Government websites as data: A methodological pipeline with application to the websites of municipalities in the United States

Markus Neumann Fridolin Linder Bruce Desmarais

September 28, 2018

Abstract

A local government's website is arguably the most important general source of information about policies and procedures for residents and other community stakeholders. Accordingly, government websites have become prominent sources of data for a variety of research agendas in public administration, public policy, and political science. Existing research has relied on manual methods of website data collection and processing. However, reliance on manual collection and processing limits the scale and scope of website content analysis. We develop a methodological pipeline that researchers can follow in order to gather, process, and analyze website content with established text analysis techniques. First, for the acquisition of website data, we cover approaches to automated scraping methods. Second, pre-processing is a particularly vital step in text analysis, but when websites are concerned, additional measures need to be taken in order to guard against potential sources of bias. We propose a new method for dealing with the types of duplicated and boilerplate contents that are commonly found in government websites. We illustrate our methodological pipeline through the collection and analysis of a new and innovative dataset—the websites of over two hundred municipal governments in the United States. We build upon recent research that analyzes how variation in the partisan control of government relates to content made available on the government's website. Using a structural topic model to analyze municipal website contents, we find that websites of cities with Democratic mayors include more information about policy deliberation and crime control, whereas websites from cities with Republican mayors include more information about the provision of basic utilities and services such as water, electricity, garbage removal and fire safety; and information on municipal ordinances and policies.

1 Introduction

Local governments convey voluminous information about all aspects of their policymaking, policy implementation, and public deliberation, via their official websites. The vital role of official websites in connecting the government and the governed has motivated a wave of research on the contents of government websites (e.g., Grimmelikhuijsen 2010; Wang, Bretschneider and Gant 2005; Osman, Anouze, Irani, Al-Ayoubi, Lee, Balcı, Medeni and Weerakkody 2014). Despite the potential for automated scraping of website contents, the conventional approach to data collection in projects focused on government websites involves manual content extraction from each website in the dataset. Though highly accurate, the manual approach to data collection is costly, and cannot be scaled to capture even a fraction of the volume of content available on government websites. In this paper, we present a methodological pipeline that can be used to automatically scrape government websites in order to build datasets that can be used for text analysis. We provide an illustrative application in which we explore the ways in which the textual contents on city government websites in six American states (IN, LA, NY, WA, CA and TX) correlate with the partisanship of the city mayor.

Though there exists a variety of software tools that are designed to automatically scrape all of the files available at a website (Glez-Peña, Lourenço, López-Fernández, Reboiro-Jato and Fdez-Riverola 2013), raw website downloads have to be processed significantly before the files are adequately prepared for content (e.g., text, image) analysis. We describe and provide solutions to two central challenges in automatically gathering and analyzing website textual contents. First, plain text must be extracted from the files. This involves purging the files of syntax in HTML or other markup and programming languages and discarding any character encoding errors that result from reading in files with incorrectly specified character encodings. This challenge would arise in any context in which researchers sought to study the textual contents of websites, and is not unique to comparative analysis of government websites. The second challenge we address in our methodological pipeline is, however, specific to the research objective of comparing websites on the basis of a common lexicon. For any two governments, the textual signatures that most dramatically differentiate the textual contents of their websites consist of what we can call "boilerplate" text—header, footer, or other titling text that is designed to identify the website as being associated with a specific government entity (e.g., "Welcome to the city of Santa Cruz", "The City of Los Angeles welcomes you"). This boilerplate text is replicated across many files that are associated with a government's website, but it provides little information regarding the form and/or function of the government. The second methodological innovation we offer in our pipeline is designed to minimize the impact of this boilerplate text on the comparative analysis of government website content.

Government websites provide information about how public policies shape the lives of local residents, and how local residents can engage with government to shape public policy. As such, government websites reflect both the results of and inputs to, the political leadership in the city. In our illustrative application, we explore the ways in which the contents of city government websites differ on the basis of the partisanship of the city's elected executive. A substantial body of research has found that the partisanship of the mayor affects city governance along multiple dimensions, including city budget priorities (de Benedictis-Kessner and Warshaw 2016), policies affecting inequality in cities (Einstein and Glick 2016), and framing of criminal justice policy (Marion and Oliver 2013). Furthermore, recent media coverage of changes to government websites that follow transitions in party control suggest that changes in web content are salient government actions, as perceived by the general public (Sharfstein 2017; Kirby 2017; Duarte 2017). We study whether significant differences between city governments based on mayoral partisanship are reflected in the contents of city websites.

2 The Significance of Government Website Content

According to Mayhew (1974), politicians engage in advertising, credit claiming and position taking in order to get re-elected. Official city websites allow mayors to perform all three of these functions. Their offices frequently take a prominent position on the front page, and many websites also feature a picture of the mayor. We present an example of this in Figure 1. The Erie, Pennsylvania website homepage presents an image of the Democratic mayor, Joseph Schember, along with a list of laudable attributes of the city. In local politics, where campaign funds are low, this lends

Figure 1: Screenshot from the homepage at http://www.erie.pa.us/, accessed on 06/14/2018. Image depicts Democratic mayor of Erie, PA, Joseph Schember.

incumbents a crucial advantage in becoming more well-known among their constituencies (Stanyer 2008). Furthermore, municipal politics gives incumbents clear and tangible achievements they can point to, such as completed infrastructure projects, the acquisition of federal or state funding, or the hosting of city-wide events. City websites present an opportunity for local officials to brandish these accomplishments. Finally, they also give mayors a platform from which they can advertise their political beliefs. On municipal websites, this may not manifest in the form of brazen partisanship, but more subtle avenues are available. As noted by Einstein and Glick (2016), there are stark differences in the spending preferences of Democratic and Republican mayors. City websites can then be used to communicate the stance of a mayor on social or economic programs. Another advantage of websites with regard to communication is that unlike direct social interactions, officials have full control over them.

Members of the public visit municipal government websites for a wide variety of purposes Sandoval-Almazan and Gil-Garcia (2012), and with significant regularity. In a survey conducted among a random sample of citizens in the state of Georgia in 2000—nearly two decades ago—found that 25% of internet users reported visiting a local government website in the previous twelve months (Thomas and Streib 2003). Furthermore, the use of a local government website is associated with an individual's perspective on government. Tolbert and Mossberger (2006) find that users of local government websites are more likely to trust local governments, and hold other positive attitudes related to local and federal governments. Lastly, in a study of residents of Kansas City, Missouri, Ho and Cho (2017) find that participants' perceived quality of the city website is strongly associated with their perceptions of the overall effectiveness of the City's communication with the public.

The existing research that uses scraped websites provides an indication of the theoretical value of empirical analysis of web contents produced by governments, public officials, and candidates for office. The most pertinent literature to our research is the e-governance literature, which focuses on the online presence of governments from a usability and public service point of view. For the most part, research in this category develops a classification scheme to rate websites in terms of accessibility, ease-of-use, and function, and then hand-codes a set of websites according to these criteria (e.g., Urban 2002; Armstrong 2011; Feeney and Brown 2017). As an example, Grimmelikhuijsen and Welch (2012) study local government websites with the goal of uncovering how they aid the goal of transparency. To this end, they analyze a set of Dutch municipalities in which air quality had deteriorated. The authors test whether local governments provide citizens with information about potential complications and solutions associated with this issue. Like most e-government studies, however, this publication does not make any use of automated text analysis. Branching out of the topic of e-government service provision, McNutt (2010) uses Canadian government websites to study how well-connected government websites are to non-profits and intergovernmental organizations in different public policy domains.

The websites of politicians and their parties have also been the object of research. Researchers have found that in order to identify the constituencies, motives, and modes of communication among these actors, their websites can be very illuminating sources of information (Druckman, Kifer and Parkin 2009; Druckman, Hennessy, Kifer and Parkin 2010; Cryer 2017; Esterling, Lazer and Neblo 2011; Esterling and Neblo 2011; Norris 2003; Therriault 2010). Druckman, Kifer and Parkin (2009); Druckman et al. (2010) rely on the *National Journal* to find the websites, then hand-coded them. Cryer (2017) mentions that she relied on Archive-it, a web service of the Internet Archive. Though the Internet Archive may provide extensive coverage of high profile officials or national governments, we found that its coverage of municipal government websites was sparse and irregular. Esterling, Lazer and Neblo (2011); Esterling and Neblo (2011) rely on hand-coded data by the Congressional Management Foundation, a nonprofit organization which aims to assist Congress. Therriault (2010) use automated text analysis, and also has the most extensive overview of the associated methodology. However, the division of the website into sections (home page, topics, issues, details) is done by hand. The author acquired the websites from the Library of Congress (which only collected them from legislators who actually consented, and Therriault notes that this causes nonrandom missingness).

