Computational Text Analysis for Public Management Research: An Annotated Application to County Budgets

L. Jason Anastasopoulos* ljanastas@uga.edu

Tima T. Moldogaziev† timatm@uga.edu

Tyler A. Scott‡ tascott@ucdavis.edu

September 20, 2017

Abstract

Organizations produce copious volumes of written documents, including position papers, meeting summaries, minutes from hearings, presentations, and budget justifications. These documents present a wealth of untapped information, which can shed light on a variety of organizational factors–individual and group behaviors, managerial and policy choices, and other key inter- and intra-organizational dynamics that are of great interest to public management scholars. Computational text analysis methods offer a highly generalizable means of tapping into these documents in order to generate objective organizational data. We demonstrate a general method for analyzing public texts by applying the Latent Dirichlet Allocation (LDA) approach to measuring budget orientations in county budget documents. LDA is a nonparametric Bayesian method, which is used to extract topical content from collections of documents. We demonstrate how this method can be utilized to measure the functions of

*Assistant Professor, Department of Public Administration and Policy, Department of Political Science, University of Georgia. Microsoft Visiting Professor, Center for Information Technology Policy, Princeton University (2017–2018).

†Assistant Professor, Department of Public Administration and Policy, University of Georgia

‡Assistant Professor, Department of Environmental Science and Policy, University of California, Davis

budgets in county budget narratives in the state of California, highlighting both within– and between– county variation. This annotated computational analysis of documents is an example of how machine-learning techniques can greatly enhance longitudinal, comparative studies in public management and governance research.

Keywords; machine learning, topic model, public document, budget function

1 Introduction

Public sector organizations produce and retain a trove of procedural documents, either in physical locations, electronic depositories, or both. Documents such as position papers, meeting agendas, minutes from hearings, plans, and budget statements–which one might collectively refer to as "procedural documents"–present a wealth of information about the structure and function of an organization. Accordingly, students of organizational analysis have long relied upon procedural documents to tell important stories about individual and group behaviors, goal and agenda setting, and other essential dynamics both within and between organizations. Reading and systematically coding procedural documents takes a great deal of time, effort and specialized expertise. Computational limitations have rendered large scale, systematic analysis of these documents virtually impossible. Advanced computational text analysis methods offer a means of tapping into these documents in order to generate objective organizational data by replacing methods that rely primarily on expert human coders (Morris, 1994).

Such techniques go beyond simple exercises of word counting, with current applications in social sciences venturing to more complex tasks such as extracting particular parts of speech or identifying commonalities in different texts (Lowe et al., 2009; Wiedemann, 2013; Grimmer and Stewart, 2013). For instance, several recent studies have evaluated the policy positions of legislatures in the US, Europe, and elsewhere (Lowe et al., 2009; Greene and Cross, 2017; Laver, Benoit, and Garry, 2003). Further, Pandey, Pandey, and Miller (2017) offered a natural language processing technique for public management researchers to measure innovativeness within pub-

lic organizations and compared them to expert assessments. These types of applications have the potential to transform–and radically expand–how we study public sector organizations. To be clear, computational analysis–and in particular, the text mining applications that we focus on in this paper–are not a substitute for proper theory development or valid research design. Rather, these techniques enhance the methodological capabilities of public management scholars and open new horizons for their research productivity. The large volume and extensive variety of available documents and the technological capacities of computational text analysis offer a way to greatly expand the breadth and scope of data inputs for governance research. Of course, text mining is not a panacea and the promises of computational document analysis are not without pitfalls (Grimmer and Stewart, 2013). Popping (2012), for instance, warns that there always remains a need for a qualitative assessment of quantitative data to ensure validity. Apropos of this concern, computational techniques for document analysis are rightly considered a highly useful–and potentially transformative–tool (Grimmelikhuijsen, Tummers, and Pandey, 2017), but one that requires a careful adaptation and application to the context of governance research. Therefore, it is clear that the time is ripe for a methodological guide on machine learning techniques for computational analysis of documents for public management and governance research.

As an example of how text analysis integrates with public management theory, our annotated application uses a database of proposed budgets by California counties for fiscal years 2012 through 2017. During these five fiscal years, county executive/administrative officers (CEO) have prepared and proposed budget documents in

excess of 125,000 pages. Schick (1966) seminal article identifies three main functions, or orientations, of these types of budgetary documents. Empirical studies applied this framework and showed that budgets can be used for the purposes of control, management, or planning by state and city governments (Daley, 1985; Friedman, 1975). One of the principal concerns necessary for understanding different types of budgetary systems are "whether greater emphasis is placed at the central levels on planning, management or control" Schick (1966). In the recommended budget documents from California counties, text generally comes in two main and distinct forms. The first form is a macro–narrative, often an overarching summary and message that the county CEO wishes to communicate. The second is a micro–narrative that accompanies individual functional departments, such as administration office, public works, fire, and so on. These micro–narratives often concern messages that are relevant in a given narrow area. The macro and micro messages may be accompanied by generic information about counties and their history, financial and economic reports, and actual budget numbers. For the purposes of this particular annotated application, we single out macro- and micro- budgetary narratives and utilize the study of Schick (1966) as a theoretical guide for validating topics extracted through computational document analysis. We demonstrate how this method can be used to measure the functions of county budget narratives in the state of California, highlighting both within– and between–county variations along Schick's spectrum. In the next section we provide a brief introduction to the two major components of computational text analysis methods: natural language processing and machine learning techniques. We then proceed to the discussion of the annotated application

to budget documents.

2 Overview of Natural Language Processing and Machine Learning for Text Analysis

2.1 Natural Language Processing

The field of natural language processing, or NLP for short, is fundamentally concerned with representing natural language units (words, sentences, paragraphs, etc.) in a form that can be understood by and manipulated through a machine for the purpose of systematically analyzing large *corpora* or groups of documents. Once the features of documents can be represented on a machine, primarily as numerical values, all of the computational and statistical methods which can be applied to numerical data can then also, in theory, be applied to texts. Natural language processing methods have been used with a high degree of success to solve a variety of practical problems including email filtering, speech recognition systems, search engines (Google, Bing, etc.) and even artificial intelligence systems. In the social sciences, natural language processing methods combined with machine learning techniques have been used to estimate media bias (Young and Soroka, 2012), identify the politically relevant features of texts (Barberá, 2014; Bond and Messing, 2015; Lowe et al., 2011; Monroe, Colaresi, and Quinn, 2008; **?**) and to measure agendas in political texts (Grimmer, 2009).

2.1.1 Terminology

In natural language processing, the most basic unit of analysis is the "term." Terms can be of two types: *words* or *n–grams*. Words in NLP are the same as they are in common usage while n–grams are typically two or more words that are treated as a single unit. For example, names of places such as "New York", "Los Angeles", "Ohio State University" etc. are n–grams that might be treated as a single term for purposes of analysis. *Documents* are comprised of groups of terms and are typically the main unit of analysis when analyzing texts. Documents can be anything from sentences and paragraphs to entire works of literature. Here, the documents that we are analyzing are macro- and micro- budget narratives or statement from each California county for fiscal years 2012–2017. Finally, a *corpus* is a collection of documents that are analyzed. A corpus can be thought of as the equivalent to a "data set." In our case, the single corpus that we are analyzing is the *set* of budget statements from California counties. Equation 1 provides an overview of the NLP processing hierarchy. Here we see that the corpus is comprised of multiple documents, which in turn are each comprised of terms. When it comes to analyzing data, documents are almost always the unit of analysis.

$$terms \subseteq document \subseteq corpus \tag{1}$$

2.2 From Text to Data

The process of converting text to data typically involves four major steps whose final goal is the creation of a *Document–Term Matrix* which will be described in

Tokenization → *Formatting* → *Stop words* → *Stemming*

Figure 1: Example of a pre–processing text–analysis pipeline

more detail in the next section. To prepare the texts for analysis, they must be pre–processed to ensure that they have comparable terms across documents and also that they do not include extraneous information which can introduce noise in the subsequent analyses. Figure 1 presents an overview of a common text to data pre–processing steps also referred to as a "pipeline". *Tokenization* is the process of dividing a document into its constituent terms as discussed above. Documents are almost always represented as strings and must be *tokenized* in order to identify the terms. The default tokenization algorithms of most software packages divide documents into words rather than n-grams.

After documents are tokenized, they must be formatted so that all words are in lowercase and marks such as punctuation, etc., which are not relevant to the analyses, are removed. Formatting provides us with a standardized means of comparing terms in documents. After formatting, two processes known as *stop word removal* and *stemming* are used to further increase the signal–to–noise ratio when comparing documents.

