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

Markus Neumann Fridolin Linder Bruce Desmarais

July 13, 2018

Abstract

A local government's website is arguably the most important general source of information about city policies and processes 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. Relying on manual data collection requires that researchers focus on a limited number of websites and/or limited types of site content. 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.

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 text 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 and other programming languages, and discarding any other character encoding errors that result from reading the files. 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 Democratic mayor, Joseph Schember, along with a list of laudable attributes of the city. In local politics, where campaign funds are low, this lends

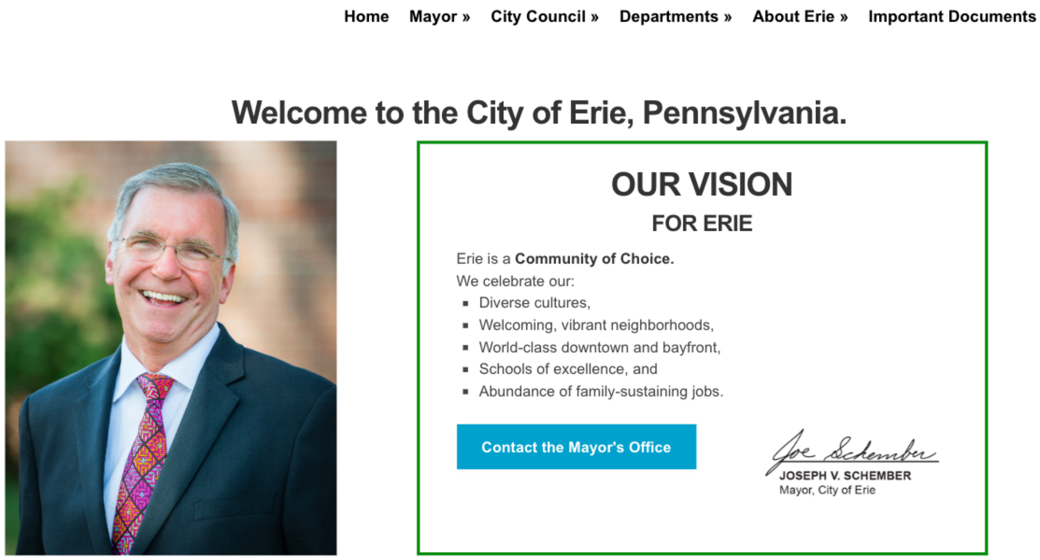


Figure 1: Screenshot from the homepage at http://www.erie.pa.us/, accessed on 06/14/2018. Image depicts Democratic mayor of Erie, PA, Joseph Schember.

the incumbent a crucial advantage in becoming more well-known among her constituents. 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) finds that users of local government Web sites 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 literature making use of scraped websites clusters into a number of categories. One, and most pertinent to our own endeavors, the e-governance literature which discusses 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 (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.

Websites have also played a major role in the field of media studies, as scholars have scraped and analyzed the online presence of newspapers, as well as the more diffuse world of online political blogs (Adamic and Glance 2005; Gentzkow and Shapiro 2010). Lin, Bagrow and Lazer (2011) provide a good example for a study which makes extensive use of automated content analysis - a necessity arising from its dataset of 66830 blog posts and 57221 online news articles. The authors estimate the political slant of these entities by counting the frequencies with which politicians of either side are mentioned and determine that blogs are generally more biased. Unfortunately for us, the authors don't go into the details of their text analysis, and offer no information on the acquisition and pre-processing of the data.

Another well-known example fitting into this area of study is the set of studies conducted by King et al. (King, Pan and Roberts 2013, 2014, 2017), in which the authors study censorship by the country's government on its lively blogosphere. However, the authors also provide no information on how their data was collected "our extensive engineering effort, which we do not detail here for obvious reasons [...]".

The websites of politicians and their parties have also fallen under scholarly scrutiny. Researchers have found that in order to identify the constituencies, motives and modes of communication of 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) provides fairly little information, but does mention the fact that she relied on Archive-it, a webservice of the Internet Archive. Unfortunately we found the data provided by the Internet Archive to not be sufficiently reliable and well-documented for our own purposes. 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) (a working paper) actually portends to 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, and the actual analysis is incomplete. 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).

Importantly for us, research analyzing and improving the scraping, pre-processing and analysis methods of this literature is scarce. Eschenfelder, Beachboard, McClure and Wyman (1997) provide something of an overview of how 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

In this section we introduce the data we use in our application—the analysis of municipal websites in six states - Indiana, Louisiana, New York, Washington, California and Texas. These states provide us with a sample that is well-balanced on a number of theoretically important indicators. One, each of the four geographic regions is 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 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 (**?**)), but also some of the poorer ones (IN and LA). In terms of pure GDP per capita, the sample is on the less affluent side - however, wealth is also correlated with poverty: CA is the state with the highest poverty rate in the country, and LA, NY and TX follow closely (**?**).

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. This method proved to be very reliable. Two, the General Services Administration (GSA) maintains all .gov addresses, and provides a complete[1] list of all such domains to the public through

[1]Domains used for testing and internal programs are excluded.

GitHub[2][3]. Naturally, this list does not contain cities 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.

