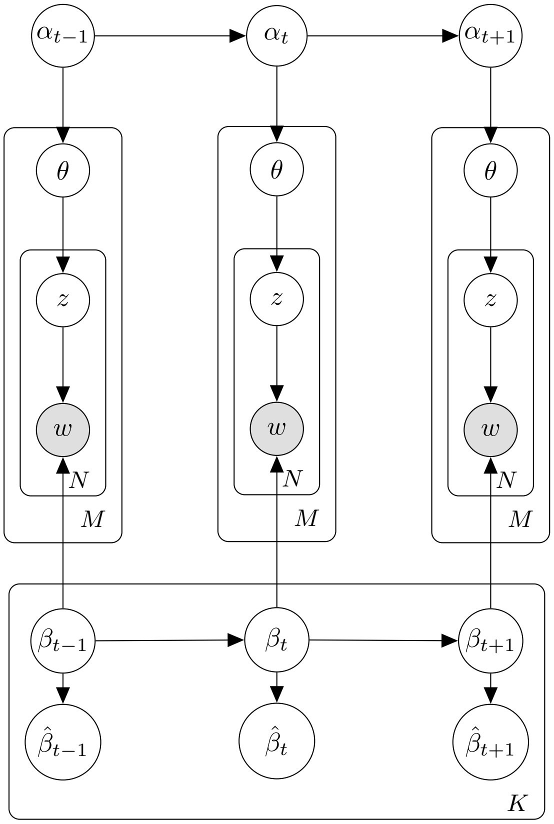


Figure 3. Dynamic Topic Model for $t = 3$

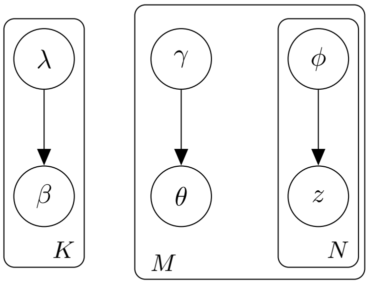


Figure 4. Variational distribution for LDA

relationship between the β_t parameters. The β parameters are allowed to evolve subject to a Gaussian distribution on their *natural* parameters. That is,

$$\beta_{tk} \mid \beta_{(t-1)k} \sim \mathcal{N}(\beta_{(t-1)k} \mid \sigma^2 I). \tag{5}$$

In order to draw words from this distribution (to go from the natural parameters to the mean parameters of a categorical distribution) we apply the softmax function to β_t and words are drawn from the categorical distribution with mean parameters $\sigma(\beta_t)$.

Most of the inference on this model can proceed as in standard LDA, but we will handle the β_t parameters differently, fitting variational parameters $\hat{\beta}_t$ in a way that will preserve the smooth evolution described in Equation 5. Inference on this model is described in Section 6.2.

6. Inference

6.1. Inference for LDA

In training these models we want to infer the model parameters (topic and word distributions) as well as the latent variables. The posteriors for both LDA and DTM are infeasible to evaluate exactly, therefore we will perform variational Bayes training. Our variational distributions for both models will take the form of a mean field approximation; however for DTM we will make an exception for the chained word-topic distributions and their dependencies.

Mean field updates for LDA are given in a variety of sources[7, Ch. 27][3][6]. The variational distribution for LDA is illustrated in Figure 4.

Variational inference proceeds by maximizing ELBO (a lower bound on the log-likelihood) which, denoting the mean field approximation as q is given by

$$\begin{aligned}\mathcal{L}(w, \phi, \gamma, \lambda) = \sum_{n=1}^{N} \{ &\mathbb{E}_q[\log p(w_n|\theta_n, z_n, \beta)] \\ &+ \mathbb{E}_q[\log p(z_n|\theta_n)] - \mathbb{E}_q[\log q(z_n)] \\ &+ \mathbb{E}_q[\log p(\theta_n)|\alpha)] - \mathbb{E}_q[\log q(\theta_n)]\} \\ &+ \mathbb{E}_q[\log p(\beta|\eta)] - \mathbb{E}_q[\log q(\beta)]\end{aligned}$$