Stop words are the most common words of a language. In English, stop words tend to be words like "the", "who", "them", "they", etc. (Wilbur and Sirotkin, 1992). While there is no universally accepted list of stop words, all natural language processing software includes a standard list of the most commonly used words in English and other commonly used languages. The final pre–processing step in converting text to

data is known as *stemming*. Stemming is the process of reducing words to their roots, which typically involves removing suffixes from all of the words in each document. For example, the suffixes "–ing", "–s", "–ed" in the words "governing", "governs" and "governed" would all be removed after stemming and these words would be in a document would all be reduced to their root stem "govern–." Stemming often is useful as a pre–processing step, especially in the area of unsupervised learning, because it tends to increase the interpretability of the estimated clusters (Collobert et al., 2011). This, however, is not always the case and it is often worthwhile to explore results with and without stemming (Lai and Tsai, 2004). For further exposition, we provide the reader with code which implements each of these pre–processing steps in the **R** statistical environment.

2.3 The Document–Term Matrix

After text pre–processing, the final step before preparing the text data for analysis is the transformation of the terms and documents in a corpus into a *document–term matrix*. As the name suggests, the document–term matrix is a matrix whose rows are documents and whose columns are terms. The entries of the matrix correspond to the documents' *term frequency*, or the number of times each term appears in each document. Formally, if the total number of documents in a corpus are represented by S and the total number of terms or words in the corpus are represented by T, a document–term matrix is always a matrix $\mathcal{D}$ with the following properties:

$$\mathcal{D} \in \mathbb{R}_{+}^{SxT} \tag{2}$$

$$T \geq S \tag{3}$$

From the equations above, we can see that a document term matrix has entries corresponding to the positive real numbers and is always of dimension SxT. It is also clear that the number of terms T is always greater than or equal to the number of documents S. Indeed, in most cases the number of terms is significantly larger than the number of documents, even after accounting for all of the pre–processing steps mentioned above. Figure 2 is a sample document–term matrix.

$$\mathcal{D} = \begin{array}{c} \\ Document_1 \\ Document_2 \\ Document_3 \\ Document_4 \\ Document_5 \\ \vdots \\ Document_S \end{array} \begin{array}{c} \begin{array}{ccccccc} Term_1 & Term_2 & Term_3 & Term_4 & Term_5 & \cdots & Term_T \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & 1 & \cdots & 0 \end{pmatrix} \end{array}$$

Figure 2: A sample document–term matrix (dtm) for a corpus. The dtm contains S documents and T terms and elements of the matrix are term–frequencies, or the number of times that each term appears in the document.

There are two implications of Equation 3 which should be noted. First, entries of the document–term matrix $\mathcal{D}$ need not only be text frequencies. Indeed, a weighting method known as text frequency–inverse document frequency (TF–IDF) is commonly

used as a means of weighting text frequencies especially in the supervised machine learning context. Second, because there are generally far more terms than there are documents, the majority of entries in the document–term matrix tend to be 0's. In linear algebra, this property of document–term matrices is referred to as *sparsity*. Matrix sparsity poses unique challenges for prediction with statistical models and there are many instances in text analysis where sparsity reduction is desirable to improve model performance (Tibshirani et al., 2005; Zhang and Huang, 2008).

2.4 Text Analysis with Machine Learning Algorithms

Natural language processing methods provide the necessary background for understanding how texts can be converted to numerical data to prepare them for analysis. But to understand how we can use text data to learn more about public officials and their decision-making processes, we must first understand some of the computational tools which make use of text data. Here we provide a very brief introduction to machine learning algorithms which can be used to analyze texts.

Perhaps the best way to broadly think about machine learning algorithms is that they are a set of statistical models and procedures whose ultimate goals are directed toward *prediction* rather than *inference*. For example, ordinary least squares (OLS) or logistic regression are often used in the context of public administration research to determine whether an independent variable X affects or is related to some dependent variable Y. In this case, the goal of parameter estimation for OLS or logistic regression is *statistical inference*, which, in ideal circumstances can give us information about the *causal effect* of X on Y. In the context of machine learning,

parameter estimation for OLS and logistic regression models is not done with the goal of statistical inference, but rather is done with the goal of prediction. Thus, OLS and logistic regression models can be thought of as *machine learning algorithms* if they are used to predict new values of Y using a particular set of predictors X.

In general, there are two types of machine learning algorithms: *supervised* and *unsupervised* algorithms. *Supervised algorithms* are models which use a set of independent variables or *features* to accurately predict a dependent variable which can either be a binary or multi–class label or a continuous measure. Examples of popular supervised machine learning algorithms for text analysis include sparse logistic regression, naive Bayes and support vector machines. When applied to text data in each case, each algorithm utilizes the *terms* (columns) in the document term matrix as features which are used to predict a document *label* or class. For example, consider the case where the "documents" are tweets containing mentions of federal agencies such as the FBI, NASA, etc. and we are interested in classifying them into one of two classes C: "Positive" + or "Negative' −. For sparse logistic regression, we would model the classifications using logistic regression with the term vectors (W) in the document–term matrix as independent variables:

$$logit(\mathbb{E}[C|W] = \theta_0 + \theta_1 w_1 + \cdots + \theta_n w_n + \epsilon$$

Since we are dealing with text data in which the number of predictors always outnumbers the number of observations, we would use *sparse logistic regression* which is simply logistic regression with a *regularization parameter* also known as the LASSO which penalizes variables that do not contribute to predicting the outcome. When

dealing with high–dimensional problems in which there are a large number of covariates, LASSO methods generally reduce mean squared error and classification error and in most contexts and allow for parameter estimation in high dimensional spaces in which estimation would be intractable. Thus, while in ordinary logistic regression we would estimate parameter values by minimizing the following loss function:

$$\arg\min_{\theta} \sum_{i=1}^{n} - \left[D_i \theta^T T_i - log(1 + exp(\theta^T T_i)) \right] \tag{4}$$

for *sparse logistic regression* we estimate parameters using a loss function similar to the one above but with a ℓ_1 regularization norm[1] which penalizes or shrinks the parameter values in the document term matrix to zero by λ:

$$\arg\min_{\theta} \sum_{i=1}^{n} - \left[D_i \theta^T T_i - log(1 + exp(\theta^T T_i)) \right] + \lambda ||\theta||_1 \tag{5}$$

After estimating the parameters in this equation using the *training* data, which would consist of labeled tweets in this case, we would then apply the algorithm to the *test* data to see how well the model performs when it classifies new data. In this case, the classification criteria for $k = \{+, -\}$ for tweet i is:

$$C_{ki} = \arg\max_{k} P(C_i = k | W_i)$$

Unsupervised algorithms are models which do not use class labels to classify data, but rather automatically generate groups or clusters from independent variables (fea-

[1]There are several choices of regularization norms for the lasso but the ℓ_1 has typically been found to produce the best solutions with the lowest mean squared error in a number of supervised machine learning problems.

tures). Unsupervised algorithms include k–means clustering, hierarchical clustering, principal components analysis, multidimensional scaling and the Latent Dirichlet Allocation (LDA) of which there are a number of variants[2]. In the context of text analysis, unsupervised learning methods compute the similarities between documents using the document–term matrix based on some distance metric.

Here, we propose the use of LDA, an unsupervised algorithm, as a general method for measuring the budget orientations of institutional budget documents. LDA is a nonparametric Bayesian method and a feature of the LDA approach, which makes it an excellent method for analyzing texts, is that it has the ability to cluster texts and also provides information about how to *interpret* the clusters (Blei and Lafferty, 2007). For this reason, LDA is often colloquially referred to as a "topic model" because it scales text into a multidimensional set of topics that reflect underlying document themes (Blei and Lafferty, 2007; Grimmer, 2009). These unique features also allow us to identify theoretically relevant aspects of institutional documents as we do with budget statements below. In what follows, we focus primarily on the LDA method, but do emphasize that it is a specific example of a more general category of unsupervised classification methods.

[2]These include: dynamic topic models (Blei and Lafferty, 2006; Wang, Blei, and Heckerman, 2012) to analyze the evolution of topics over time and structural topic models which incorporate topic metadata for enhanced interpretability (Roberts et al., 2014)

3 Measuring the Functions of Budgets in California Counties with Computational Text Analysis

3.1 Overview

Before describing the topic modeling process in more technical detail, it is useful to first understand what the topic model does. The topic model allows us to do two things with texts. First, it allows us to get a sense of the common latent thematic elements across a corpus or collection of documents. For example, in our corpus of budget statements the topic model would be able to tell us what thematic aspects are common across budget statements. In addition to this, topic models allow us to measure how much of each topic is contained within each document. Suppose we discover that the common themes across our corpus of budget statements are related to Schick (1966)'s analytical elements: control, planning and management; we then would conceptualize the corpus as having three underlying dimensions, or topics, that correspond to these budgetary functions.