To test whether the websites we found actually work, we use a webdriver-controlled browser (Firefox/Selenium/Geckodriver). This is necessary because a) some city websites simply don't work, and more often, b) cities sometimes change their websites' URLs, in which case they redirect from the old to the new URL. A webdriver-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.

The partisanship 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 LEAP project[5]. 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 recorded candidates who received money from party committees. For California and Texas, where our data consists of major

[2]https://github.com/GSA/data/tree/gh-pages/dotgov-domains

[3]This list is updated once per month - we rely on the version released on January 16, 2017. The data from the GSA contains the following data: One, domain name, specifically, the all-uppercase version of domain and top-level domain (for example, 'ABERDEENMD.GOV'). Two, the type of government entity to which the domain is registered, such as city, county, federal agency, etc. Three, for federal agencies, the name is specfied. Finally, the city in which the domain is registered, is noted.

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

[5]http://www.leap-elections.org/

cities, partisanship information was acquired from Ballotpedia[6]. 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) this method once again 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.

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[7].

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.

For some cities, whose websites make heavy use of JavaScript, this method does not lead to satisfying results. Consequently we restricted our corpus to cities with at least 3 documents.

4 The Web to Text Pipeline

In this methodological pipeline from native website files to text data that is appropriate for comparative analysis, 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

[6]https://ballotpedia.org/List_of_current_mayors_of_the_top_100_cities_in_the_United_States
[7]https://www.census.gov/data/developers/data-sets.html

Filetype	current	before	after
	51455	13866	19199
pdf	9646	5489	7544
jpg	5216	1988	3512
html	3767	17842	17596
aspx	2832	4356	3271
png	2714	2327	3684
gif	1068	664	1077
JPG	478	182	263
1	443	61	54
css	390	265	518
js	350	255	468
htm	264	295	256
docx	203	106	120
doc	167	70	130
asp	161	201	211
svg	87	55	69
php	83	157	241

Table 2: The most common file types in scraped websites

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 process differences between governments. We describe these methodological steps in this section.

4.1 Site to Text Conversion

For the most part, the file type of a document can be correctly determined through its ending. However, there are exceptions to this, which, if ignored, can lead to large amounts of garbage 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, which is an R interface to the Unix library libmagic, which determines the type of a file on the basis of its file signature. 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[8] - which determines a document's type solely through its ending - to convert the files to plain text.

The text documents are then read into R line by line, converted to UTF-8 and then stripped of dates, punctuation, 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 pdfs, doc and docx files. This is an issue, because it clusters documents around the cities from which they originate in a way that has nothing to do with their actual content. In other words, the signal would be drowned out by the noise. Our solution to this problem is described in more detail in section 4.2. 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 OCR. To combat this problem, we employ two solutions. One, we use spellchecking, implemented through the hunspell R package, to remove all non-English words.[9] 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

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

[9]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 relying on multilingual word embeddings. Since the removal of non-English words is very computationally-intensive, we only take this step at the end of the preprocessing process, the result of which might be a slightly adverse effect on the accuracy of the boilerplate classifier.

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 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 (none of these words are of any substantive relevance to our research question). Since these pre-processing steps reduce documents which are largely unsuitable to only a few words of texts that don't make much sense, 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 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.

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 correct analysis by automatic text processing methods. 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 this level.

Boilerplate Classification

In order to determine whether a line should be discarded, we train a simple classifier. We sampled 100 lines from documents 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. To account for the higher likelihood of some lines being part of the training set, we use inverse probability weights in the classifier.[10]

These 500 lines were then hand-coded as either substantively useful or useless. Then we trained a random forest with this usefulness measure as the dependent variable. The independent variables were: (1) number of times the line was duplicated within the city, (2) length of the line, in characters, (3) 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 following:

The length of the line and the number of tokens are a way 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 of zero substantive interest to us. These terms also happen to be fairly common, which causes them to be overweighted by the topic model.

To directly address the latter problem, a measure for the number of times a line is duplicated within a city is included. 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 create topics that are highly correlated with cities.

Finally, the distance measure: 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

[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.

of documents might have a score of 0.45, whereas a line that tends to be found more in the center, and thus be indicative of more relevant content, might be scored with a 0.11 instead.

We rely on random forests as a classifier, which offer slightly better performance than logit[11] and have the added benefit of giving estimates of variable importance. Performance of this classifier was assessed through five-fold cross-validation, the results of which can be found in table 3.

	Value
Percent Correctly Predicted	0.89
Precision	0.89
Recall	0.94
F1-Score	0.92

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

This classifier is then used to flag and remove all lines that are not classified as substantively useful. 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 looks more like what it should be - lines of medium length now occur more frequently than extremely short ones (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 within 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). After all the preprocessing is set and done, our corpus consists of 259,099 documents.

[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.

Figure 2: Effects of the boilerplate classifier on the corpus of the city of Anchorage, AK.