Then, coordinate ascent on this loss function yields the following update rules for ϕ and γ [3]:

$$\phi_{nk} \propto \beta_{kw_n} \cdot \exp\Big\{\psi(\phi_k) - \psi\Big(\sum_{j=1}^{K} \phi_j\Big)\Big\} \tag{6}$$

$$\gamma_m = \alpha_m + \sum_{n=1}^{N_m} \phi_{nm*}. \tag{7}$$

where ψ is the digamma function. Repeating similar analysis for λ yields the update rule:

$$\lambda_{kv} = \eta + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dnk} \cdot \mathbf{1}(w_{dn} = v). \tag{8}$$

Note that η and α are prior hyperparameters to the LDA model. We can fit the variational distribution applying the EM algorithm until convergence[7, Sec. 27.3.6.3]. During the E-step we will use the updates from Equations 6 and 7. During the M-step we apply the update in Equation 8 and repeat this process until convergence.

6.2. Inference for DTM

Inference on DTM proceeds similarly to inference on LDA. However when we are fitting the variational parameters $\hat{\beta}_t$ we do not want to use a pure mean field approximation as this would sever the links between the topic-word distributions across time. To do this, Blei and Lafferty make use of Kalman filters[2]. We can view the $\hat{\beta}_t$ as observations from a the hidden process described by the β_t. Because these parameters evolve subject to a Normal distribution (Equation 5), this precisely matches the form of a Kalman filter. Therefore this approach enables exact inference. In our implementation we used a Python Kalman filter implementation, PyKalman[9] to perform portion of the inference on the DTM.

7. Methods

We began by running our implementation of a time-static LDA on the large vocabulary with 9 471 words. For 30 epochs, we performed a Variational Bayes EM step, generating a word distribution for $K = 20$ topics. The results can be found in Table 2. Our LDA implementation yielded results in about 2 hours.

Unfortunately, running the DTM on the same vocabulary size was not a feasible option. Due to the Kalman smoothing step, we introduce extra computation cubic in the size of the vocabulary over several epochs. [7, Sec 18.3.2][3] As a result, even by substituting segments of our code with Cython, our implementation was too slow for the 9471^3 number of operations required to calculate the smoothing.

After experimenting with numerous configurations of vocabulary sizes and partitions of time ranges, we opted to analyze the results for the 472-word vocabulary with one-year and one-decade time range granularity using the PyKalman off-the-shelf filter in conjunction with our personal implementation of the LDA. The results can be found in Tables 3 and 4. With the reduced vocabulary size, the model training process took 5 hours for the unpartitioned dataset and 3 hours on a dataset partitioned into decades.

8. Results

We now discuss the results of our models and explain the motivations behind the presented results.

8.1. Quantitative Measurements

We first experimentally justify choosing a reasonable number of topics. In our static-time LDA model, we calculated the approximate per-word log-likelihood for different values of K on a 10% holdout to get an estimate for an appropriate value of K.

We find that the log-likelihood is maximized at $K = 20$:

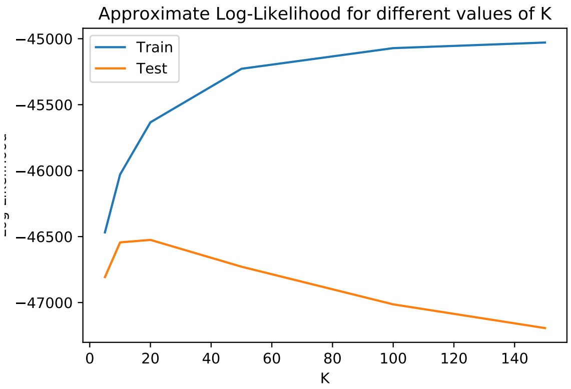


Figure 5. Log-Likelihood for varying K. Log-likelihood is maximized at $K = 20$

With this measurement, we now proceed to evaluate the words for $K = 20$. Because the model relies on performing inference on unlabeled data, there are few metrics that quantitatively describe the improvement or performance of our data. As a result, we will continue by presenting a qualitative analysis of our findings.

8.2. LDA Results

The LDA's results in Table 2 show that the topic model creates a reasonable distribution of words over topics. The table lists a subset of the 20 topics generated by training. In each column, the top five most representative words (based on having the highest word probability in the word-topic distribution) in each topic are listed from top to bottom. Many of the topics have words that would be expected from a corpus of documents on school segregation. Topic 3, for example, seems to contain the umbrella terms associated with school segregation.