For each county budget micro and macro statement in that corpus, then, the topic model would be able to tell us what proportion of each county budget in each year is devoted to control, planning and management functions (i.e., in essence how a given budget scores on each of the three topical dimensions). For example, an exploration of the distribution of topics in the Los Angeles County macro budget statement in 2012 might reveal that, in discussions of the Los Angeles County budget in that year, 20% of the text was devoted to discussions of control functions, 50% management and 30% to planning. Empirical evaluation of different dimensional reductions might further

reveal a more optimal number of topics; for instance, perhaps management content diverges starkly into management of public utilities and management of health and human services and is thus better represented by two different topics. In this way, topic modeling is much like exploratory factor analysis as a means of dimensionality reduction. We discuss methods for testing and selecting the number of topics further in the section to follow. In summary, the topic model gives us two major pieces of information for any collection of documents: (1) a number of topics which are contained within a corpus and; (2) for each document contained within the corpus, what proportion of each of the topics is contained within the each document. Below, we describe how the topic model treats texts from a more technical perspective.

3.2 Elements of Topic Models

As with all text analysis problems which we discussed above, the fundamental unit of data used in topic models are *terms* as represented in the document–term matrix. Terms are treated as items from a *vocabulary*, indexed by a set of numbers $\{1, ..., V\}$. The vocabulary are all of the terms in a given *corpus* or collection of documents as discussed above.

A *document* is a bag of N terms. We describe a document as a "bag of terms" rather than a series or sequence of terms in a particular order because the topic model does not take the order of terms or words into account. These N terms can be represented by a vector $\mathbf{w} = (w_1, w_2, ..., w_N)$. A *corpus*, as above, is a collection of M documents which can be represented by $D = \{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_M}\}$. The topic model treats each document within a corpus as a mixture of a fixed number of k latent

topics which is represented by a distribution over words.

3.3 A Topic Model of Budget Functions

As we discussed above, the LDA, with careful theoretical guidance and the appropriate data, can shed light on many theoretically relevant questions in the public administration and public management literature which have been limited by the inability to systematically analyze large quantities of documents. One of these questions relates to understanding variation in budget systems as they relate to each of the three major functions of budgets described by Schick (1966): planning, management and control. Specifically, we are interested whether macro- and micro- budget narratives place greater emphasis on planning, management or control (Schick, 1966). By modeling budget documents using the LDA, we are able to directly address this question both within and between organizations or counties in the State of California.

The three functions are defined as follows: *planning* functions relate to the goals and objectives of an organization; *management* functions relate to overarching strategic plans which are means of accomplishing the objectives of the organization; and *control* functions are the means of binding superiors to the strategic plans set out by the organizations. Different types of budgetary systems in organizations can be defined by the different levels of emphasis that they place on each of these three areas (Schick, 1966). For example, macro-narratives may emphasize overall control functions, while micro-narratives may present more planning and management orientations. Furthermore, some organizations may have incentives to focus primarily on control functions if they face significant fiscal and political constraints while other

organizations may find it necessary to focus primarily on broader goals vis–a–vis planning and less so on management and control orientations. Such variations in state and city budget functions are reported in prior research by Daley (1985) and Friedman (1975).

Unlike financial data, budget narratives can give us a window into which functions organizations seek to emphasize and how these emphases change over time. Earlier studies used surveys to ask budget directors of how they used their budgets along Schick's three analytical functions. In more recent studies that analyze text– a more direct approach to understanding organizations, traditionally, one would have research assistants or other coders read each of these documents and use their own personal judgment to figure out which of these three functions budgets emphasize and to what extent they emphasize them. There are two major issues with such an approach: time and expertise. First, time constraints make gathering and reading potentially hundreds or thousands of texts unfeasible, thus requiring that inferences be made on a small subset of budget statements which may not be representative. Second, systematic analysis of budgets requires the researcher to rely exclusively on coders and requires that these coders consistently apply the same method to classifying each document. Consistent application of methods among coders often requires a high degree of expertise and subject matter knowledge that is typically unavailable.

The topic model solves both of these problems simultaneously. Topic modeling can extract latent themes from thousands of budget documents in a matter of seconds in a systematic manner and it puts the power back in the hands of the skilled public

management expert to determine which of these latent themes relate to each of the functions of budget. Below we describe how budget statements are modeled with topic models and then move on to analyses of California counties budget statements between fiscal years 2012–2017

3.4 Modeling Budget Functions in California Counties

The essential first step toward modeling any set of texts using the LDA is division of these texts into *corpora* and documents. For our purposes, we define:

- **Document** - A budget statement a within a California county d, for fiscal year t which is represented as a sequence of N terms $\mathbf{w} = (w_1, w_2, ..., w_N)$.
- **Corpus** - The collection of all budget statements in $D = 58$ counties in California between fiscal years 2012–2017. This corpus includes 96 budget macro statements and 76 budget micro statements.

Figure 4 is a plot of the most frequent words found in the California budget macro–statements between 2012–2017. The most frequent stemmed term is "fund" which refers to a number of elements of macro statements including "funding" for various projects and the "general fund".

The LDA is a generative probabilistic model of the corpus of budget documents treated as a random mixture over k latent topics. Each topic as a distribution over the terms. But how do we know how many topics a corpus contains? Because the LDA does not automatically select the number of topics that a corpus is comprised of, the researcher must decide how many topics the corpus is comprised of on the

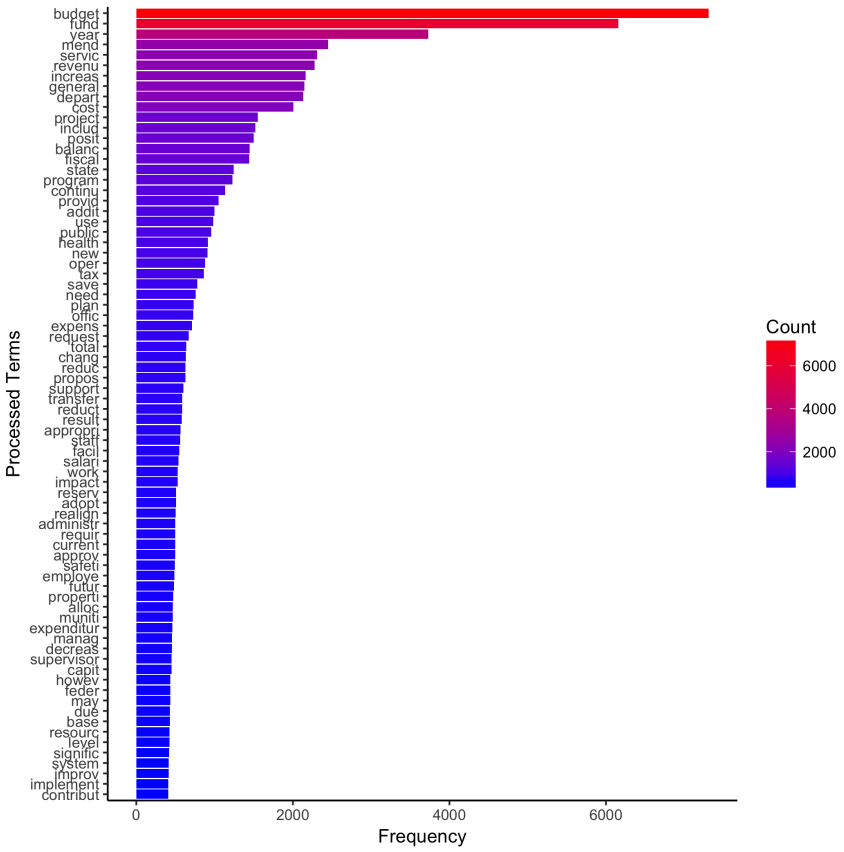


Figure 3: Most frequent terms found in the processed California county budget macro–statements, 2012–2017. The most frequent stemmed term is "budget" followed by "fund" and "year" suggesting that macro statements tend to discuss the current state of the budget.

basis of a number of factors. In most cases, theoretical guidance provided by expert knowledge combined with human interpretability should always serve as the first guide.