Important to our methodological objectives, research analyzing and improving the scraping, pre-processing and text analysis pipeline that is applicable to government websites is still in its infancy. Eschenfelder, Beachboard, McClure and Wyman (1997) provide something of an overview of how federal websites should be assessed from an e-governance point of view, but they largely focus on the substantive criteria that should be fulfilled, rather than the technical aspects of website acquisition and analysis.

3 Data: US Municipal Government Website Text

For data availability reasons, we focus our analysis of municipal websites on six states—Indiana, Louisiana, New York, Washington, California, and Texas. The selection of states and cities is largely dictated by the presence of partisan mayors and availability of the relevant data. Municipal elections in Indiana and Louisiana are partisan across the board,[1] so our sample is primarily focused on these two states. Here, all cities with a website are included, resulting in a considerably larger sample than for the other four states. New York and Washington do not have nominally partisan elections, but for a subset of cities, partisanship can be determined through contribution data (see below for more detail on how we use this data). Finally, California and Texas contain a number of large cities, whose mayors are sufficiently well-known for their partisanship to be available. These six states also provide us with a sample that is well-balanced on a number of theoretically important indicators. One, each of the four geographic regions are represented with at least one state. Two, we have a fairly well-balanced sample with respect to the urban/rural cleavage, as both major cities and less densely populated areas are covered. Furthermore, the sample is politically balanced—we have three blue states (CA, WA, NY) and three red states (TX, IN, LA). Finally, our dataset contains some of the wealthiest states (NY, CA, WA and TX are #2, #8, #9 and #16 respectively, by GDP per capita (Bureau of Economic Analysis 2017)), but also some of the poorer ones (IN and LA).

We acquired the website URLs from two sources: One, we scraped the URLs of city websites from their respective Wikipedia pages, which we found from lists of cities contained within each state. Two, the General Services Administration (GSA) maintains all '.gov' addresses, and provides a complete list of all such domains to the public.[2] The data from the GSA contains the following variables: (1) domain name, specifically, the all-uppercase version of domain and top-level domain (for example, 'ABERDEENMD.GOV'); (2) the type of government entity to which the domain is registered, such as city, county, federal agency, etc; (3) for federal agencies, the name is specified; (4) the city in which the domain is registered. Naturally, the GSA's list does not contain cities

[1] In Indiana, this includes only cities - incorporated municipalities with at least 2,000 inhabitants - as opposed to towns.

[2] The dataset is made available at https://github.com/GSA/data/tree/gh-pages/dotgov-domains. This list is updated once per month—we rely on the version released on January 16, 2017.

which do not use a '.gov' website (or, in many cases, a city owns a registered '.gov' address, but uses a different one). Furthermore, some of the links are non-functional, and some of the county websites on the list are incorrectly marked as city websites (and vice versa). Since the GSA data is less complete and less reliable than the URLs found on Wikipedia, we mainly rely on the former and only supplement them with the GSA data if a specific city doesn't have a URL recorded on Wikipedia, or our tests (see below) find it to be non-functional.

Not all of the URLs contained in these archives are functional. To test the URLs' functionality, we use a web driver-controlled browser - a browser that is automatically controlled by a program rather than a human user. We use the Python bindings for the program `Selenium`, which we use to control `Firefox` through the web driver `Geckodriver`. This is advantageous compared to conventional scraping tools such as `Beautiful Soup` or `Rvest` because most websites are designed to be explored by browsers. Modern browsers perform a lot of actions behind the scenes, such as URL resolution and redirection. The use of a web drive-controlled browser is necessary in our case because a) some city websites simply don't work, but they don't always output an error code correctly (this can fail, for example, if a webmaster simply stops maintaining a site without removing it entirely) which would throw off an automatic scraper, and more often, b) cities sometimes change their websites' URLs, in which case they redirect from the old to the new URL. A web driver-controlled browser, unlike the more rigid conventional scraping tools, will simply follow this redirection. This allows us to subsequently record and use the new URL for the actual website scraping. Consequently, an automated browser allows us to robustly answer the following questions: Is the website actually there? Does it work? If not, is it somewhere else or is it broken? We record this information and construct a list of verified URLs.

To download the websites, we rely on the Unix command line tool `wget`. This program is used to download files from the Internet, and with the use of a recursive option, acts like a web crawler and scraper. This means that `wget` downloads HTML files, parses them and then follows the links contained therein. Then it follows those links and repeats the process until it has constructed a complete tree of the website (note that the program is instructed to stay on the same domain, i.e. it does not follow external links). This way, all the files that make up a website are downloaded. For some cities, whose websites make heavy use of JavaScript to serve content dynamically, such content is not reachable with our methodology and would require additional steps to obtain. For this paper, we ignore such sites and restricted our corpus to cities with at least three successfully downloaded pages.[3]

The partisanship of the mayor of each city is coded in different ways, depending on the state. For Indiana, where elections are nominally partisan, this information is accessible through the state government's website[4]. For Louisiana, we received data on the outcomes of mayoral elections from the Local Elections in America Project (LEAP) (Marschall and Shah 2013). For the other states, where mayoral elections are not nominally partisan (but the partisanship of the mayor is still well-known), we employed different means: For New York and Washington, we searched the state campaign finance websites, and coded the parties of the candidates based on the party committees from which they received donations. For California and Texas, where our data consists of highly populated cities, partisanship information was acquired from Ballotpedia[5]. Finally, we also scraped mayoral partisanship from the cities' Wikipedia pages. When compared to the other data sources above, (and manual searches in case of conflicts) Wikipedia proved to be very reliable and added additional cases to our dataset even for Indiana and Louisiana. Generally speaking, we found data scraped from Wikipedia, aided by manual corrections in case of missing or conflicting data, to be more reliable than data from governmental sources. The partisan breakdown of city websites is depicted in Table 1. The dataset is has a relative party balance, with variation in each state.

[3]There is a possibility that this leads to a small bias in selecting against cities with the resources to build more elaborate websites. However, given that our sample is generally more on the wealthy side, this, if anything, should lead to a more balanced sample.

[4]http://www.in.gov/apps/sos/election/general/general2015?page=office&countyID=1&officeID=32&districtID=-1&candidate=

[5]https://ballotpedia.org/List_of_current_mayors_of_the_top_100_cities_in_the_United_States

State	Democratic	Republican
California	9	6
Indiana	46	54
Louisiana	28	17
New York	36	16
Texas	2	7
Washington	11	2

Table 1: Descriptive statistics on the partisanship of the cities in the corpus.

One of the more subtle aspects of local government is the presence of different types of government structures. Between council-manager governments and mayor-council governments (Morgan and Watson 1992)—either in the weak or strong mayor variant (DeSantis and Renner 2002)—there is a certain degree of variance in where a city's executive authority lies. Unfortunately, we do not have access to information about the type of governments across the breadth of our dataset and therefore cannot explore heterogeneity in the relationship between mayoral partisanship and municipal website contents based on the executive system variant. Given the prominent place that mayors tend to have on their cities' websites, we feel that any bias arising from this nuance should be minor. (Gerber and Hopkins 2011), whose model is somewhat comparable to ours in the sense that they also test the effect of mayoral partisanship on city policy priorities (on a sample of large cities, for which this information is more easily available), find that the inclusion of this potential confounder does not affect the results. Information on other covariates (population and median household income - from the American Community Survey 5) was acquired through the API of the U.S. Census Bureau[6].

[6] https://www.census.gov/data/developers/data-sets.html

4 The Web to Text Pipeline

Once we have gathered the website files, we have the raw data necessary for text analysis, but it is not yet formatted effectively. As is common in text analysis applications, we need to pre-process our data (Denny and Spirling 2018). In this section, we describe our pre-processing pipeline, with which we take an archive of website files, and output a corpus of formatted plain text files that are suitable for comparative analysis with text as data methods. In this methodological pipeline, we address two methodological challenges. First, though they contain significant amounts of text, websites are not comprised of clean plain text files. Rather, the files available at websites are of multiple types, including HTML, PDF, word processor, plain text, and image files. The first step in the methodological pipeline is aimed simply at extracting clean plain text from this heterogeneous file base. The second step in our methodological pipeline is to process the text to remove boilerplate language—language that is effective at differentiating one website from another but is uninformative regarding policy or political differences between governments.

4.1 Site to Text Conversion

WEB For the most part, the file type of a document can be correctly determined through the filename ending—its extension. However, there are exceptions to this, which, if ignored, can lead to large amounts of improperly formatted text, arising from incorrectly converted documents, which leads to a general decrease in the amount of usable data. Two issues, in particular, need to be addressed: One, HTML files on city websites frequently do not have an ending but are still perfectly readable if correctly identified as such. Second, some documents contain the incorrect file ending. For example, we found thousands of documents that ended in .html, when they were actually PDFs. To accurately assess their type, we rely on the R package `wand` (Rudis, Zoulas, Rullgard and Ong 2016), which is an R interface to the Unix library `libmagic` (Darwin 2008), which determines the type of a file on the basis of its file signature - or “magic number”. This short sequence of bytes at the start (and sometimes end) of files is unique for each file type and therefore allows its correct identification through computer forensics tools such as `libmagic`.