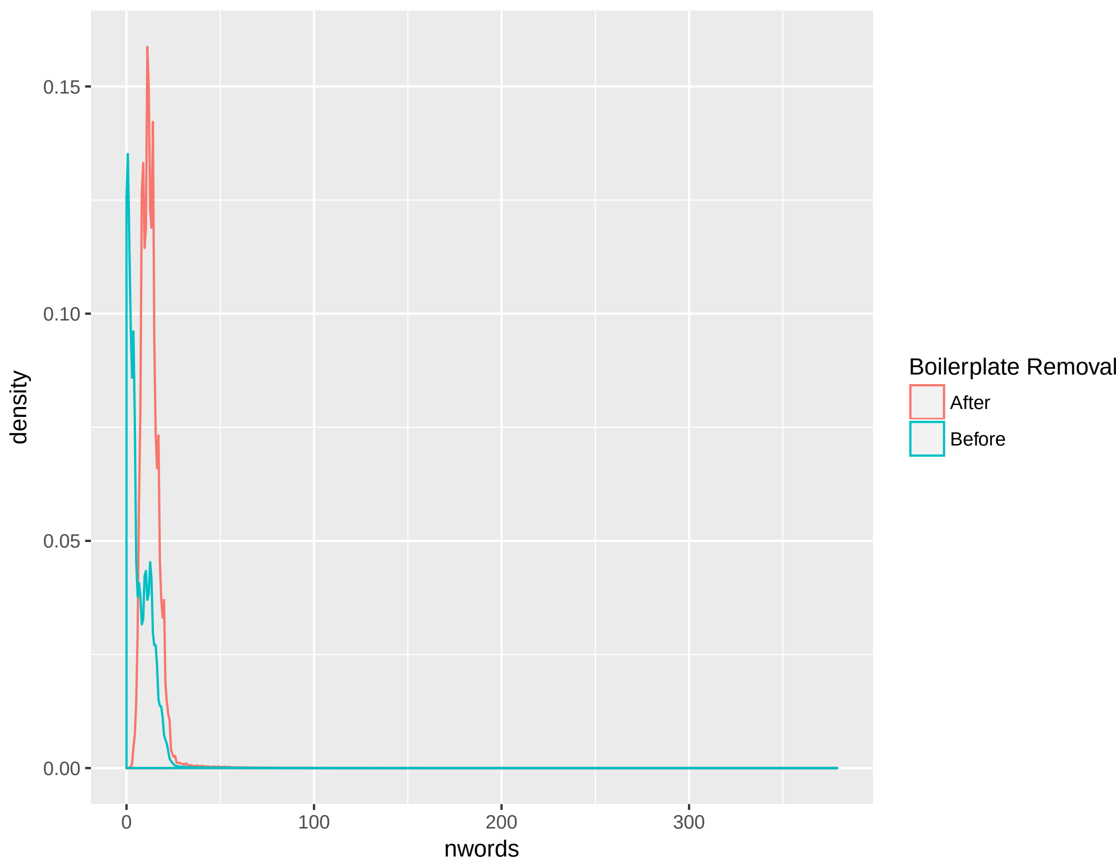


Figure 3: Effects of the boilerplate classifier on the corpus of the city of Anchorage, AK.

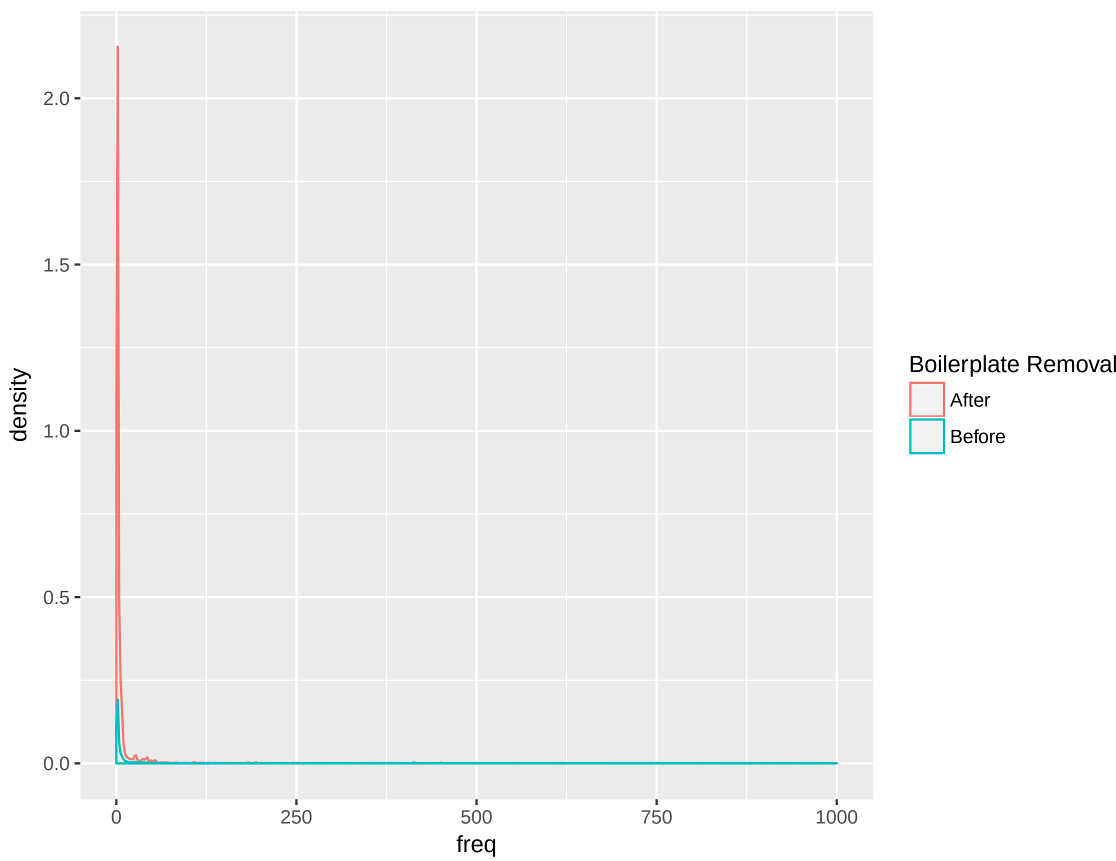


Figure 4: Effects of the boilerplate classifier on the corpus of the city of Anchorage, AK.