In addition to this, another popular method for topic selection involves estimating several topic models using $k = \{2, \cdots, n\}$ topics, measuring the perplexity of each model and choosing the model for which the marginal perplexity stops de-

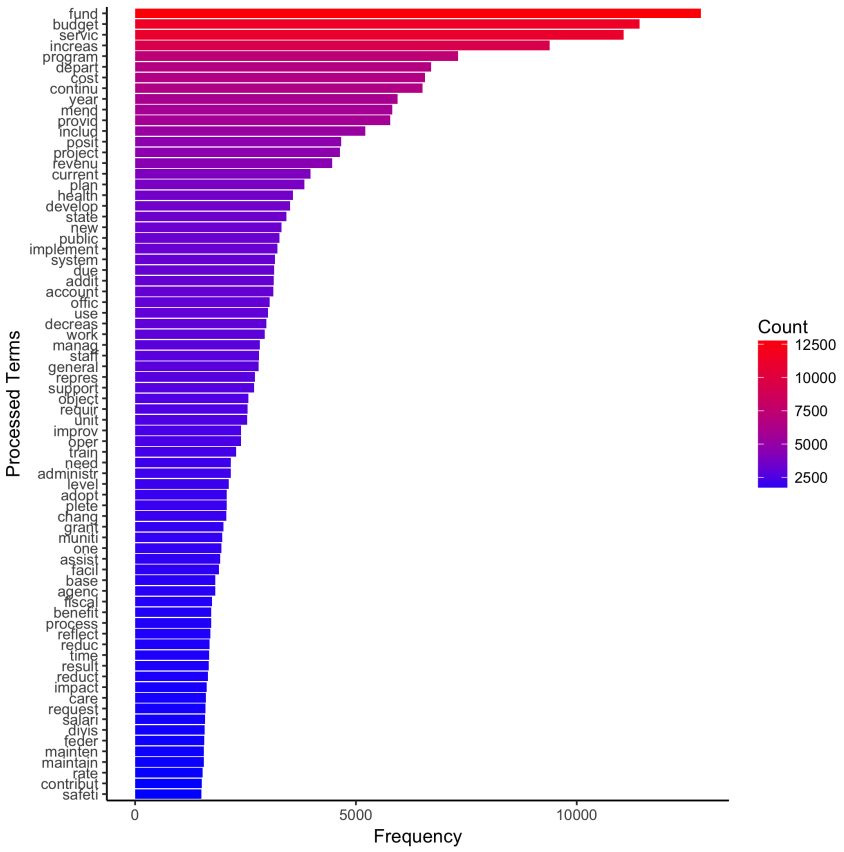


Figure 4: Most frequent terms found in the processed California county budget micro–statements, 2012–2017. The most frequent stemmed terms are "fund", "budget" and "servic" suggesting that micro statements tend to discuss funding for particular programs and services.

creasing (Blei and Lafferty, 2007; He et al., 2013; Hinton and Salakhutdinov, 2009). Perplexity is an information theoretic metric which measures how well probability models predict a sample which we describe in further detail below. Lower values of perplexity imply models that better fit the data. While perplexity often provides a good means of guiding researchers, many argue that it should only be used as a guide rather than the sole means of choosing the appropriate number of topics.

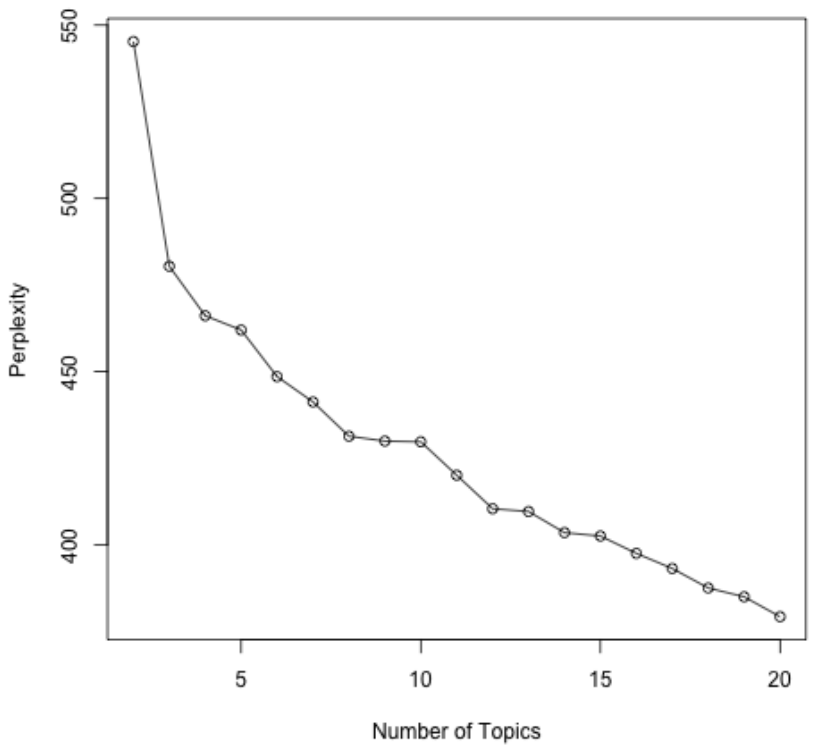


Figure 5: Perplexity for $k = 2, \cdots, 20$ topic models of California county budget macro statements, 2012–2017.

Accordingly, we use both theoretical guidance and perplexity as a guide to determine the number of topics present in our corpus. Looking at figures 5 and 6 we see that perplexity continues to improve after a roughly 5–topic model, but by very little after that. From a theoretical standpoint this makes perfect sense since we do not expect the number of topics to be significantly greater than the three budgetary analytical functions as defined by (Schick, 1966).

Thus, the first cut of this data involved setting $k = 5$ for budger statements.

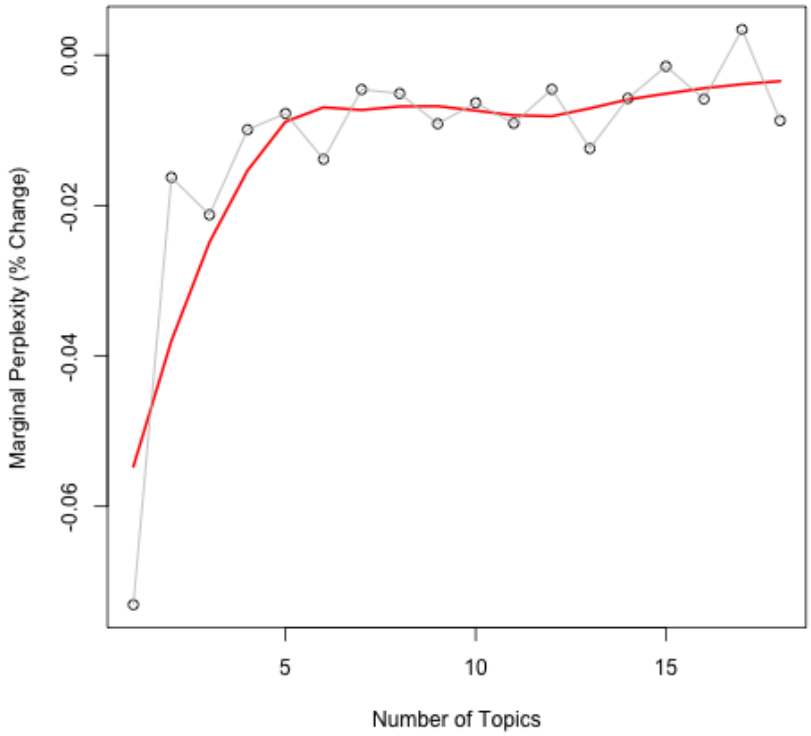


Figure 6: % change in perplexity for $k = 2, \cdots, 20$ topic models.

In words, this implies that there are a total of 5 latent topics contained within the corpus of the 96 budget macro statements and 76 micro statements, and that each of these macro–statements is a mixture over these latent topics. The proportion of the budget statement devoted to each topic is represented by θ_{ad}, the distribution of topic proportions for each budget statement.