Consequently, we rename all documents so that their file ending reflects their actual file type. This is strictly necessary because we rely on the `readtext` R package (Benoit and Obeng 2018), which determines a document's type solely through its ending—to convert the files to plain text.[7] The breakdown of the files by type is given in Table 2. The most frequent file types are HTML and PDF, from which we are able to extract a substantial amount of usable text. Files of type XML, DOC, TXT, and DOCX, also occur regularly in our corpus and offer a considerable volume of textual data.

Filetype	Occurances Before	Occurances After
html	211682	887362
pdf	464842	638802
jpg	0	36958
xml	0	29638
Other	162681	9475
ics	435	8950
png	0	8863
doc	6972	8430
txt	317	6025
	793990	5234
docx	3137	4319
TOTAL	1644056	1644056

Table 2: Number of files per type, before and after detecing them via their magic number. The table shows that a lot of files originally have the wrong type, and that converting them correctly has a large impact on how many of them end up being usable.

We then take several steps to pre-process the data as required for the subsequent analysis. Preprocessing choices should be contingent on the analysis being conducted with the text later and can have significant effects on the outcomes of an analysis (Denny and Spirling 2018).

STM The text documents are converted to UTF-8 and then stripped of dates, punctuation,

[7]We have also experimented with several Unix-based alternatives, but found that they largely led to the same results as `readtext`.

numbers, and words connected by underscores. At this point, the documents of one city still closely resemble one another in the form of boilerplate content, be it website elements (i.e. "You are here", "Home", "Directory" etc.) in HTML documents, or commonly used forms or phrases in pdf, doc and docx files. This is an issue, because this boilerplate content causes the results of analyzing this data with text analysis methods to characterize documents primarily by the cities from which they originate (through their unique boilerplate structure, e.g. a menu with certain terms repeated on every site of the domain), and not the substantive features of their contents. Our solution to this problem is described in more detail in Section 4.2.

STM Preprocessing further includes setting every character to lowercase, as well as the removal of bullet points which frequently occur in HTML documents, extraneous whitespace, XML documents mislabeled as HTML files, and empty documents. Furthermore, some documents contain gibberish, often as a result of faulty or impartial optical character recognition applied to text that was produced through a non-machine-readable medium. To combat this problem, we employ two solutions. One, we use spellchecking, implemented through the `hunspell` R package (Ooms 2017), to remove all non-English words.[8] However, `hunspell` does not cover everything, either because some tokens are not actual words (for example artifacts from defective encoding), or because random sequences of characters just so happen to form words that exist in a dictionary (for example "eh" or "duh"). Since we rely on a bag-of-words model in which syntax does not matter, we can ameliorate these problems by removing all text except for whitespaces and the characters that appear in the English alphabet. Since a lot of the nonsensical text tends to be quite repetitive, we also delete all documents in which the proportion of unique to the total number of tokens is less than 0.15. Furthermore, `hunspell` does not spellcheck individual characters or two-character words, so we remove these token types entirely. Since these pre-processing steps

[8]Some of the cities, for example, Los Angeles, do contain a sizable proportion of Spanish content. The analysis of this content is beyond the scope of this paper but could be explored in future work, for example using methods of text processing that are applicable to multilingual corpora (Lucas, Nielsen, Roberts, Stewart, Storer and Tingley 2015).

reduce documents which are largely unsuitable to only a few tokens (i.e., word occurrences), we also remove all remaining documents containing less than 50 tokens. Finally, to remove words that are extremely rare (which also has the advantage of eliminating any remaining oddities) and thus add nothing substantive[9] to our models while increasing their computational cost, we also discard any token types that occur in only one document. We also conduct lemmatization to reduce words to their basic form. Lemmatization is similar to stemming but works in a somewhat more sophisticated manner by taking grammar and surrounding words into account to identify the dictionary form of a word. For example, the lemma of the word "lemmatization" would be "lemmatize", whereas most stemmers would simply chop off the ending, which would yield "lemmatiz". Thus, lemmatization makes the results more easily comprehensible. To this end, we rely on the R package `spacyr`, which provides an R implementation of the Python library `spaCy`.

4.2 Boilerplate Removal

As noted above, city websites contain a large amount of text that is uninformative for its actual content, and therefore a hindrance to understanding through algorithmic text analysis. This is a common issue with textual data in which informative content is embedded in technically structured documents. See, e.g., Burgess, Giraudy, Katz-Samuels, Walsh, Willis, Haynes and Ghani (2016); Wilkerson, Smith and Stramp (2015) and Linder, Desmarais, Burgess and Giraudy (Forthcoming) for examples of boilerplate removal in the analysis of legislative text. In the case of websites, lines in documents are generally quite informative, so all of our boilerplate removal efforts are done at the line level.

[9]Topic models essentially do not pick up on extremely rare words, so their inclusion is a waste of computational resources. Removing them in this manner is also the default preprocessing choice in the `stm` package.

Boilerplate Classification

In order to determine whether a line should be discarded, we train a classifier on a human-coded sample. We sampled 500 lines from documents in each of the following five cities: Los Angeles, CA, Indianapolis, IN, New York, NY, Shreveport, LA, and Seattle, WA. To ensure that lines which occur more frequently in these cities (sometimes hundreds of thousands of times) had a higher probability of being scrutinized by the classifier, we use sampling weights equivalent to the proportion of total lines in a city's corpus made up by each specific line type. As an example, the most common line throughout all pages of the city of Seattle consists only of the word "total" and occurs 103,068 times. Similarly, the line "page" occurs 58,833 times. Even something completely nonsensical such as "a a" still appears on 376 occasions. To account for the higher likelihood of some lines being part of the training set, we use inverse probability weights in training the classifier—the weight of each line in the sample is 1/[number of occurrences in the corpus].[10]

These 500 lines were then hand-coded as either substantively informative (210 lines) or not (290 lines). We then trained a number of different classifiers with this informativeness measure as the dependent variable. The independent variables we use are: (1) number of times the line was duplicated within the city, (2) the length of the line, in characters, (3) the number of tokens in the line, and (4) the median distance from the document midpoint to the position of the line itself. The purpose of these covariates is as follows:

- **Line length:** The length of the line and the number of tokens are ways to find lines consisting of only a word or two. This is highly predictive of lines which are used as website headers and navigational elements, which are of zero substantive interest to us but are very effective at differentiating cities. These terms also happen to be fairly common, which causes them to be overweighted by the topic model.

[10]Note that the performance of the classifier is robust to the use of these weights and only changes by about one percentage point if they are not used.

- **Number of line duplications:** To directly address the latter problem, we include a measure of the number of times a line is duplicated within a city. Many lines occur hundreds or even thousands of times on a single website and therefore are terms that are highly predictive of the website, which causes the topic model to find topics that are highly predictive of cities, but not substantively informative.

- **Line position in the document** Since boilerplate terms such as navigational elements, headers, footers, and so on, should occur more frequently at the beginning and the end of websites, we attempt to identify such content as following: We measure the distance between the midpoint of a document and the position of a line, expressed as quantiles (to account for differing document lengths). Since lines can occur in multiple documents or multiple times in the same document, we take the median of these measures. Thus, for example, a line which often occurs at the beginning of documents might have a score of 0.45, whereas a line that tends to be found more in the center, and thus is indicative of more relevant content, might be scored with a 0.11 instead.

	Value
Percent Correctly Predicted	0.87
Precision	0.87
Recall	0.91
F1-Score	0.89

Table 3: Performance metrics for random forest boilerplate classifier, with inverse probability weights.

We rely on a random forest as the final classifier, which offers slightly better performance than logistic regression.[11] We assess the performance of this classifier through five-fold cross-validation. This means that the classifier is trained on 400 samples and then tested on the held-out set of 100, measuring metrics such as percent correctly predicted, precision, recall, and F1

[11]We also tried SVM, boosted trees and AdaBoost, with similar results and chose the random forests because this method has a probabilistic basis and is more intuitive.

score. This procedure is carried out five times so that each sample is part of the test set once. The aggregated (mean) results of this process can be found in Table 3. For the implementation of this method, we rely on the R package `caret`, whose random forest classifier is based on the package `ranger`. We use this classifier to flag and remove all lines that are not classified (based on a threshold of $p = 0.5$) as substantively meaningful. The effect of this process on the corpus is illustrated with the corpus of Anchorage, AK (i.e. a city that isn't part of our sample used in the analysis) as an example in Figures 2 to 5. Before the lines identified by the classifier as boilerplate are removed, lines with very few characters and words are the most common. After the removal, the distribution has changed—lines of medium length now occur more frequently than extremely short ones, which are unlikely to be substantively meaningful (see figures 2 and 3). Furthermore, lines that are duplicated only a few times rather than dozens, hundreds or even thousands are now more common (see figure 4). Finally, the position of the line in the documents is not as important to the random forest, and this also shows in the results. However, this feature still has a positive effect, as lines at either end of the document are a bit less common now (see figure 5). Table 4 provides further illustration by listing the top 10 most likely boilerplate lines (in Anchorage, AK) – all of which were flagged as such with a probability of 1. After all the preprocessing is set and done, our corpus consists of 259,099 documents.