More interesting are topics that seem to indicate underlying issues indirectly related to segregation. Topic 1, for example, indicates interest in the criminal system in the court cases. It is possible that this arises from concerns pertaining to inequalities in education leading to increases in criminal activity. Topic 17 indicates a geographical interest in school segregation. The prevalence of words related to Texas and the Southwestern Reporter (s.w.2d) hint at school segregation being a more contentious issue in the southern states.

It is worth noting that there were also topics that were less insightful. As an example, topic 0 contained several numbers as the most representative words. This is to be expected, however. Given our conservative vocabulary trimming, it is no surprise that we have uninformative topics.

Of course, these comments are speculative, and further investigation into the court opinions is required to confirm these ideas. In short, we find that our static-time LDA model reports reasonable results.

8.3. DTM: Year-Granularity

We now examine the results of incorporating topic evolution over time with a DTM. In this model, an LDA is trained on the small vocabulary corpus for each year from 1950 to 2017, and the Kalman Filter smooths over each of these years individually. Table 3 lists the top 5 words at every decade for topic 1.

Looking at the words alone, it can be difficult to discern the trends of each word from decade to decade. To better assist with analysis, in Figure 7, we also show a visualization of the top 10 words at decade intervals, linking their movement in rank over time. As the figure shows, the topic evolution is quite noisy over time. There are 45 words that occur in the top 10 most representative words over the decades. Furthermore, many of the earlier years have only a few documents associated with them, causing the LDA pillar to be overrepresented relative to the rest of the model. To reduce the amount of inconsistency in the DTM, we tried instead to bucket documents into decades.

8.4. DTM: Decade-Granularity

We now discuss a time-bucketing DTM. Each document was grouped into the next lowest decade from 1950 onward. The resultant document frequency distribution is now shown in Figure 6.

1	3	5	8	10	11	12	15	17	19
inmates	school	trial	voting	city	her	state	state	tex	discrimination
have	students	his	district	housing	child	federal	public	texas	title
prison	education	he	county	plaintiffs	children	act	us	s.w.2d	evidence
inmate	state	jury	black	its	she	funds	we	no	her
defendants	program	evidence	election	have	his	resources	religious	trial	see

Table 2. Top 5 Words from selected topics for static LDA model

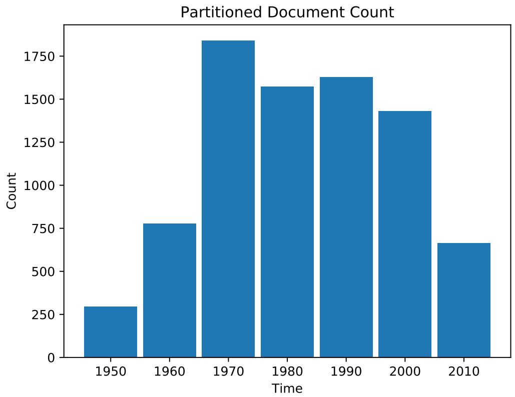


Figure 6. DTM Partition Histogram

With better document balance and a decrease in the number of LDAs learned, our results were much smoother. Figure 8 shows the words in 4 follows much smoother interpolation between decades. For topic 2, we see that the most representative words, *state* and *we*, dominate for the entire time range. However, for lower-ranked words, we see a general trend upward in importance for the words *constitutional* and *constitution*, as well as a slight decrease in the importance of the word *act*.

With the decrease in vocabulary size and aggregation of datapoints into larger time buckets, however, it seems that the diversity and specificity of representative words for different topics seems to have decreased. We find that most topics found from the partitioned DTM share similar most popular words.

From these results, we see that the dynamic topic model tracks changes in word-topic distribution over time in a reasonable way.

9. Discussion

From the results, we see that our dynamic topic model implementation yields reasonable results. Here, we discuss the successes and shortcomings of this project as well as introduce further work to improve the results.

9.1. Qualitative Verification

In addition to comparing the top words in each topic distribution, we also took some time to verify that the topic distributions for each document were reasonable. In our informal sampling, we found that the documents we read did tend to have appropriate representation of the most significant words in the document's most important topics.