Figure 7 is a graphical model of the LDA as applied to the California budget statements using plate notation to denote replicates of the budgets D and the terms within each budget N. Each of the nodes represents a random variable. The only

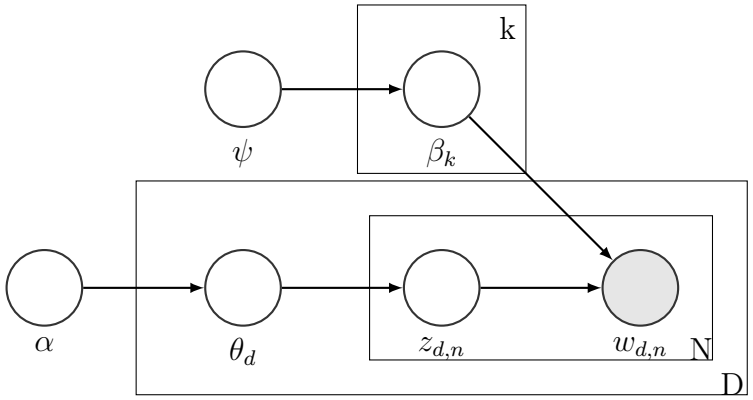


Figure 7: Graphical representation of the LDA applied to California budget documents using plate notation

observed variable is the collection of terms (budget macro statements for each county–year) which comprise the corpus. All other variables are unobserved latent variables which are estimated by the LDA. The graphical model above assumes that $w_{d,n}$, each term in each budget document in the corpus is generated from both a distribution over latent topics, which corresponds to each budget statement and a distribution over words which comprises the topic.

We define:

1. $\beta_k \sim Dir(\psi)$, where $k \in \{1, ..., 5\}$ - the distribution over words that defines each of the $k = 5$ latent topics assumed to the California budget statements between 2012–2017.

2. $\theta_d \sim Dir(\alpha)$, where $d \in \{1, ..., N_{st}\}$ – the distribution over topics for each budget county–year. For micro–statements $N_{st} = 76$, for macro–statements $N_{st} = 96$.

3. $z_{d,n}$ - topic assignment of the n^{th} word in the d^{th} budget statement.

4. $w_{d,n}$ - the n^{th} word of the d^{th} budget statement.

The probability distributions of topic proportions for each budget statement $p(\theta_d|\alpha)$ and of each topic in all budget statements $p(\beta_k|\psi)$ are distributed Dirichlet with hyperparameters α and ψ respectively. Thus, topic proportions for each budget macro–statement in a California county–year d has the distribution:

$$p(\theta_d|\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma\big(\sum_{i=1}^{K} \alpha_i\big)} \prod_{i=1}^{5} \theta_{di}^{\alpha_i - 1}$$

And each topic k across all articles has the distribution over words:

$$p(\beta_k|\psi) = \frac{\prod_{i=1}^{N} \Gamma(\psi_i)}{\Gamma\big(\sum_{i=1}^{N} \psi_i\big)} \prod_{i=1}^{N} \beta_{ki}^{\psi_i - 1}$$

The remaining distributions that we need in order to specify the model including topic assignment conditional on topic distribution $p(z_{d,n}|\theta_d)$ and word conditional on topic assignment $p(w_{d,n}|z_{d,n}, \beta_k)$ are multinomial with:

$$z_{d,n} \sim Multinom(\theta_d) \tag{6}$$

$$w_{d,n} \sim Multinom(\beta_k) \tag{7}$$

Putting all this together, we arrive at the fully specified model over all California budget macro–statements collected for fiscal years 2012–2017:

$$p(\theta, \mathbf{z}, \mathbf{w}, \beta|\psi, \alpha) = \prod_{k=1}^{5} p(\beta_k|\psi) \prod_{d=1}^{D} \left(p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta_k) \right) \quad (8)$$

Estimating $p(\theta_d|\alpha)$, which we use for understanding how planning, management and control functions as defined by Schick (1966) vary within counties over time and all other relevant hidden parameters requires posterior inference using the variational expectation-maximization algorithm (VEM) algorithm (Blei and Lafferty, 2007) which is implemented in **R** packages such as *topicmodels* and *lda*. The same routine is then repeated for 76 micro budget narratives.

3.5 Results

Here, we demonstrate how the LDA can be used to explore emphases on planning, control and management functions using empirical analyses of California micro– and macro– budget statements using the model. We first begin by describing the topics estimated by the model in each set of documents and then go on to demonstrate how this method can produce estimates of planning, control and management function emphases within time periods and over time in two example California counties: (1) Contra Costa, a suburban county of San Francisco and; (2) Yolo County, which is mostly a rural county, with a large college town, Davis, CA accounting for much of its population and revenue.

Table 1: Labeled micro–budget statement topics in California counties, 2012–2017

Topic #	1	2	3	4	5
Function	Planning	Control	Control	Control	Management
Sub–Function	Services	Internal	General Fund	Services	Accounting
	continu	fund	budget	servic	increas
	program	servic	fund	fund	fund
	servic	depart	mend	increas	mend
	object	budget	revenu	depart	year
	provid	cost	unit	budget	account
	develop	posit	increas	continu	budget
	plan	mend	servic	year	servic
	implement	increas	cost	cost	repres
	project	includ	posit	program	current
	system	program	general	state	due

3.5.1 Estimating Budget Functions Emphases in Micro–Budget Statements

We begin our analysis by first labeling the latent topics extracted by the model.

Table 1 contains the top words from each topic estimated from the micro–budget statements. As we mentioned above, topics must be interpreted and labeled by subject matter experts. Accordingly, we labeled each of the topics using the top words from each topic according to whether each topic represented a planning, control or management function and then also labeled the sub–functions which we believed were specific to the California counties. When labeling topics, we draw upon empirical investigations of control, planning and management functions in Friedman (1975) and Daley (1985), which tested Schick's analytical functions using surveys of state and city budget officials.

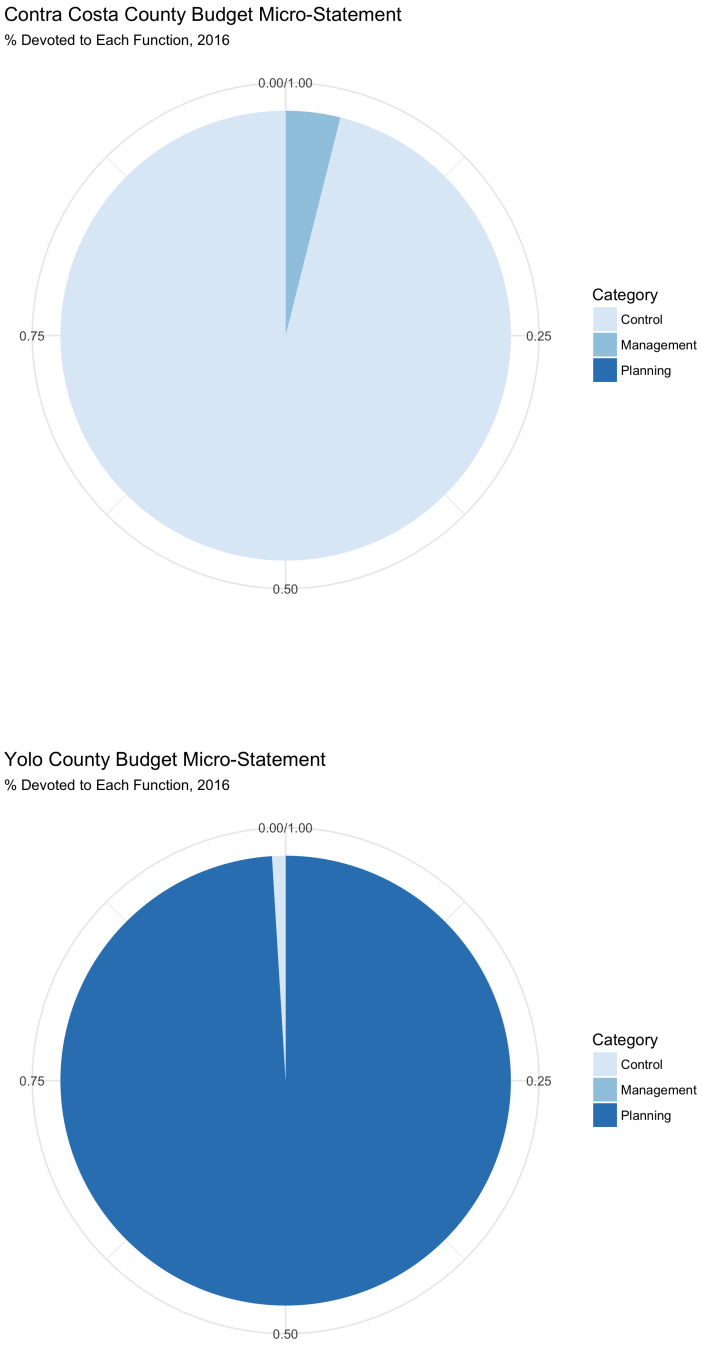
In Table 1, Topic 1 is clearly related to planning functions because it contains the terms “plan”, “objective”, “servic” and “continu”. This suggests that this topic is related to planning for continuation of services. For example, in the 2016-17 Yolo County budget statement they discuss continuation of and planning for services:

> “*...Continue to improve coordination between Planning, Building, Public Works and Environmental Health divisions (Operational Excellence) Conduct outreach to stakeholders regarding development of a retail food placarding (grading) program and develop program proposal and plan. (Safe Communities & Thriving Residents)*”

Topic’s 2, 3 and 4 were labeled as related to control functions primarily because the words "increas", "servic", "fund" and "budget" were found together and control budgeting indicators typically discuss budgetary changes and comparisons with previous years (Friedman, 1975). Finally, Topic 5 is related to "management" functions “account,” “budget,”“due” and “current,” all of which are common management budgeting indicators according to Friedman (1975) such as “existence of a cost accounting system” and “narrative descriptions of activities included in the budget document.”