5 Partisan Language on Municipal Websites

We illustrate the analysis of municipal website content by studying differences in website content based on the party of the mayor. As we reviewed above, the partisanship of the mayor has been found in past research to affect several features of city governance. However, Gerber and Hopkins (2011) note that, due to the constraints of state and national policies, municipalities lack discretion in many domains of governance. They find that cities with Democratic mayors spend a smaller share of their budgets on public safety, but that mayoral partisanship does not appear to be

Figure 2: Effects of the boilerplate classifier on the corpus of the city of Anchorage, AK. After the boilerplate content is removed, extremely short lines are less common.

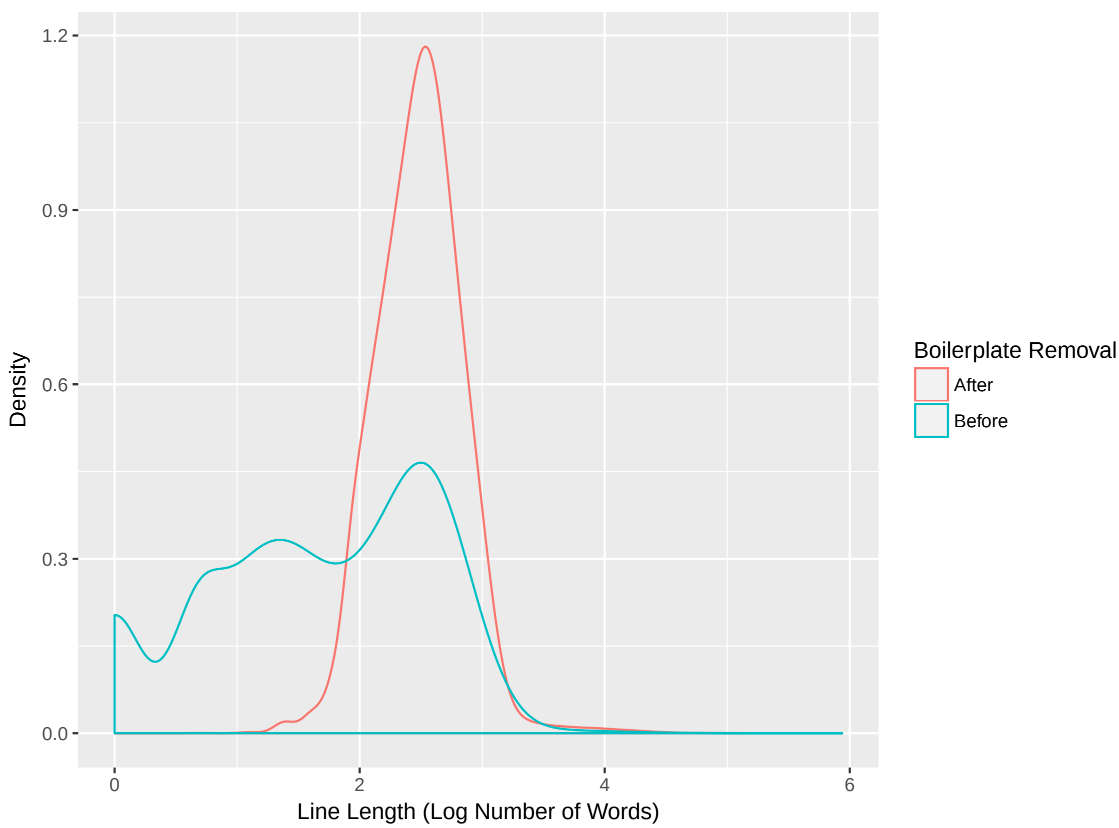


Figure 3: Effects of the boilerplate classifier on the corpus of the city of Anchorage, AK. After the boilerplate content is removed, extremely short lines are less common.

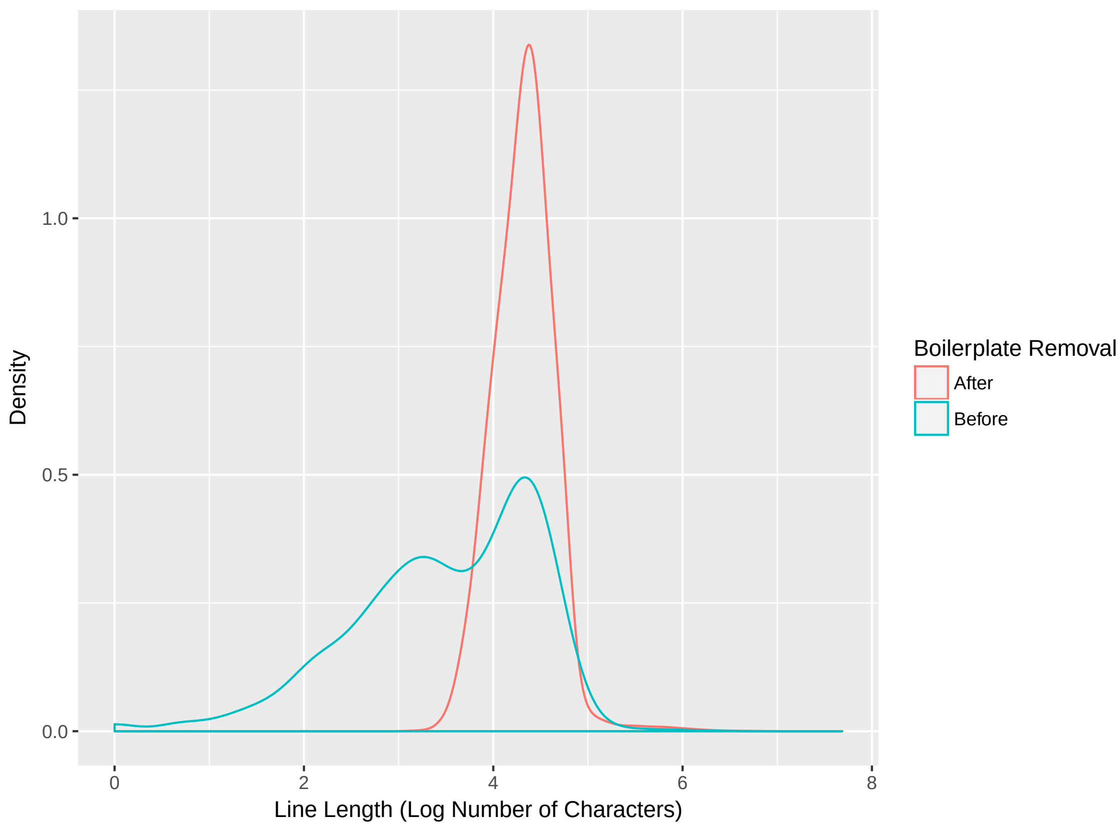


Figure 4: Effects of the boilerplate classifier on the corpus of the city of Anchorage, AK. After the boilerplate content is removed, lines that are duplicated hundreds or thousands of times are less common.