9.2. Computational Constraints

The final dynamic topic model tended to yield similar representative words over many of the topics. We suspect this is due to the significantly truncated vocabulary used in the final model. Again, because of the cubic increase in time needed to Kalman smooth, it quickly became intractable to fit a larger model. Given enough computation power and a graphics card-enabled implementation, the model would be more reasonable to test on larger vocabularies.

9.3. C Implementation

While browsing for the existence of other implementations of the dynamic topic model, we were only able to find one implementation by Blei [8] himself. The dynamic topic model, written in C and GPU-enabled, is likely to be a more stable solution than our proposed Python implementation. As far as we know, there does not seem to be a Python-native implementation of the DTM, likely due to the issues of speed that make this model difficult to train in Python.

9.4. Theoretical Contributions

In contrast with black-box strategies and straightforward applications of well-discussed models to domain specific problems, our major challenge in the project was to understand the mathematical underpinnings behind a relatively complex, unsupervised machine learning model. While Latent Dirichlet Allocation is a popular strategy for topic modeling, much of the details regarding Variational Bayes EM and ELBO optimization are lost in application. In this project, we had the opportunity to incorporate many of the topics covered near the end of the course in an easy-to-understand LDA implementation and link it to Kalman Filters in order to introduce dynamics of time into our topic model. Due to

1950	1960	1970	1980	1990	2000	2010	2017
city	school	he	inmates	inmates	have	he	he
no	state	school	prison	have	state	defendant	defendant
have	schools	no	have	prison	defendants	his	his
there	public	have	inmate	defendants	school	trial	trial
case	act	were	he	has	its	evidence	evidence

Table 3. Top 5 words over time for topic 1, year-granularity DTM

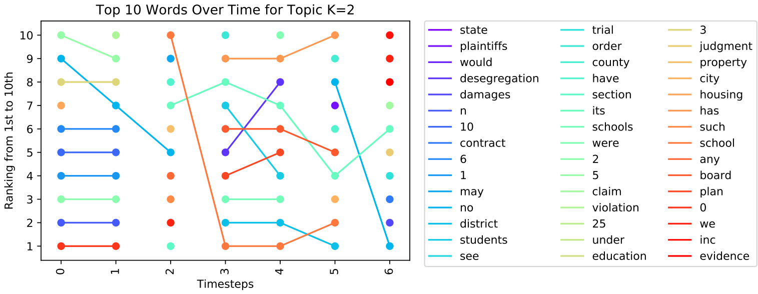


Figure 7. DTM word ranking plots for year-granularity

1950	1960	1970	1980	1990	2000	2010
state	state	state	state	state	state	state
we	we	we	we	we	we	we
act	section	section	section	section	section	section
section	act	act	act	law	law	law
constitution	law	law	law	act	act	act

Table 4. Top 5 words over time for topic 2, decade-granularity

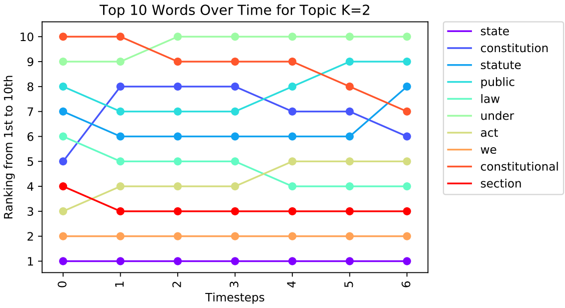


Figure 8. DTM word ranking plots for decade-granularity

its complexity and the lack of mathematical clarity in Blei's original paper, there are few attempts at an entire implementation of the dynamic topic model, so we are excited to have tried to decipher and reconstruct the original findings.

10. Conclusion

To conclude, we created a Python implementation of dynamic topic models and applied it to a text corpus of United States federal and state court opinions on school desegregation. This implementation uses a variational Bayes EM approach to maximizing the ELBO between a posited variational distribution and the true distribution of the model.