Figure 8 are estimates of emphases on control, management and planning functions in the Contra Costa and Yolo county budget micro statements. This figure reveals a great deal about these counties’ budget orientations and agendas. While Contra Costa county emphasizes control functions with a minor emphasis on management functions, Yolo county focuses almost entirely on planning functions with a minor emphasis on control functions. These emphases are likely due to institutional differences in each counties’ governance structure or may even be the product of responses to external events such as natural disasters. To get a better sense of

Figure 8: LDA estimated emphasis on control, management and planning functions in the Contra Costa and Yolo County Budget Micro–Statements in 2016.



changes in emphasis over time, we can also explore changes in control, management and planning functions in each county as well.

Figures 9 and 10 contain LDA estimated budget orientations for Contra Costa and Yolo over time. These data allow us to see what each of these different counties find important and how this changes over time. For example, in Figure 9 we see that the control emphases in the Contra Costa county budget statement regarding the general fund and internal issue fluctuate significantly over time, while in the Yolo county budget the emphasis on planning for service provision remains relatively constant over time.

3.5.2 Estimating Budget Function Emphases from Macro–Budget Statements

Here we explore macro-budget statements using the LDA model and discover some interesting results.

Table 2 contains estimated topics for macro–budget statements. We can observe here that all of these topics emphasize the budget and funding. As such, the macro statements appear to primarily emphasize different "shades" of control functions, suggesting that they serve as a means of providing the legislature and others with a broader overview of current county revenue and projections for the future programs.

Specifically, we find that macro–statement topics relate primarily to revenue projections from a variety of sources and address concerns about funding for the public sector workforce. Figure 11 contains LDA estimated emphases on different shades of control functions within the Contra Costa and Yolo county budgets. These charts

Figure 9: LDA estimated emphasis on control, management and planning functions and subfunctions in the Contra Costa and Yolo County Budget Micro–Statements, 2012–2017.

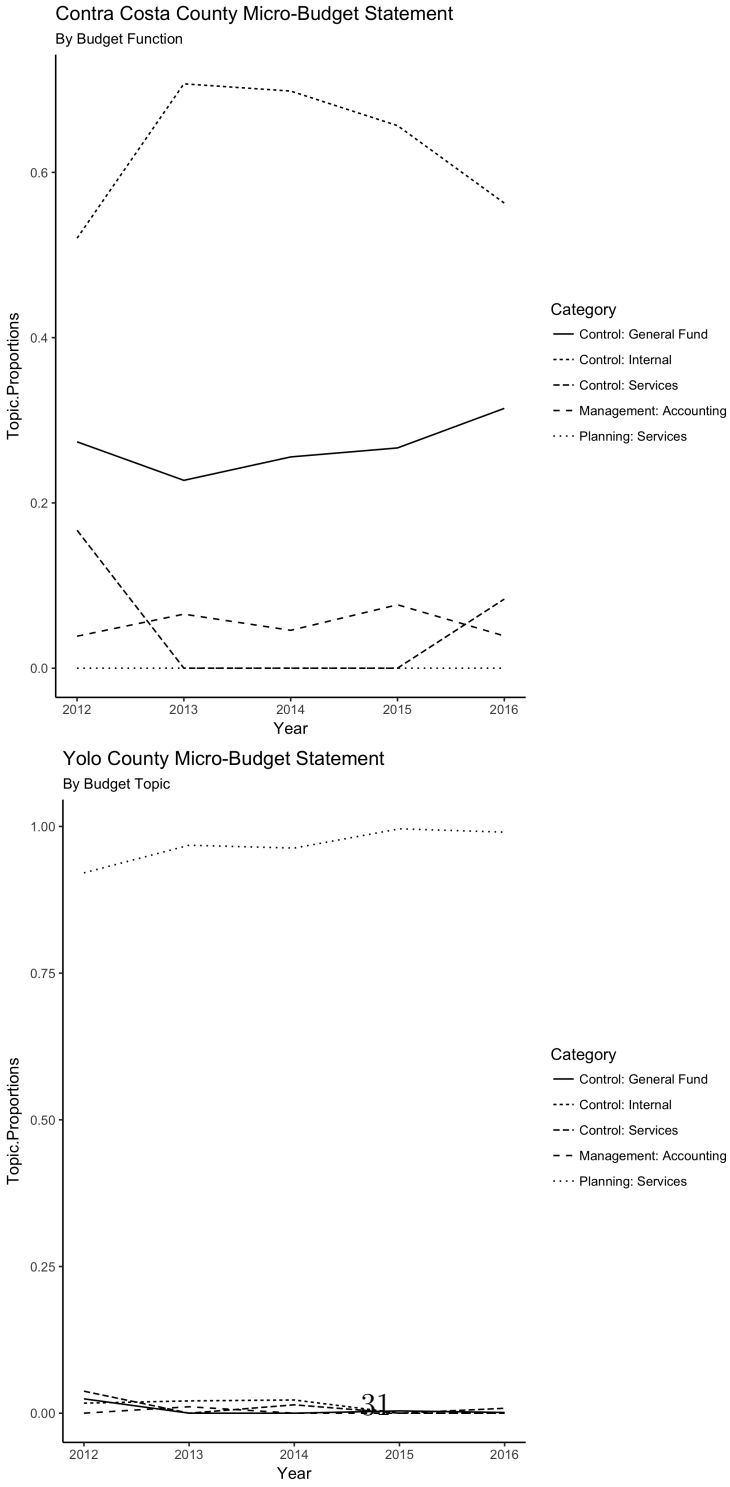


Figure 10: LDA estimated emphasis on control, management and planning functions in the Contra Costa and Yolo County Budget Micro–Statements, 2012–2017.

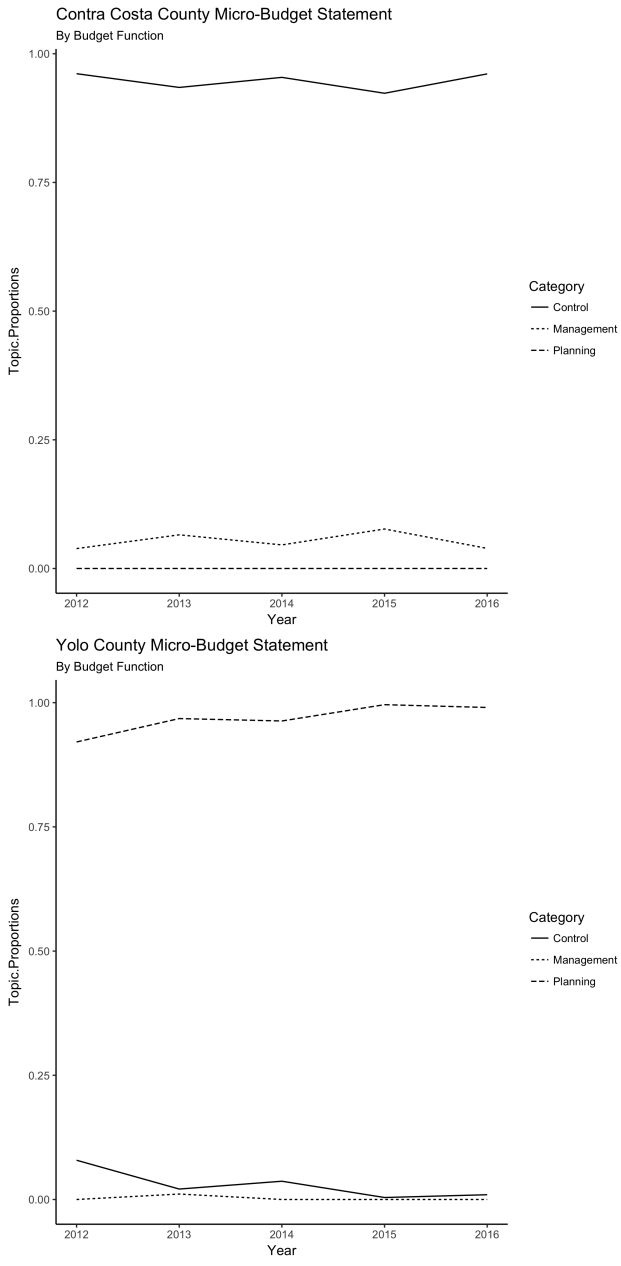


Figure 11: LDA estimated emphasis on control, management and planning functions in the Contra Costa and Yolo County Budget Macro–Statements in 2017.