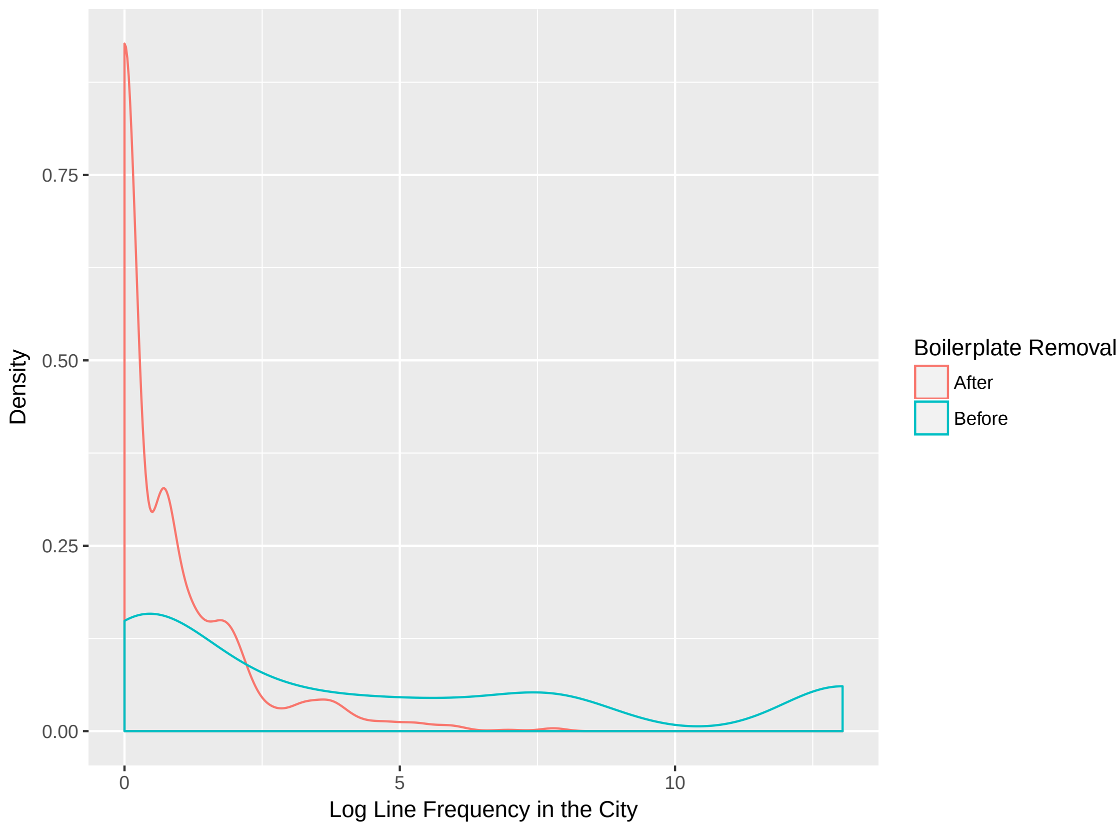
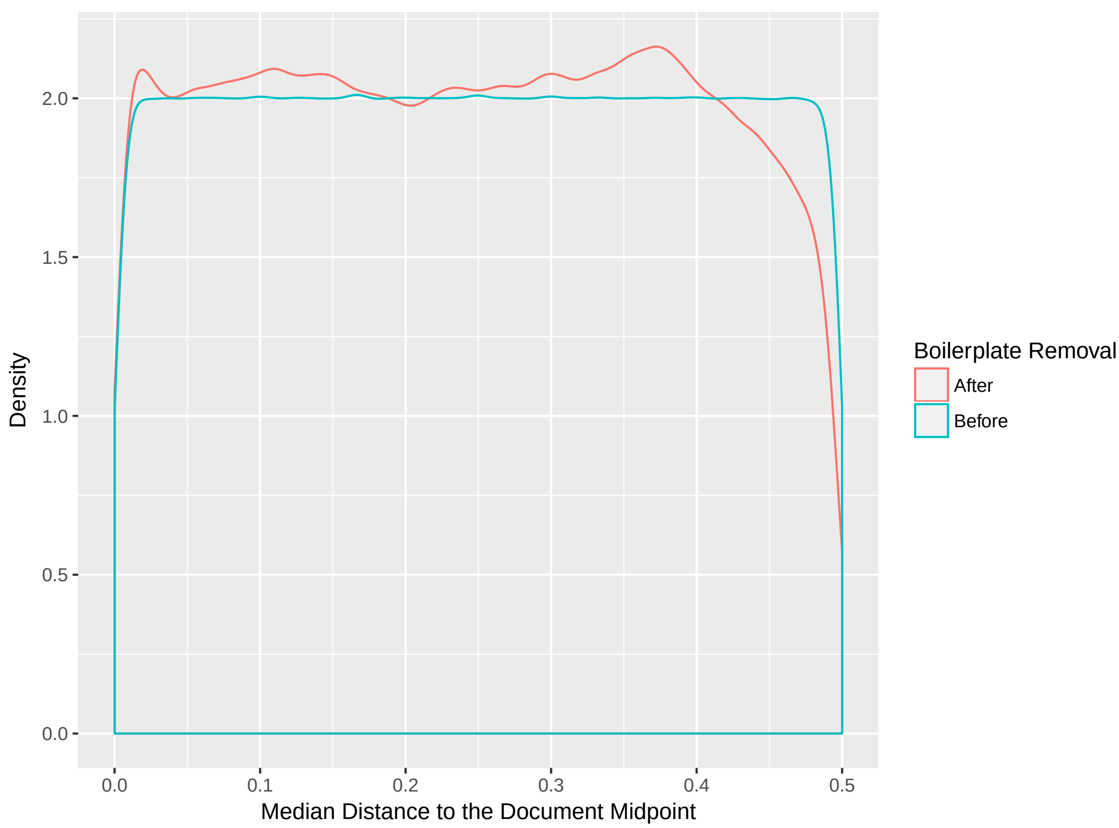


Figure 5: Effects of the boilerplate classifier on the corpus of the city of Anchorage, AK. After the boilerplate content is removed, lines at the beginning and end of documents are less common.



significantly related to spending in most other areas, where cities have less discretion. As opposed to budget allocation or even the presence/absence of policies, cities have a great deal of discretion regarding what is emphasized on their websites, and how policies are framed. Since (1) city governments have great discretion in composing their websites, (2) modifying website content is low cost relative to other policy changes, and (3), as reviewed above, city websites provide an effective and often-used means of communication with city residents.

Boilerplate Probability	Text
1.00	ri ng h i l
1.00	any others
1.00	premium for
1.00	ra tr
1.00	oa od
1.00	te pos
1.00	design to
1.00	i pli
1.00	wherever possible
1.00	public nuisances
1.00	prime movers
1.00	of the offense
1.00	d nonconformities
1.00	yearshasexceeded costs
1.00	f street

Table 4: Lines in the corpus of Anchorage, AK, with a probability of 1 of being classified as boilerplate. This table illustrates that lines which get classified as boilerplate largely consist of irrelevant nonsense which needs to be removed.

5.0.1 Analytical approach: Structural topic modeling

In order to analyze content differences between government websites based on mayoral partisanship, we draw upon a recently-developed class of text-as-data methods, the structural topic model, developed by Roberts, Stewart, Tingley, Lucas, Leder-Luis, Gadarian, Albertson and Rand (2014). Building on the conception of "topics" in Latent Dirichlet Allocation, in the structural topic model a topic is a multinomial distribution defined on the word types in the dictionary. Each word occurrence in a document is attributed with a single topic label. The word topic assignments are also drawn from a multinomial distribution. The log-odds of the topic probabilities in each document-specific multinomial distribution over topics are drawn from a multivariate normal distribution in which the topic-specific means are determined by a linear regression function that associates document-attributed covariates with topics. For example, in the context of municipal website content, the structural topic model can be used to estimate a regression coefficient that defines the linear relationship between the log-odds of the municipality's population and the log-odds of each topic. For our primary empirical investigation, the structural topic model provides with a tool with which to estimate the relationship between the party of the city's mayor and the prevalence of each topic we estimate. A topic interpreted through post-hoc analysis of the collective meaning of the most likely words to be drawn from the multinomial distribution defined by the respective topic.[12]

The structural topic model is implemented in the R package `STM` (Roberts, Stewart and Tingley 2018). We use 60 topics—the number recommended by the authors[13] for medium- to large-sized corpora. Since our corpus is at the larger end of that spectrum, the appendix also contains the results of a model with 120 topics, which corroborates the findings of the one presented here. We use four covariates: First, *party*, to estimate the difference in topic prevalence based on whether mayors are Republican or Democratic. Second, *city population*, which the literature frequently emphasizes as a determinant of the issues a city faces (see, for example, Guillamón, Bastida and Benito (2013)). Third, we control for wealth by relying on *median income* as a covariate, which we use as a proxy for the tax base in a city. Fourth and finally, we include state dummy variables, which should account for language that is associated with state-specific issues, and general background variables

[12]The "Fightin' Words" methodology developed by Monroe, Colaresi and Quinn (2008) could also be used to analyze word-frequency differences between cities based on mayors' partisanship, but we elected to use the structural topic model since, unlike "Fightin' Words" , the structural topic model enables us to adjust for several other features through multiple regression.

[13]For this recommendation, see the documentation for the function `stm()` in version 1.3.0 of the R package `stm` (Roberts, Stewart and Tingley 2018).

that vary across states.

5.0.2 Structural topic model results

The results are shown in Table 5. The rows of the table are sorted so that the most Republican topics (marked by a deeper red color) appear at the top, and the most Democratic ones (blue) at the bottom. The degree of partisanship shown in the table is determined by the size and direction of the coefficient of the party variable for each topic. The topics that are entirely white have 90% credible intervals on the effect of the mayoral party that include zero.