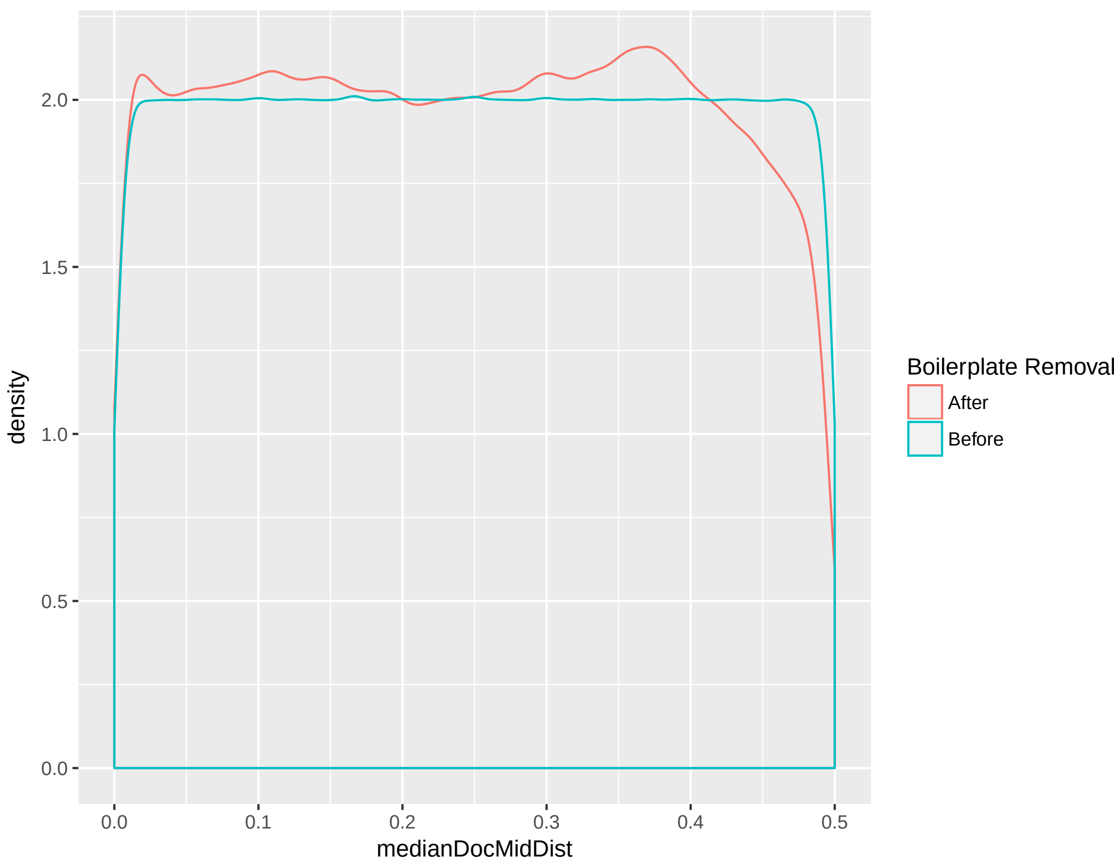
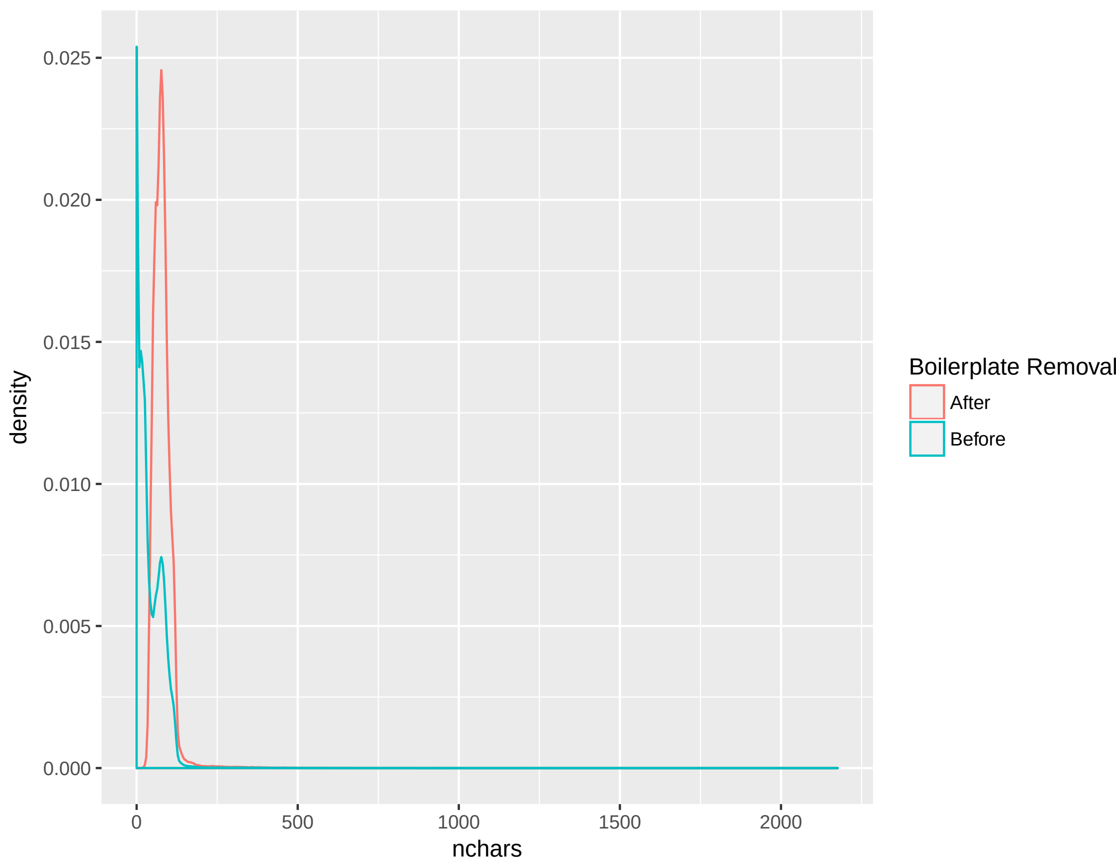