The results seem to indicate success in the dynamic topic implementation. Topics generated by the model seem to group reasonably together in the static model, and extension to the dynamic model shows smooth evolution over time when bucketed time ranges are introduced. Overall, it seemed that topics did not vary that highly over time. In order to conclusively say that arguments about school segregation have not changed substantially since Brown v. Board of Education in 1954, however, it would be necessary to run our models with a larger vocabulary (using greater computational resources,) to experiment further with varying the imposed level of smoothness between topics determined by the hyperparameters of the Kalman filtering, and to do more followup on specific topics of court cases with domain experts to verify the consistency of our topic modeling with general knowledge in the legal domain.

Overall, analyzing unlabeled data is still a relatively challenging task in machine learning. As methods in modeling and summarizing unlabeled data improve, we will be better able to interpret and control the topics being generated by various topic models. Until then, more work needs to be done on detailing the implementation and statistical theory behind the dynamics of topic models.

References

[1] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, and A. Jaimes, "Sensing trending topics in twitter," eng, *Multimedia, IEEE Transactions on*, vol. 15, no. 6, pp. 1268–1282, Oct. 2013.

[2] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06, Pittsburgh, Pennsylvania, USA: ACM, 2006, pp. 113–120, ISBN: 1-59593-383-2. DOI: 10.1145/1143844.1143859. [Online]. Available: http://doi.acm.org/10.1145/1143844.1143859.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, Jan. 2003.

[4] B. Bonikowski and P. DiMaggio, "Varieties of american popular nationalism," *American Sociological Review*, vol. 81, no. 5, pp. 949–980, Oct. 2016, ISSN: 0003-1224.

[5] *CourtListener*. [Online]. Available: https://www.courtlistener.com/.

[6] C. Geigle, *Inference methods for latent dirichlet allocation*, Oct. 2016.

[7] K. P. Murphy, *Machine Learning, A Probabilistic Perspective*. MIT Press, 2012, ISBN: 978-0-262-01802-9.

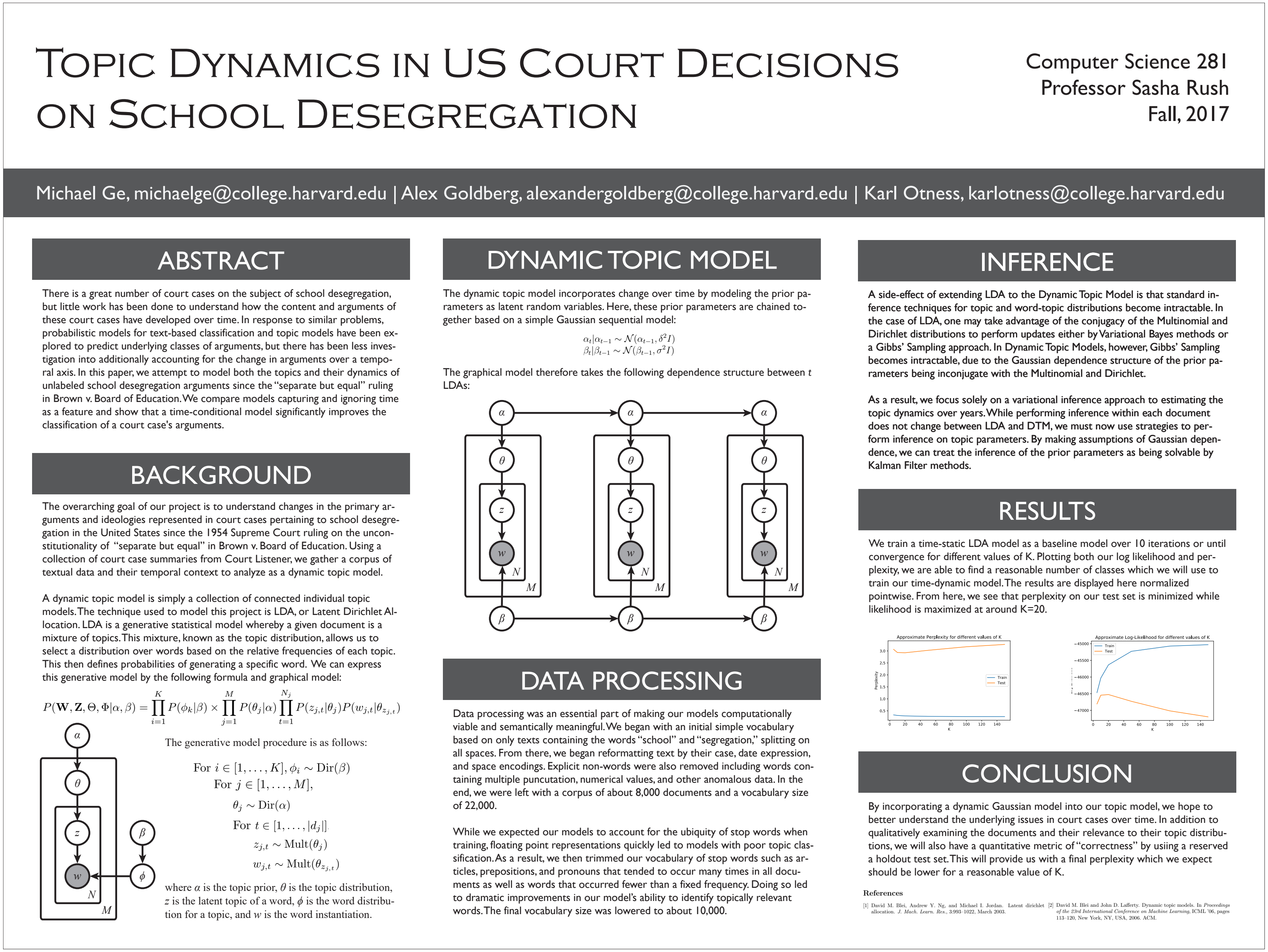
[8] *Princeton statistical learning dtm implementation*. [Online]. Available: https://code.google.com/archive/p/princeton-statistical-learning/downloads.