Many of the topics associated with Democrats fit with what we understand to be national party priorities. Topic 21, on affordable housing, clearly resonates with the Democratic party's appeal to low-income voters. Similarly, employee rights are represented in Topic 47. Democrats also exhibit a strong preference for words related to public finances, such as Topic 32 ('budget', 'revenue', 'expenditure') as well as Topic 19 ('debt', 'bond', 'financial'). We suspect that the association of Democratic mayors with finance-related terms is indicative of a greater willingness to emphasize the city's efforts to raise and spend money. This finding is consistent with (Einstein and Kogan 2015), who show that Democratic mayors tend to favor greater spending. A second, consistent Democratic focus appears to be law enforcement: The most Democratic topic, 55 ('robbery', 'homicide', 'sergeant') (a comparable topic is also the most Democratic topic in the model with 120 topics in tables 6 and 7 of the Appendix) depicts Democrats' complicated relationship with law enforcement. On the one hand, Democratic partisans have a more negative perception of the police, rating it considerably more negatively on the appropriate use of force and the equal treatment of minorities (Brown 2017). On the other hand, the literature has also shown that cities with a higher Democratic vote share spend more on the police, even after controlling for crime (Einstein and Kogan 2015). Finally, Democrats also focus more on the deliberative process of policymaking, as topics 31 ('agenda', 'committee'), 34 ('comment', 'draft', 'feedback'), 48 ('absent', ‘aye’, ‘nay’), and 37 (’audit’, ’procedure’, ’oversight’) attest to. This openness regarding the policy process on behalf of cities with Democratic mayors fits with the findings of Grimmelikhuijsen and Welch (2012), which are that left-wing local governments exhibit greater transparency via website content.

City websites with Republican mayors, meanwhile, exhibit a pronounced focus on the essential functions of government. Basic utilities such as energy (Topic 7), fire protection (Topic 17), drinking water (53), and garbage removal (Topic 49) are included among those topics that are more prevalent in cities with Democratic mayors. Similarly, protecting citizens from natural disasters is a focus in topics 1 (’storm’, ’runoff’, ’drainage’) and 42 (’breastfeed’, ’infection’, ’mosquito’ – and so, essentially, about the Zika virus), which may reflect the greater prevalence of Republican mayors in the southeast, a region which is more often affected by hurricanes and tropical diseases.

6 Conclusion

We have developed a methodological pipeline for automatically gathering and preparing government websites for comparative content analysis. This methodology holds the potential to vastly scale up the data collection efforts underpinning the growing body of research that is focused on government website analysis. The methods involved in the pipeline include checking the integrity of the site addresses, extracting English language from website contents, identifying document types, and removing boilerplate language. Government website contents can, and have been, used to study questions related representation, government transparency, and public policy priorities. The methods we develop can be used to scale up projects involving website data collection by automating the process. Through an application to the analysis of municipal websites in six different states, we show how our pipeline is capable of gathering corpora that shed light on the forms and functions of local government. We find that government website contents are associated with the partisanship of the mayor in ways that would be expected based on the parties’ national priorities

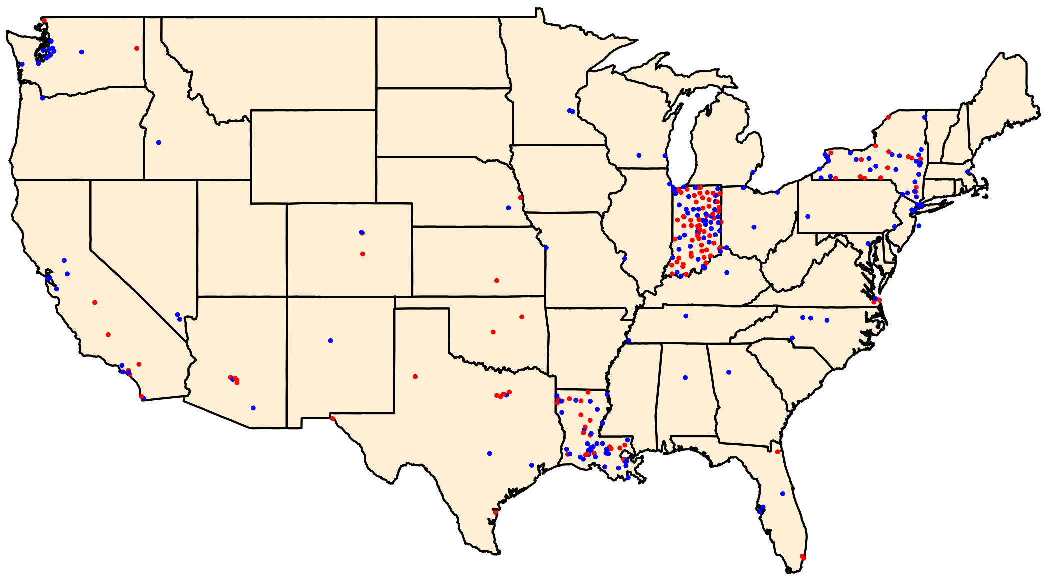


Figure 6: Map of the cities in the corpus in the contiguous U.S. The corpus also includes Alaska. In the analysis, only cities in California, Indiana, Louisiana, New York, Texas and Washington are used. The colors represent the partisanship of the mayor (blue corresponding to Democrats and red to Republicans).

and past research on the effects of mayoral partisanship on city governments.

We offer several contributions that will be valuable in future research endeavors. First, the data collected in the current study can be used for comparative analysis of US city website contents. Second, the pipeline we present can be used as a set of procedures to follow in gathering large-scale datasets of textual contents from other samples of governments. For example, the pipeline we have developed could be used to build comparative datasets of state or federal bureaucratic agencies. Third, our findings regarding the effects of mayoral partisanship on city website contents advance the literature on the role of partisan leadership in local government, and reinforce the finding of Gerber and Hopkins (2011) that the effects of mayoral partisanship can be best observed through the analysis of domains of government (e.g., website contents) that are not heavily constrained by state or national governments.

References

Armstrong, Cory L. 2011. "Providing a clearer view: An examination of transparency on local government websites." *Government Information Quarterly* 28(1):11–16.

Benoit, Kenneth and Adam Obeng. 2018. *readtext: Import and Handling for Plain and Formatted Text Files*. R package version 0.71.
URL: *https://CRAN.R-project.org/package=readtext*

Brown, Anna. 2017. "Republicans more likely than Democrats to have confidence in police.".
URL: *http://www.pewresearch.org/fact-tank/2017/01/13/republicans-more-likely-than-democrats-to-have-confidence-in-police/*

Bureau of Economic Analysis. 2017. "Per capita real GDP by state (chained 2009 dollars).".
URL: *https://www.bea.gov/iTable/drilldown.cfm?reqid=70&stepnum=11&AreaTypeKeyGdp=1&GeoFipsGdp=XX& 1&YearGdpEnd=-1&UnitOfMeasureKeyGdp=levels&RankKeyGdp=1&Drill=1&nRange=5*

Burgess, Matthew, Eugenia Giraudy, Julian Katz-Samuels, Joe Walsh, Derek Willis, Lauren Haynes and Rayid Ghani. 2016. The Legislative Influence Detector: Finding Text Reuse in State Legislation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM pp. 57–66.

Cryer, J. E. 2017. "Candidate Identity and Strategic Communication." pp. 1–42.

Darwin, IF. 2008. "Libmagic.".

de Benedictis-Kessner, Justin and Christopher Warshaw. 2016. "Mayoral partisanship and municipal fiscal policy." *The Journal of Politics* 78(4):1124–1138.

Denny, Matthew J and Arthur Spirling. 2018. "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it." *Political Analysis* 26(2):168–189.

DeSantis, Victor S and Tari Renner. 2002. "City government structures: An attempt at clarification." *State and Local Government Review* 34(2):95–104.

Druckman, James N., Cari Lynn Hennessy, Martin J. Kifer and Michael Parkin. 2010. "Issue Engagement on Congressional Candidate Web Sites, 2002—2006." *Social Science Computer Review* 28(1):3–23.
URL: *http://journals.sagepub.com/doi/10.1177/0894439309335485*

Druckman, James N., Martin Kifer and Michael Parkin. 2009. "Campaign Communications in U.S. Congressional Elections." *American Political Science Review* 103(03):343–366.
URL: *http://www.journals.cambridge.org/abstract_S0003055409990037*

Duarte, Eugenio. 2017. "The Un/Deniable Threat to LGBTQ People." *Contemporary Psychoanalysis* pp. 1–6.

Einstein, Katherine Levine and David M Glick. 2016. "Mayors, partisanship, and redistribution: Evidence directly from US mayors." *Urban Affairs Review* p. 1078087416674829.

Einstein, Katherine Levine and Vladimir Kogan. 2015. "Pushing the City Limits: Policy Responsiveness in Municipal Government." *Urban Affairs Review* pp. 1–30.

Eschenfelder, Kristin R, John C Beachboard, Charles R McClure and Steven K Wyman. 1997. "Assessing U.S. federal government websites." *Government Information Quarterly* 14(2):173–189.
URL: *http://www.sciencedirect.com/science/article/pii/S0740624X97900186%5Cnpapers2://publication/doi/10.1016 624X(97)90018-6*

Esterling, Kevin M, David Lazer and Michael A Neblo. 2011. "Representative Communication: Website Interactivity & " Distributional Path Dependence " in the U.S . Congress.".

Esterling, Kevin M. and Michael A. Neblo. 2011. "Explaining the Diffusion of Representation Practices among Congressional Websites." *Working Paper* pp. 1–42.