Figure 5: Effects of the boilerplate classifier on the corpus of the city of Anchorage, AK.



5 Bag-of-Words Text Analysis

We illustrate the analysis of municipal website content using bag-of-words (BoW) methods. BoW methods are methods of text analysis that do not take into account the sequence or placement of words in text—just the presence and frequency of words. As noted by Grimmer and Stewart (2013), for most applications, bag-of-words approaches have been found to be more than sufficient. Furthermore, there is reason to believe that city government websites are a particularly 'safe' case for bag-of-words methods due to their informative, manner-of-fact based language. It is extremely unlikely for these pages to feature ambiguous language such as an abundance of negation or even sarcasm.

5.0.1 Structural topic model

A powerful and frequently used approach within the family of BoW methods is the use of topic models. This class of clustering methods relies on the co-occurrence of words within documents to form a set of semantically coherent topics. In order to compare the degree to which Republicans and Democrats prefer specific topics, we rely on the structural topic model, developed by (Roberts, Stewart, Tingley, Lucas, Leder-Luis, Gadarian, Albertson and Rand 2014). Theoretically, the most widely-used form of topic model, latent dirichlet allocation, can also be used to test for the impact of a single covariate through a post-hoc comparison, but the structural topic model allows for multiple covariates, and also produced more meaningful topics in our experiments.

We use 60 topics - the number recommended by the authors 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 test our main hypothesis. 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. Fourth and finally, we include a *state* variable.

The results are shown in table 4. 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 topics that are entirely white are not statistically significant for the party covariate. In order to test statistical significance, we calculated credible intervals - the threshold chosen here is the 1% level.

Many of the topics associated with Democrats - one related to education, one to recycling - clearly seem to match the party brand. For example, topic 44 on affordable housing clearly resonates with the Democratic party's appeal to low-income voters. Similarly, employee rights are represented in topic 46. Democrats also exhibit a strong preference for words related to public finances, such as topic 22 ('budget', 'revenue', 'expenditure') as well as topic 17 ('debt', 'bond', 'financial'). We hypothesize that this 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. Democrats are the party of big government, and this clearly also shows at the city level. A second, conistent Democratic focus appears to be law enforcement: The most Democratic topic, 57 ('robbery', 'homicide', 'sergeant') (a comparable topic is also the most Democratic topic in the model with 120 topics in tables 5 and 6 of the Appendix) depicts Democrats' complicated relationship with law enforcement - a finding that is not entirely without precedent in the literature (see (Einstein and Kogan 2015)). Finally, Democrats also focus more on the deliberative process of good governance, as topics 29 ('election', 'agenda', 'committee') and 37 ('audit', 'procedure', 'oversight') attest to.

Republicans, meanwhile, live up to their reputation as the party of small government. Many topics prevalent in Republican cities focus on the bare essentials. Basic utilities such as energy (topic 7), fire protection (topic 33), drinking water (51) and garbage removal (topic 49) are ubiquitous here. Similarly, protecting citizens from natural disasters is a focus in topics 1 ('storm', 'runoff', 'drainage') and 20 ('snow', 'hurricane', 'tornado') is an important issue.