[9] *PyKalman version 0.9.5*. [Online]. Available: https://pykalman.github.io/.

[10] A. Rule, J.-P. Cointet, and P. S. Bearman, "Lexical shifts, substantive changes, and continuity in state of the union discourse 1790-2014," *PNAS*, vol. 112, no. 35, 2015. DOI: 10.1073/pnas.1512221112.

[11] J. B. Slapin and S.-O. Proksch, "A scaling model for estimating time-series party positions from texts," *American Journal of Political Science*, vol. 52, no. 3, pp. 705–722, 2008, ISSN: 1540-5907. DOI: 10.1111/j.1540-5907.2008.00338.x. [Online]. Available: http://dx.doi.org/10.1111/j.1540-5907.2008.00338.x.

A. Poster



Topic Dynamics in US Court Decisions on School Desegregation

Computer Science 281
Professor Sasha Rush
Fall, 2017

Michael Ge, michaelge@college.harvard.edu | Alex Goldberg, alexandergoldberg@college.harvard.edu | Karl Otness, karlotness@college.harvard.edu

ABSTRACT

There is a great number of court cases on the subject of school desegregation, but little work has been done to understand how the content and arguments of these court cases have developed over time. In response to similar problems, probabilistic models for text-based classification and topic models have been explored to predict underlying classes of arguments, but there has been less investigation into additionally accounting for the change in arguments over a temporal axis. In this paper, we attempt to model both the topics and their dynamics of unlabeled school desegregation arguments since the "separate but equal" ruling in Brown v. Board of Education. We compare models capturing and ignoring time as a feature and show that a time-conditional model significantly improves the classification of a court case's arguments.

BACKGROUND

The overarching goal of our project is to understand changes in the primary arguments and ideologies represented in court cases pertaining to school desegregation in the United States since the 1954 Supreme Court ruling on the unconstitutionality of "separate but equal" in Brown v. Board of Education. Using a collection of court case summaries from Court Listener, we gather a corpus of textual data and their temporal context to analyze as a dynamic topic model.

A dynamic topic model is simply a collection of connected individual topic models. The technique used to model this project is LDA, or Latent Dirichlet Allocation. LDA is a generative statistical model whereby a given document is a mixture of topics. This mixture, known as the topic distribution, allows us to select a distribution over words based on the relative frequencies of each topic. This then defines probabilities of generating a specific word. We can express this generative model by the following formula and graphical model:

$$P(\mathbf{W}, \mathbf{Z}, \Theta, \Phi | \alpha, \beta) = \prod_{i=1}^{K} P(\phi_k|\beta) \times \prod_{j=1}^{M} P(\theta_j|\alpha) \prod_{t=1}^{N_j} P(z_{j,t}|\theta_j) P(w_{j,t}|\theta_{z_{j,t}})$$