Feeney, Mary K. and Adrian Brown. 2017. "Are small cities online? Content, ranking, and variation of U.S. municipal websites." *Government Information Quarterly* 34(1):62–74.
URL: *http://dx.doi.org/10.1016/j.giq.2016.10.005*

Gerber, Elisabeth R and Daniel J Hopkins. 2011. "When mayors matter: estimating the impact of mayoral partisanship on city policy." *American Journal of Political Science* 55(2):326–339.

Glez-Peña, Daniel, Anália Lourenço, Hugo López-Fernández, Miguel Reboiro-Jato and Florentino Fdez-Riverola. 2013. "Web scraping technologies in an API world." *Briefings in bioinformatics* 15(5):788–797.

Grimmelikhuijsen, Stephan G. 2010. "Transparency of Public Decision-Making: Towards Trust in Local Government?" *Policy & Internet* 2(1):5–35.

Grimmelikhuijsen, Stephan G and Eric W Welch. 2012. "Developing and testing a theoretical framework for computer-mediated transparency of local governments." *Public administration review* 72(4):562–571.

Guillamón, Ma Dolores, Francisco Bastida and Bernardino Benito. 2013. "The electoral budget cycle on municipal police expenditure." *European Journal of Law and Economics* 36(3):447–469.

Ho, Alfred Tat-Kei and Wonhyuk Cho. 2017. "Government Communication Effectiveness and Satisfaction with Police Performance: A Large-Scale Survey Study." *Public Administration Review* 77(2):228–239.

Kirby, Reid. 2017. "The Trump?s administration?s misaligned approach to national biodefense." *Bulletin of the Atomic Scientists* 73(6):382–387.

Linder, Fridolin, Bruce A Desmarais, Matthew Burgess and Eugenia Giraudy. Forthcoming. "Text as Policy: Measuring Policy Similarity Through Bill Text Reuse." *Policy Studies Journal* .

Lucas, Christopher, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer and Dustin Tingley. 2015. "Computer-assisted text analysis for comparative politics." *Political Analysis* 23(2):254–277.

Marion, Nancy E and Willard M Oliver. 2013. "When the Mayor Speaks... Mayoral Crime Control Rhetoric in the Top US Cities: Symbolic or Tangible?" *Criminal justice policy review* 24(4):473–491.

Marschall, Melissa and Paru Shah. 2013. "Local Elections in America Project." *Center for Local Elections in American Politics. Kinder Institute for Urban Research, Rice University.(Database)* .

URL: *http://www.leap-elections.org/*

Mayhew, David. 1974. *Congress: The Electoral Connection*. Yale University Press.

McNutt, Kathleen. 2010. "Virtual policy networks: Where all roads lead to Rome." *Canadian Journal of Political Science/Revue canadienne de science politique* 43(4):915–935.

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 SPEC. ISS.):372–403.

Morgan, David R and Sheilah S Watson. 1992. "Policy leadership in council-manager cities: Comparing mayor and manager." *Public Administration Review* pp. 438–446.

Norris, P. 2003. "Preaching to the Converted?: Pluralism, Participation and Party Websites." *Party Politics* 9(1):21–45.

Ooms, Jeroen. 2017. *hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker*. R package version 2.9.
URL: *https://CRAN.R-project.org/package=hunspell*

Osman, Ibrahim H, Abdel Latef Anouze, Zahir Irani, Baydaa Al-Ayoubi, Habin Lee, Asım Balcı, Tunç D Medeni and Vishanth Weerakkody. 2014. "COBRA framework to evaluate e-government services: A citizen-centric perspective." *Government Information Quarterly* 31(2):243–256.

Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. 2018. *stm: R Package for Structural Topic Models*. R package version 1.3.3.
URL: *http://www.structuraltopicmodel.com*

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.

Rudis, Bob, Christos Zoulas, Mans Rullgard and Jonathan Ong. 2016. *wand: Retrieve 'Magic' Attributes from Files and Directories*. R package version 0.2.0.
URL: *https://CRAN.R-project.org/package=wand*

Sandoval-Almazan, Rodrigo and J Ramon Gil-Garcia. 2012. "Are government internet portals evolving towards more interaction, participation, and collaboration? Revisiting the rhetoric of e-government among municipalities." *Government Information Quarterly* 29:S72–S81.

Sharfstein, Joshua M. 2017. "Science and the Trump Administration." *Jama* 318(14):1312–1313.

Stanyer, James. 2008. "Elected representatives, online self-presentation and the personal vote: Party, personality and webstyles in the United States and United Kingdom." *Information, Community & Society* 11(3):414–432.

Therriault, Andrew. 2010. "Taking Campaign Strategy Online: Using Candidate Websites to Advance the Study of Issue Emphases." pp. 1–23.
URL: *http://poseidon01.ssrn.com/delivery.php?ID=58812509611308010110700710910410112103503107705401*

Thomas, John Clayton and Gregory Streib. 2003. "The new face of government: citizen-initiated contacts in the era of E-Government." *Journal of public administration research and theory* 13(1):83–102.

Tolbert, Caroline J and Karen Mossberger. 2006. "The effects of e-government on trust and confidence in government." *Public administration review* 66(3):354–369.

Urban, Florian. 2002. "Small town, big website? Cities and their representation on the internet." *Cities* 19(1):49–59.

Wang, Lili, Stuart Bretschneider and Jon Gant. 2005. Evaluating web-based e-government services with a citizen-centric approach. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on.* Ieee pp. 129b–129b.

Wilkerson, John, David Smith and Nicholas Stramp. 2015. "Tracing the flow of policy ideas in legislatures: A text reuse approach." *American Journal of Political Science* 59(4):943–956.

Appendix

#	Top Word 1	Top Word 2	Top Word 3	Top Word 4	Top Word 5	Top Word 6	Tokens assigned
43	fun	player	dream	celebration	favorite	blog	3460
5	please	email	contact	copy	mail	click	201
42	breastfeed	vaccine	infection	symptom	asthma	mosquito	2497
17	alarm	disaster	fire	rescue	preparedness	evacuation	989
53	drinking	wastewater	water	pipeline	pump	disinfection	461
50	buffalo	news	honor	warren	announce	lovely	1106
52	reappoints	digest	cat	leg	legislator	sander	997
33	really	think	something	thing	somebody	anybody	1873
44	shall	herein	forth	deem	thereof	pursuant	405
8	invoice	card	amt	filer	debit	officeholder	527
26	fee	charge	billing	per	meter	monthly	233
2	yon	borough	comm	gen	sou	spec	709
49	bin	recycling	garbage	recyclables	recyclable	bag	1791
7	energy	garland	renewable	solar	electricity	climate	742
23	bid	proposer	bidder	contractor	subcontractor	contract	447
57	duct	conduit	bolt	splice	valve	fitting	1373
13	server	wireless	software	telecommunication	subscriber	desktop	1092
54	motion	adjourn	second	unanimously	ayes	carry	474
1	storm	runoff	infiltration	discharge	drainage	drain	516
38	youth	student	parent	teacher	immigrant	literacy	714
35	artist	rouge	baton	art	artwork	exhibition	1632
59	sampling	sample	analytical	concentration	hydrocarbon	toxicity	1241
3	portfolio	yield	jun	maturity	investment	rating	544
45	premise	licensee	violation	license	permit	inspection	509
9	para	persona	ante	horas	junta	largo	1469
60	exhaust	fugitive	aircraft	airport	aviation	diesel	731
30	fort	thence	blvd	worth	ave	west	681
58	councilor	auburn	plain	ward	beech	glen	480
51	whereas	councilman	alderman	ordain	hereby	resolution	420
16	recreation	park	golf	playground	picnic	zoo	682
36	retiree	retirement	actuarial	deductible	dental	pension	470
27	exam	incumbent	supervise	supervision	examination	knowledge	687
56	historic	landmark	revival	archaeological	century	historian	2587
12	parking	hotel	garage	space	retail	square	321
41	tax	exemption	abatement	real	estate	property	310
4	facade	awning	porch	roof	balcony	exterior	1108
28	census	population	respondent	figure	percent	margin	541
18	prune	tree	deer	forestry	shrub	bulrush	2522
15	complainant	defendant	allegation	complaint	allege	discrimination	1384
20	noise	mitigation	impact	adverse	significant	vibration	325
14	yes	agency	federal	recipient	compliance	entity	205
46	variance	setback	plat	zoning	yard	fence	289
29	learn	neighborhood	graffito	event	resident	online	196
25	cannabis	marijuana	senate	dispensary	ballot	cultivation	1188
22	priority	strategic	ongoing	goal	implementation	implement	141
6	project	improvement	phase	replacement	upgrade	capital	174
11	shoreline	beach	marina	coastal	waterfront	salmon	1069
24	attract	economy	workforce	innovation	sector	economic	748
47	employee	overtime	sick	wage	grievance	bargaining	511
39	tab	accessibility	mode	var	alt	false	259
10	density	village	urban	us	mixed	corridor	358
37	audit	auditor	internal	procedure	accountability	oversight	420
21	housing	affordable	homeless	homelessness	affordability	landlord	318
34	comment	draft	feedback	stakeholder	suggest	discussion	289
19	debt	bond	governmental	obligation	financial	accounting	251
40	bicycle	bike	lane	crosswalk	pedestrian	bicyclist	574
32	budget	revenue	expenditure	appropriation	fund	million	242
48	absent	aye	khan	nay	berry	voting	528
31	chair	agenda	commission	speaker	chairperson	committee	314
55	robbery	homicide	arrest	sergeant	suspect	burglary	1395

Table 5: Top words from a structural topic model with 60 topics and FREX scoring. Colors depict partisanship based on coefficient size. White cells are non-significant topics.