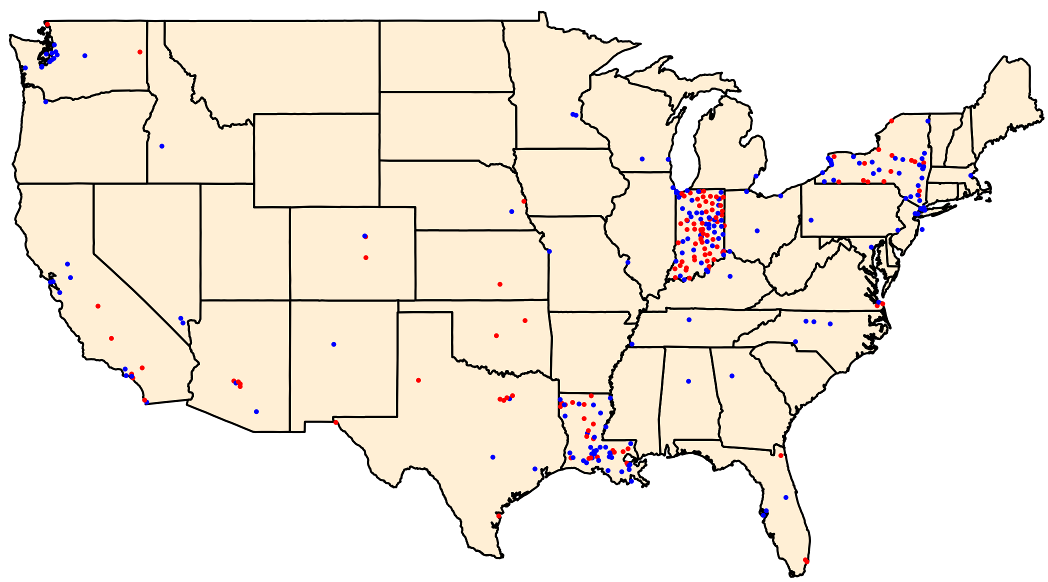


Figure 6: Cities in the corpus, by partisanship of mayor. **REVISE to remove everything not in our six states.**

6 Conclusion

We have developed a methodological pipeline for automatically gathering and preparing government websites for comparative analysis. This methodology holds the potential to vastly scale up the data collection efforts underpinning the rapidly growing body of research that is focused on government website analysis. Through an application to the analysis of municipal websites in six

#	Top Word 1	Top Word 2	Top Word 3	Top Word 4	Top Word 5	Top Word 6	Tokens assigned
38	artist	poetry	music	fun	dance	exhibition	3770
5	please	email	mail	copy	contact	click	260
43	epidemiology	infection	vaccine	antibody	asthma	hygiene	2469
20	snow	hurricane	tornado	plow	evacuate	pothole	1290
52	reappoints	legislator	cat	leg	sander	dog	1152
51	drinking	wastewater	water	pump	sludge	sewage	487
32	think	really	okay	thing	something	seem	1940
36	shall	herein	forth	deem	thereof	hereunder	433
27	library	branch	learn	book	online	view	302
58	buffalo	announce	warren	lovely	honor	ceremony	1298
33	fire	fort	worth	beach	alarm	firefighter	459
34	fee	charge	per	billing	bill	refund	241
35	youth	student	parent	school	teacher	academic	710
56	garland	auburn	councilor	plain	hall	ward	229
21	bid	proposer	subcontractor	bidder	contractor	subcontract	485
59	motion	adjourn	unanimously	second	ayes	carry	487
49	garbage	recycling	bin	recyclable	recyclables	cart	1635
31	deductible	dental	medicare	coinsurance	copay	aircraft	706
54	duct	conduit	bolt	splice	valve	piping	1477
14	immigrant	discrimination	gender	immigration	racial	refugee	1095
1	storm	runoff	drainage	infiltration	drain	discharge	490
2	yon	ave	blvd	greenwood	suite	comm	1317
4	para	persona	ante	horas	junta	largo	1377
55	alderman	whereas	hereby	ordain	resolution	resolve	457
26	sampling	petroleum	sample	concentration	hydrocarbon	pesticide	1278
48	premise	marijuana	permit	licensee	license	cannabis	489
12	server	wireless	software	digital	telecommunication	technology	917
7	energy	renewable	solar	climate	electricity	greenhouse	740
16	recreation	golf	playground	park	picnic	zoo	702
60	exhaust	air	boiler	diesel	ozone	fuel	316
3	rouge	baton	issuer	maturity	parish	jun	502
23	economic	attract	downtown	economy	industry	revitalization	862
25	incumbent	exam	supervise	supervision	knowledge	examination	683
8	actuarial	retirement	pension	contribution	retiree	valuation	289
9	facade	awning	roof	porch	balcony	exterior	1103
15	shoreline	marsh	coastal	habitat	wetland	salmon	1454
42	tax	exemption	taxable	real	abatement	property	343
30	population	census	respondent	figure	trend	comparison	540
53	historic	landmark	revival	century	historian	archaeological	2518
11	parking	vehicle	passenger	tow	garage	taxicab	435
18	prune	tree	forestry	shrub	deer	planting	2279
45	variance	plat	setback	zoning	fence	yard	300
19	noise	mitigation	impact	fugitive	adverse	significant	360
13	agency	yes	federal	entity	recipient	deficiency	239
6	improvement	project	upgrade	capital	appropriated	replacement	189
46	employee	overtime	sick	bargaining	wage	salary	398
28	allegation	complainant	defendant	misconduct	allege	bankruptcy	1747
40	tab	mode	accessibility	false	focus	else	257
10	density	us	mixed	village	urban	orient	336
39	comment	draft	review	preliminary	planning	propose	274
37	audit	auditor	internal	procedure	implement	oversight	402
44	housing	affordable	homeless	homelessness	landlord	affordability	340
17	debt	governmental	bond	obligation	financial	accounting	259
41	bicycle	bike	lane	intersection	pedestrian	crosswalk	527
24	strategy	goal	outreach	priority	strategic	stakeholder	313
50	aye	absent	nay	councilman	khan	voting	674
22	budget	revenue	expenditure	million	appropriation	forecast	236
47	digest	authorize	inc	consolidated	contingency	agreement	215
29	chair	election	agenda	committee	speaker	ballot	353
57	robbery	homicide	sergeant	arrest	suspect	crime	1255

Table 4: 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. Based on data preprocessed with the classifier.