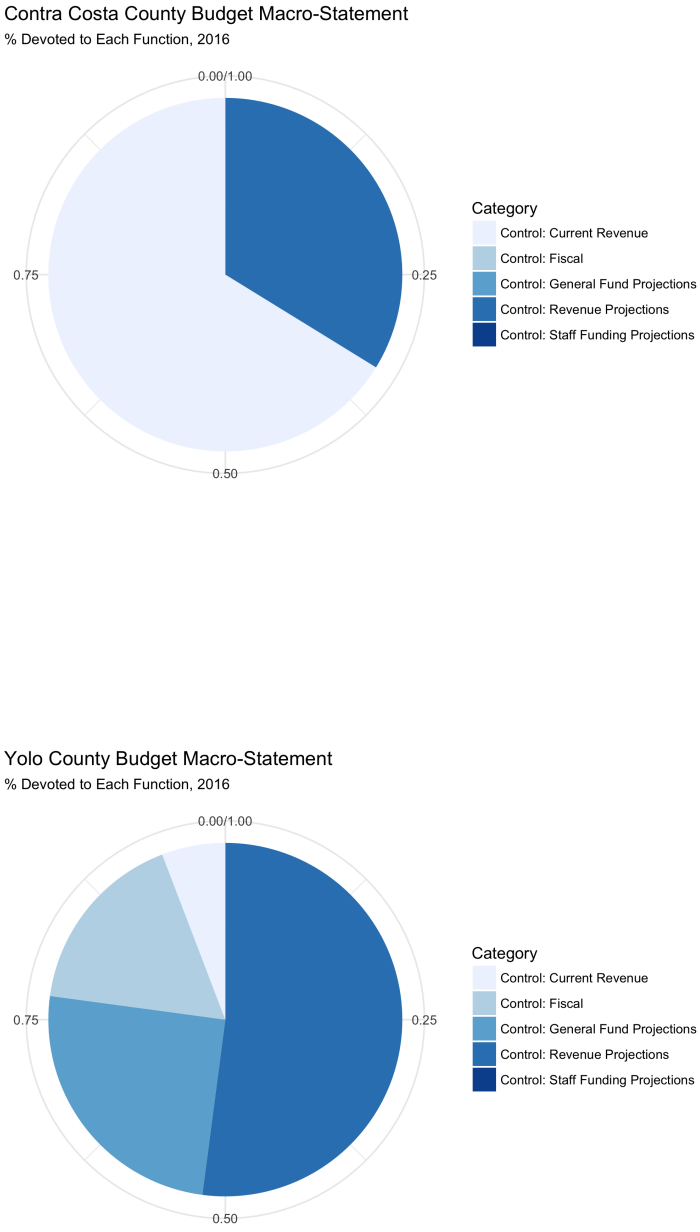


Table 2: Labeled macro–budget statement topics in California counties, 2012–2017

Topic	1	2	3	4	5
Function	Control	Control	Control	Control	Control
Subfunction	Fiscal	Current Revenue	Revenue Projections	General Fund Projections	Staff Funding
	budget	budget	budget	fund	budget
	fund	fund	fund	budget	fund
	year	year	servic	increas	posit
	mend	servic	increas	servic	year
	fiscal	revenu	year	general	cost
	general	state	mend	depart	depart
	revenu	depart	revenu	revenu	increas
	balanc	project	state	cost	general
	expens	continu	includ	includ	servic
	depart	program	public	mend	revenu

make it clear that both counties are very concerned with revenue projections and current revenue, but Yolo county also has significant concerns about California general fund projections, most likely because Yolo county's economy is directly tied to external decisions.

4 Conclusion

Computational text analysis is an increasingly hot topic in the public management world, but there remains a good deal of confusion about what these methods are good for–and their limitations. Text mining tools offer a unique opportunity to undertake large scale, systematic assessments of these documents by replacing methods that rely primarily on human coders. Given the vast array of procedural documents

and text-generating processes–including meeting minutes, public comments, planning documents, and many other texts–that contain information of direct interest to public management scholars, these methodological tools will only grow in importance as a means to expand the scope and detail of research that was once limited by the ability of researchers to collect and code documents by hand. This paper has provided a tour "behind the curtain" to clarify the steps by which documents are turned into analytical data. Using California county proposed budgets for fiscal years 2012-2017, we demonstrate the LDA approach for document analysis. In an annotated application, we show how to extract data from macro- and micro- budget narratives to investigate organizational dynamics within and between counties. This application to county budgets, and the ability to extract Schick (1966)'s analytical functions in budgets and confirm the findings in empirical studies by Daley (1985); Friedman (1975), illustrate the potential value that public management and governance researchers can attain through computational text analysis.

While highlighting the benefits of computational text analysis, we also underline the challenges of quantitative text methods and important practices for researchers in employing such methods. For instance, the LDA approach demonstrated above requires careful validation of the topical clusters. We believe that this annotated application and accompanying code will help to increase the analytical rigor of computational text analysis methods in public management research while at the same time making these methods more accessible to those who have little experience in this area. There are troves of public sector documents that public management and governance research could benefit from. Documents from public sector organiza-

tions contain a wealth of data on dynamics within and between organizations. We encourage researchers to apply the LDA method, which we describe in this article using California county budget documents, in other public management contexts and improve or augment its capabilities for machine learning from public sector texts.

References

Barberá, Pablo. 2014. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political Analysis* 23 (1): 76–91.

Blei, David M, and John D Lafferty. 2006. "Dynamic topic models." In *Proceedings of the 23rd international conference on Machine learning*. ACM pp. 113–120.

Blei, David M, and John D Lafferty. 2007. "A correlated topic model of science." *The Annals of Applied Statistics*: 17–35.

Bond, Robert, and Solomon Messing. 2015. "Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook." *American Political Science Review* 109 (1): 62–78.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. "Natural language processing (almost) from scratch." *Journal of Machine Learning Research* 12 (Aug): 2493–2537.

Daley, Dennis M. 1985. "Control, management, and planning: A state-level replication of the friedman study of budget practices." *International Journal of Public Administration* 7 (3): 291–304.

Friedman, Lewis C. 1975. "Control, Management, and Planning: An Empirical Examination." *Public Administration Review* 35 (6): 625–628.

Greene, Derek, and James P Cross. 2017. "Exploring the Political Agenda of the Eu-

ropean Parliament Using a Dynamic Topic Modeling Approach." *Political Analysis* 25 (1): 77–94.

Grimmelikhuijsen, Stephan, Lars Tummers, and Sanjay K. Pandey. 2017. "Promoting State-of-the-Art Methods in Public Management Research." *International Public Management Journal* 20 (1): 7-13.

Grimmer, Justin. 2009. "A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases." *Political Analysis* 18 (1): 1–35.

Grimmer, Justin, and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21 (3): 267–297.

He, Yulan, Chenghua Lin, Wei Gao, and Kam-Fai Wong. 2013. "Dynamic joint sentiment-topic model." *ACM Transactions on Intelligent Systems and Technology (TIST)* 5 (1): 6.

Hinton, Geoffrey E, and Ruslan R Salakhutdinov. 2009. "Replicated softmax: an undirected topic model." In *Advances in neural information processing systems*. pp. 1607–1614.

Lai, Chih-Chin, and Ming-Chi Tsai. 2004. "An empirical performance comparison of machine learning methods for spam e-mail categorization." In *Hybrid Intelligent Systems, 2004. HIS'04. Fourth International Conference on*. IEEE pp. 44–48.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting policy positions

from political texts using words as data." *American Political Science Review* 97 (2): 311–331.

Lowe, Will, Kenneth Benoit, Slava Mikaylov, and Michael Laver. 2009. "Scaling Policy Positions from Coded Units of Political Texts." In *general conference of the European Consortium of Political Research (ECPR), Potsdam.*

Lowe, Will, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2011. "Scaling policy preferences from coded political texts." *Legislative studies quarterly* 36 (1): 123–155.