The generative model procedure is as follows:

```
For $i \in [1, \dots, K], \phi_i \sim \mathrm{Dir}(\beta)$
    For $j \in [1, \dots, M]$,
        $\theta_j \sim \mathrm{Dir}(\alpha)$
        For $t \in [1, \dots, |d_j|]$.
            $z_{j,t} \sim \mathrm{Mult}(\theta_j)$
            $w_{j,t} \sim \mathrm{Mult}(\theta_{z_{j,t}})$
```

where α is the topic prior, θ is the topic distribution, z is the latent topic of a word, ϕ is the word distribution for a topic, and w is the word instantiation.

DYNAMIC TOPIC MODEL

The dynamic topic model incorporates change over time by modeling the prior parameters as latent random variables. Here, these prior parameters are chained together based on a simple Gaussian sequential model:

$$\alpha_t | \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$$
$$\beta_t | \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$$

The graphical model therefore takes the following dependence structure between t LDAs:

DATA PROCESSING

Data processing was an essential part of making our models computationally viable and semantically meaningful. We began with an initial simple vocabulary based on only texts containing the words "school" and "segregation," splitting on all spaces. From there, we began reformatting text by their case, date expression, and space encodings. Explicit non-words were also removed including words containing multiple puncutation, numerical values, and other anomalous data. In the end, we were left with a corpus of about 8,000 documents and a vocabulary size of 22,000.

While we expected our models to account for the ubiquity of stop words when training, floating point representations quickly led to models with poor topic classification. As a result, we then trimmed our vocabulary of stop words such as articles, prepositions, and pronouns that tended to occur many times in all documents as well as words that occurred fewer than a fixed frequency. Doing so led to dramatic improvements in our model's ability to identify topically relevant words. The final vocabulary size was lowered to about 10,000.

INFERENCE

A side-effect of extending LDA to the Dynamic Topic Model is that standard inference techniques for topic and word-topic distributions become intractable. In the case of LDA, one may take advantage of the conjugacy of the Multinomial and Dirichlet distributions to perform updates either by Variational Bayes methods or a Gibbs' Sampling approach. In Dynamic Topic Models, however, Gibbs' Sampling becomes intractable, due to the Gaussian dependence structure of the prior parameters being inconjugate with the Multinomial and Dirichlet.

As a result, we focus solely on a variational inference approach to estimating the topic dynamics over years. While performing inference within each document does not change between LDA and DTM, we must now use strategies to perform inference on topic parameters. By making assumptions of Gaussian dependence, we can treat the inference of the prior parameters as being solvable by Kalman Filter methods.

RESULTS

We train a time-static LDA model as a baseline model over 10 iterations or until convergence for different values of K. Plotting both our log likelihood and perplexity, we are able to find a reasonable number of classes which we will use to train our time-dynamic model. The results are displayed here normalized pointwise. From here, we see that perplexity on our test set is minimized while likelihood is maximized at around K=20.

CONCLUSION

By incorporating a dynamic Gaussian model into our topic model, we hope to better understand the underlying issues in court cases over time. In addition to qualitatively examining the documents and their relevance to their topic distributions, we will also have a quantitative metric of "correctness" by usiing a reserved a holdout test set. This will provide us with a final perplexity which we expect should be lower for a reasonable value of K.

References

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

[2] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM.

B. Word Cloud

Word clouds are an aesthetically pleasing yet ineffective way of displaying word topic distributions. See Figure 9 for an example.

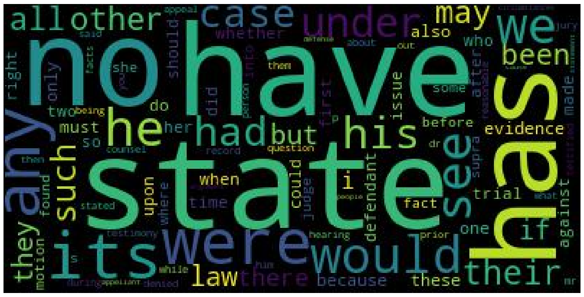


Figure 9. Word cloud for topic 2 in Partitioned DTM