#	Top Word 1	Top Word 2	Top Word 3	Top Word 4	Top Word 5	Top Word 6	Tokens assigned
93	kindness	winner	hero	famous	tribute	wager	3042
36	copy	record	request	mail	submit	fax	111
98	community	resident	mission	quality	excellent	life	78
52	county	leg	legislator	legislature	town	municipality	132
18	often	always	sometimes	never	easy	even	505
20	click	blog	email	copyright	dream	sorry	336
38	camp	yoga	library	camper	fun	librarian	1009
43	antibody	infection	hepatitis	tuberculosis	infect	viral	1551
66	drinking	water	reservoir	contaminant	irrigation	tap	228
68	spray	mosquito	pesticide	pest	repellent	soap	898
33	fire	alarm	firefighter	rescue	apparatus	emergency	271
56	holiday	weekend	parade	event	auburn	host	283
44	microchip	cat	euthanasia	spay	rabies	neuter	1229
70	election	ethic	ballot	political	candidate	lobbyist	382
60	shall	unless	except	mean	deem	forth	103
59	motion	unanimously	adjourn	prince	carry	ken	220
81	effluent	sludge	wastewater	mercury	lbs	gal	537
89	ask	explain	say	reply	suggest	ruff	427
5	home	family	homeowner	single	residence	cottage	94
105	proposer	breach	franchisee	indemnify	agree	hereunder	273
119	alderman	councilor	councilwoman	alderwoman	common	roll	260
14	borough	exam	trademark	veteran	immigrant	new	359
116	asthma	overdose	diabetes	obesity	hospitalization	prevalence	609
37	dental	medicare	deductible	coinsurance	prescription	copay	444
67	plat	thence	easement	pud	petitioner	annexation	271
35	parent	youth	child	mentor	literacy	foster	326
75	website	plain	please	online	customize	contact	98
21	bid	bidder	contractor	subcontractor	contract	procurement	238
95	duct	valve	splice	piping	conduit	conductor	850
83	storm	runoff	drainage	sewer	sanitary	infiltration	224
64	discrimination	gender	disability	race	religion	racial	437
32	think	really	thing	something	maybe	just	899
49	recycling	recycle	garbage	trash	waste	bin	405
3	maturity	portfolio	rating	jun	yield	investment	276
8	invoice	payment	card	cash	account	amt	222
12	password	header	archive	browser	folder	text	552
90	student	school	elementary	college	academic	graduate	303
48	application	applicant	must	certificate	license	proof	150
92	food	calorie	meat	vend	utensil	salad	1291
86	whereas	hereby	resolve	bond	anticipation	redemption	194
26	petroleum	spill	contamination	asbestos	contaminate	radioactive	444
4	para	persona	ante	horas	junta	sin	644
7	energy	renewable	solar	electricity	climate	efficiency	416
22	wireless	server	software	telecommunication	cable	technology	376
76	year	fiscal	five	annual	last	three	50
34	fee	charge	per	cost	plus	hourly	109
109	city	fort	manager	worth	hall	municipal	10
103	com	perm	tor	cigarette	loo	comm	1386
79	tow	plow	vehicle	trailer	motor	truck	594
54	dwell	building	remodel	unit	occupancy	alteration	132
23	name	address	description	number	list	zip	92
69	vista	ranch	suite	trinity	coliseum	mesa	657
61	cannabis	marijuana	cultivation	dispensary	collective	liquor	470
114	beach	orange	platinum	resort	ocean	angel	517
1	landlord	tenant	lease	lessee	golf	rent	288
118	flood	earthquake	floodplain	hurricane	tornado	disaster	513
99	roof	porch	awning	masonry	brick	vinyl	729
111	buffalo	player	league	ballpark	baseball	football	542
112	excavation	trench	excavate	gravel	silt	concrete	696
53	downtown	mall	hotel	midtown	uptown	shopping	531