Monroe, Burt L, Michael P Colaresi, and Kevin M Quinn. 2008. "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16 (4): 372–403.

Morris, Rebecca. 1994. "Computerized content analysis in management research: A demonstration of advantages & limitations." *Journal of Management* 20 (4): 903–931.

Pandey, Sheela, Sanjay K Pandey, and Larry Miller. 2017. "Measuring Innovativeness of Public Organizations: Using Natural Language Processing Techniques in Computer-Aided Textual Analysis." *International Public Management Journal* 20 (1): 78–107.

Popping, Roel. 2012. "Qualitative decisions in quantitative text analysis research." *Sociological Methodology* 42 (1): 88–90.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–1082.

Schick, Allen. 1966. "The road to PPB: The stages of budget reform." *Public Administration Review*: 243–258.

Tibshirani, Robert, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. 2005. "Sparsity and smoothness via the fused lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (1): 91–108.

Wang, Chong, David Blei, and David Heckerman. 2012. "Continuous time dynamic topic models." *arXiv preprint arXiv:1206.3298*.

Wiedemann, Gregor. 2013. "Opening up to big data: Computer-assisted analysis of textual data in social sciences." *Historical Social Research/Historische Sozialforschung*: 332–357.

Wilbur, W John, and Karl Sirotkin. 1992. "The automatic identification of stop words." *Journal of information science* 18 (1): 45–55.

Young, Lori, and Stuart Soroka. 2012. "Affective news: The automated coding of sentiment in political texts." *Political Communication* 29 (2): 205–231.

Zhang, Cun-Hui, and Jian Huang. 2008. "The sparsity and bias of the lasso selection in high-dimensional linear regression." *The Annals of Statistics*: 1567–1594.

5 Appendix

All macro and micro budget statment text data used in these analysis will be made available on the Harvard Dataverse https://dataverse.harvard.edu/. Below we include all of the replication *R* code for the reader's convenience.

Topic Model Replication Code in R

```
library(pacman)

# This loads and installs the packages you need at once
pacman::p_load(RTextTools, quanteda,
               tm, SnowballC, foreign, plyr, twitteR,
               slam, foreign, wordcloud, LiblineaR, e1071, topicmodels,
    readr,
               stargazer, ggplot2)

# Load all the budget statements into R

# Replace these lines with macro-budget data directory to analyze macro
# budgets

micro = "~/Dropbox/BudgetsTextAnalysis-PMRC2017/CleanData/Micro/"
setwd(micro)

statements = list.files()

```

```
# We want to create the variables: countyname,year,statementtype and
    read in the documents

countyname<-c()
year<-c()
statementtype<-c()
documents<-c()

for(i in 1:length(statements)){
  temp<-unlist(strsplit(statements[i], "[.]"))
  countyname<-c(countyname,temp[1])
  year<-c(year,temp[2])
  statementtype<-c(statementtype,"micro")
  documents<-c(documents,read_file(paste(micro,statements[i],sep="")))
}

# Use regular expressions to clean up some elements of the documents
cleandocs<-c()

for(i in 1:length(documents)){
    cleandocs[i]<-gsub("\xfc\xbe\x8d\x96\x90\xbc"," ",documents[i])
    cleandocs[i]<-gsub("\r\n"," ",cleandocs[i])
    cleandocs[i]<-gsub("\xfc\xbe\x8c\xa3\xa4\xbc"," ",cleandocs[i])
}


# Now we clean and preprocess the statements and prepare them for
    analysis
```

```
## Text pre-processing function

text_cleaner<-function(corpus){
  tempcorpus = lapply(corpus,toString)
  for(i in 1:length(tempcorpus)){
    tempcorpus[[i]]<-iconv(tempcorpus[[i]], "ASCII", "UTF-8", sub="")
  }
  tempcorpus = lapply(tempcorpus, tolower)
  tempcorpus<-Corpus(VectorSource(tempcorpus))
  toSpace <- content_transformer(function (x , pattern ) gsub(pattern,
    " ", x))

  #Removing all the special charecters and words
  tempcorpus <- tm_map(tempcorpus, toSpace, "/")
  tempcorpus <- tm_map(tempcorpus, toSpace, "@")
  tempcorpus <- tm_map(tempcorpus, toSpace, "\\|")
  tempcorpus <- tm_map(tempcorpus, toSpace, "#")
  tempcorpus <- tm_map(tempcorpus, toSpace, "http")
  tempcorpus <- tm_map(tempcorpus, toSpace, "https")
  tempcorpus <- tm_map(tempcorpus, toSpace, ".com")
  tempcorpus <- tm_map(tempcorpus, toSpace, "$")
  tempcorpus <- tm_map(tempcorpus, removeNumbers)
  # Remove english common stopwords
  tempcorpus <- tm_map(tempcorpus, removeWords, stopwords("english"))
  # Remove punctuation
  tempcorpus <- tm_map(tempcorpus, removePunctuation)

  # Eliminate extra white spaces
```

```
  tempcorpus <- tm_map(tempcorpus, stripWhitespace)
  # Stem the document
   tempcorpus <- tm_map(tempcorpus, PlainTextDocument)
   tempcorpus <- tm_map(tempcorpus,  stemDocument, "english")

  # Remove uninformative high-frequency words
   #tempcorpus <- tm_map(tempcorpus, toSpace, "budget")
   tempcorpus <- tm_map(tempcorpus, toSpace, "million")
   tempcorpus <- tm_map(tempcorpus, toSpace, "counti")
   #tempcorpus <- tm_map(tempcorpus, toSpace, "fund")
   tempcorpus <- tm_map(tempcorpus, toSpace, "also")
   tempcorpus <- tm_map(tempcorpus, toSpace, "will")
   tempcorpus <- tm_map(tempcorpus, toSpace, "board")
   tempcorpus <- tm_map(tempcorpus, toSpace, "last")
   #tempcorpus <- tm_map(tempcorpus, toSpace, "year")
  return(tempcorpus)
}

cleancorpus<-text_cleaner(cleandocs)

# Transform into a document-term matrix
dtm <- DocumentTermMatrix(cleancorpus,
                          control = list(removePunctuation = TRUE,
                                         stopwords=TRUE,
                                         weighting=weightTf))


```

```
# Only use words that are present 10 or more times (remove very rare words)

ten_words<-findFreqTerms(dtm,lowfreq = 10)

dtm10<-DocumentTermMatrix(cleancorpus,
                          control = list(removePunctuation = TRUE,
                                         stopwords=TRUE,
                                         weighting=weightTf,
                                         dictionary=ten_words
                          ))


# Run an LDA model with 5 topics looks optimal
ap_lda5 <- LDA(dtm10, k = 5, control = list(seed = 41616), method="VEM")
macrotable<-terms(ap_lda5, k=10)
stargazer(macrotable)

# What is the best model from a perplexity perspective?

model.perplexity<-c()
for(i in 2:20){
  ap_lda <- LDA(dtm10, k = i, control = list(seed = 1234))
  model.perplexity<-c(model.perplexity,perplexity(ap_lda))
}
```

```
125 setwd("~/Dropbox/BudgetsTextAnalysis-PMRC2017/Draft/figs")
126
127 png("perplexity-macro.png")
128 plot(2:20,
129      model.perplexity,
130      xlab="Number of Topics",
131      ylab = "Perplexity",
132      main = "Macro Budget Statements")
133 lines(2:20,model.perplexity)
134 dev.off()
135
136
137 # Marginal perplexity
138
139 marginal.perplexity<-c()
140 for(i in 2:length(model.perplexity)){
141   marginal.perplexity<-c(marginal.perplexity,(model.perplexity[i]-model
      .perplexity[i-1])/model.perplexity[i-1])
142 }
143
144 index<-1:length(marginal.perplexity)
145
146 png("marginal-perplex-macro.png")
147 plot(index,marginal.perplexity, xlab="Number of Topics", ylab = "
      Marginal Perplexity (% Change)",
148      main = "Macro Budget Statements")
149 lo <- loess(marginal.perplexity~index)
150 lines(predict(lo), col='red', lwd=2)
```

```
lines(index,marginal.perplexity, col='grey', lwd=1)
dev.off()


# Run an LDA model with 5 topics looks optimal
ap_lda5 <- LDA(dtm, k = 5, control = list(seed = 41616), method="VEM")
terms(ap_lda5, k=5)

### Estimate topic proportions using posterior inference

posterior_inference <- posterior(ap_lda5)
posterior_topic_dist<-posterior_inference$topics # This is the
    distribution of topics for each document
```