#	Top Word 1	Top Word 2	Top Word 3	Top Word 4	Top Word 5	Top Word 6	Tokens assigned
115	garland	celebration	blog	dream	sorry	copyright	994
52	dog	legislator	spay	neuter	animal	microchip	761
44	copy	record	mail	request	notice	notify	120
98	neighborhood	community	resident	safe	life	quality	95
88	war	professor	sister	bachelor	daughter	soldier	2516
43	camp	yoga	camper	fun	librarian	library	1080
42	infection	tuberculosis	breastfeed	hepatitis	vaccine	condom	1608
72	drinking	water	contaminant	reservoir	pipeline	irrigation	216
84	say	ask	explain	reply	horn	advise	454
18	player	coach	game	umpire	ball	shirt	1595
61	unanimously	motion	prince	adjourn	carry	ken	192
63	mosquito	spray	rodent	pesticide	repellent	pest	851
81	effluent	sludge	lbs	mercury	wastewater	gal	540
60	shall	deem	forth	unless	except	thereof	119
69	ethic	candidate	lobbyist	filer	political	officeholder	355
33	think	really	something	thing	just	go	826
119	firefighter	fire	chief	police	captain	patrol	248
37	physician	nursing	medical	nurse	outpatient	medicaid	352
5	home	homeowner	alarm	detector	monoxide	header	209
23	proposer	bidder	subcontractor	bid	contractor	subcontract	239
116	councilor	alderman	councilwoman	alderwoman	quill	councilors	268
15	trademark	borough	new	immigration	immigrant	pour	274
67	discrimination	disability	gender	religion	accommodation	origin	373
117	asthma	overdose	obesity	hospitalization	diabetes	prevalence	659
94	duct	valve	sprinkler	combustible	splice	conductor	778
58	event	firework	parade	press	holiday	troy	335
70	whereas	hereby	resolve	duly	authorize	therefore	202
30	disaster	emergency	preparedness	evacuation	dispatch	homeland	365
38	student	parent	school	teacher	academic	youth	354
93	city	fort	worth	manager	hall	charter	16
75	online	click	plain	website	download	learn	165
3	value	market	productivity	customize	yrs	index	126
49	recycling	recycle	garbage	waste	trash	landfill	408
111	franchisee	indemnify	arise	harmless	breach	party	307
17	snow	plow	tornado	flood	pothole	crew	552
89	vend	food	meat	utensil	calorie	vending	1174
45	application	applicant	certificate	must	license	permit	151
85	runoff	sanitary	infiltration	storm	drainage	drain	241
106	equipment	boiler	fleet	crane	mechanic	fuel	539
8	invoice	payment	card	credit	account	cash	187
13	class	test	adobe	embed	reader	acrobat	312
108	cigarette	senate	tobacco	consumer	smoking	ban	542
25	coal	hazard	hazardous	toxic	radiation	substance	288
86	groundwater	sample	asbestos	analytical	remediation	remedial	345
1	golf	exhibit	lessee	course	lessor	lease	401
9	para	persona	ante	horas	junta	sin	635
24	phone	name	page	address	glen	cove	158
7	energy	renewable	solar	electricity	climate	efficiency	399
66	plat	thence	easement	pud	tract	subdivision	230
57	dwell	unit	remodel	condominium	dwelling	residential	167
95	roof	masonry	porch	exterior	would	brick	611
26	fee	charge	per	cost	plus	rate	102
51	chapter	code	violation	subsection	article	sec	151
59	zoning	conditional	zone	cannabis	overlay	district	241
101	height	foot	square	feet	setback	frontage	124
96	house	cemetery	burial	butler	funeral	barber	472
65	ballot	vista	ranch	canyon	silicon	voter	518
120	bend	fir	hometown	twelfth	exceptional	rodeo	271
36	aviation	airport	airline	runway	aircraft	hangar	429
34	plan	planning	comprehensive	master	review	amendment	42

Table 6: Top words from a structural topic model with 120 topics (first 60 topics displayed here) and FREX scoring. Colors depict partisanship based on coefficient size. White cells are non-significant topics.

#	Top Word 1	Top Word 2	Top Word 3	Top Word 4	Top Word 5	Top Word 6	Tokens assigned
82	com	mar	spec	jun	est	comm	1388
22	server	software	wireless	technology	desktop	broadband	430
80	artist	art	artwork	exhibition	artistic	sculpture	1099
113	trench	thickness	compaction	concrete	slab	excavation	766
87	respondent	survey	census	racial	demographic	score	427
83	homeless	homelessness	supportive	client	transitional	encampment	229
20	noise	fugitive	receptor	exhaust	vibration	emission	376
35	landlord	tenant	owner	property	rent	lien	205
105	beach	orange	arena	rainier	ocean	resort	457
2	yon	bay	gen	sou	coliseum	estuary	385
6	redevelopment	land	developer	parcel	development	area	70
104	riparian	wetland	habitat	marsh	floodplain	grassland	968
41	tax	exemption	taxable	deduction	levy	taxpayer	172
68	economy	workforce	economic	sector	industry	innovation	332
28	figure	table	scenario	margin	analysis	appendix	207
110	bond	maturity	debt	issuer	redemption	obligation	232
102	sidewalk	curb	pole	crosswalk	ramp	sign	237
118	project	phase	construction	completion	improvement	complete	45
78	parking	tow	vehicle	garage	car	motor	210
71	actuarial	retiree	retirement	pension	deductible	unfunded	239
91	prune	tree	forestry	deer	shrub	planting	1240
114	incumbent	exam	supervision	supervise	examination	ability	432
16	park	recreation	playground	zoo	trail	picnic	290
53	waterfront	boat	shoreline	maritime	dock	barge	800
76	felony	violent	offender	gang	theft	inmate	783
4	courtyard	realm	design	facade	proponent	articulation	608
100	division	manage	staffing	oversee	management	analyst	100
97	mitigation	impact	adverse	significant	alternative	propose	132
11	historic	landmark	revival	archaeological	preservation	historical	876
77	million	fiscal	forecast	revenue	quarter	billion	138
74	board	chairperson	secretary	member	appoint	executive	118
47	allegation	complainant	misconduct	bias	complaint	allege	580
92	sick	employee	wage	overtime	grievance	bargaining	260
10	ave	avenue	south	east	west	blvd	189
112	grant	loan	funding	program	recipient	federal	85
56	downtown	mall	midtown	uptown	hotel	shopping	414
14	yes	agency	successor	oversight	attachment	describe	125
40	bicycle	bike	transit	bicyclist	lane	bus	315
62	affordable	housing	affordability	income	household	moderate	188
99	memorandum	resolution	council	legislation	entitle	commission	173
19	governmental	accounting	asset	statement	financial	net	156
103	permission	ayes	correspondence	bid	smith	demolition	203
107	appropriated	dollars	thousand	ongoing	matrix	justification	117
12	approach	difficult	achieve	challenge	critical	often	257
46	variance	fence	setback	exception	yard	applicant	136
90	audit	auditor	procedure	internal	auditing	documentation	226
64	density	urban	corridor	village	orient	transit	165
21	goal	strategy	outreach	priority	strategic	implementation	105
73	parish	rouge	baton	hogan	councilman	thereto	482
29	comment	draft	discussion	feedback	discuss	presentation	168
32	budget	expenditure	appropriation	fund	endorse	balance	129
54	aye	absent	khan	nay	berry	voting	344
39	mode	accessibility	tab	focus	else	alt	117
109	auburn	buffalo	ward	brown	announce	casino	177
50	news	warren	lovely	release	leader	proud	498
79	digest	proposal	sander	reappoints	metropolitan	gray	236
27	bankruptcy	plaintiff	creditor	trial	court	supreme	810
31	agenda	speaker	item	committee	chair	divided	146
48	consolidated	contingency	reinvestment	inc	contract	authorize	134
55	suspect	shoot	fatal	homicide	stopper	pronounce	512

Table 7: Top words from a structural topic model with 120 topics (second 60 topics displayed here) and FREX scoring. Colors depict partisanship based on coefficient size. White cells are non-significant topics.