Table 5: 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
53	downtown	mall	hotel	midtown	uptown	shopping	531
117	police	patrol	chief	lieutenant	captain	swear	283
2	page	yon	rev	sou	spec	gen	165
100	senate	house	butler	hook	rep	haven	590
55	chapter	code	section	subsection	article	amend	124
19	fugitive	noise	exhaust	receptor	coal	ozone	437
31	aviation	taxicab	airport	runway	airline	hangar	498
85	homeless	homelessness	supportive	client	transitional	encampment	232
42	tax	exemption	taxable	deduction	taxpayer	appraisal	160
80	artist	artwork	art	exhibition	gallery	artistic	1060
65	density	land	us	urban	village	growth	102
101	marsh	riparian	habitat	wetland	grassland	freshwater	1110
120	bend	rogers	walnut	grape	parenthood	shalom	315
108	owner	inspector	property	inspection	unsafe	nuisance	156
110	incumbent	ability	supervise	knowledge	supervision	essential	378
102	parking	space	height	garage	foot	lot	83
30	figure	census	population	respondent	comparison	table	240
71	economic	workforce	economy	industry	sector	job	312
91	prune	forestry	tree	planting	shrub	root	1092
28	conviction	guilty	offense	convict	misdemeanor	felony	762
115	mitigation	impact	adverse	significant	mitigate	measure	135
27	workshop	learn	tour	upcoming	get	view	119
16	park	recreation	playground	picnic	trail	zoo	253
15	landmark	historic	revival	preservation	archaeological	historical	936
10	ave	rainier	beacon	aurora	greenwood	capitol	353
51	waterfront	boat	shoreline	maritime	dock	port	788
82	avenue	east	west	north	street	south	78
73	actuarial	pension	retirement	retiree	unfunded	contribution	181
6	variance	setback	fence	exception	yard	nonconforming	122
46	allegation	complainant	misconduct	complaint	bias	allege	631
25	bankruptcy	plaintiff	examiner	creditor	trial	appeal	843
78	violent	gang	violence	inmate	crime	offender	710
97	employee	sick	wage	grievance	bargaining	overtime	243
77	board	appoint	chairperson	secretary	member	vice	137
24	grant	funding	program	fund	federal	match	49
104	project	improvement	upgrade	replacement	phase	appropriated	84
94	audit	auditing	deficiency	auditor	internal	weakness	195
13	yes	agency	successor	redevelopment	oversight	disposition	128
9	realm	design	proponent	courtyard	facade	concept	468
96	propose	draft	comment	alternative	plan	planning	68
106	sidewalk	crosswalk	signal	traffic	intersection	curb	269
41	bicycle	bike	transit	bus	route	mobility	279
63	memorandum	council	resolution	negotiation	manager	ward	132
39	commission	committee	commissioner	advisory	chair	discussion	126
84	implement	monitor	performance	inventory	process	track	146
107	budget	appropriation	fund	expenditure	adopt	levy	99
62	affordable	housing	affordability	household	income	renter	224
17	million	revenue	forecast	offset	deficit	projection	187
74	neighborhood	vision	attractive	node	amenity	corridor	351
45	zoning	district	zone	acre	dist	rezoning	71
113	debt	governmental	asset	net	statement	obligation	133
72	rouge	parish	baton	hogan	councilman	bowman	528
88	position	staffing	citywide	analyst	strategic	allocation	111
40	accessibility	mode	false	null	else	tab	105
11	strategy	goal	stakeholder	strategic	engagement	outreach	168
58	news	warren	announce	lovely	release	today	501
50	aye	absent	khan	nay	berry	voting	318
87	digest	proposal	reappoints	sander	gray	metropolitan	232
29	agenda	speaker	item	divided	speak	refrain	146
47	consolidated	reinvestment	contingency	contract	authorize	engineering	131
57	suspect	fatal	shoot	pronounce	stopper	gunshot	500

Table 6: 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.

different states, we show how our pipeline is capable of gathering corpora that shed light on the forms and functions of local government.

References

Adamic, Lada A. and Natalie Glance. 2005. "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog." *Proceedings of the 3rd international workshop on Link discovery - LinkKDD '05* pp. 36–43.

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

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.

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

Dovere, Edward-Isaac. 2018. "Government isn't working — so mayors say hire them.".
URL: *https://www.politico.com/story/2018/01/25/mayors-pitch-themselves-trump-remedy-366996*

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/10.101*
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*

Gentzkow, Matthew and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers." *Econometrica* 78(1):35–71.

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.

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.

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.

Hatemi, Peter K. and Rose McDermott, eds. 2011. *Man Is by Nature a Political Animal: Evolution, Biology, and Politics*. Chicago: University of Chicago Press.

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.

King, G., J. Pan and M. E. Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science* 345(6199):1251722–1251722.
URL: *http://www.sciencemag.org/cgi/doi/10.1126/science.1251722*

King, Gary, Jennifer Pan and Margaret E. Roberts. 2017. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument." *American*

Political Science Review 111(03):484–501.

URL: *https://www.cambridge.org/core/product/identifier/S0003055417000144/type/journal_article*

King, Gary, Jennifer Pan and Margaret Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107(02):326–343.

URL: *http://www.journals.cambridge.org/abstract_S0003055413000014*

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

Lin, Y-R, J P Bagrow and D Lazer. 2011. "More Voices than Ever? Quantifying Bias in Social and Mainstream Media." *arXiv preprint arXiv* 1111(1227).

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

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.

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

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.

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

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, 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.

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.

Schaffner, Brian F., Matthew Streb and Gerald C. Wright. 2001. "Teams without Uniforms: The Nonpartisan Ballot in State and Local Elections." *Political Research Quarterly* 54(1):7–30.

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

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

Figure 7: Five largest topic effects for the population covariate. The fact that the population and epidemiology topics are positively correlated with city size is indicative of the model's validity.

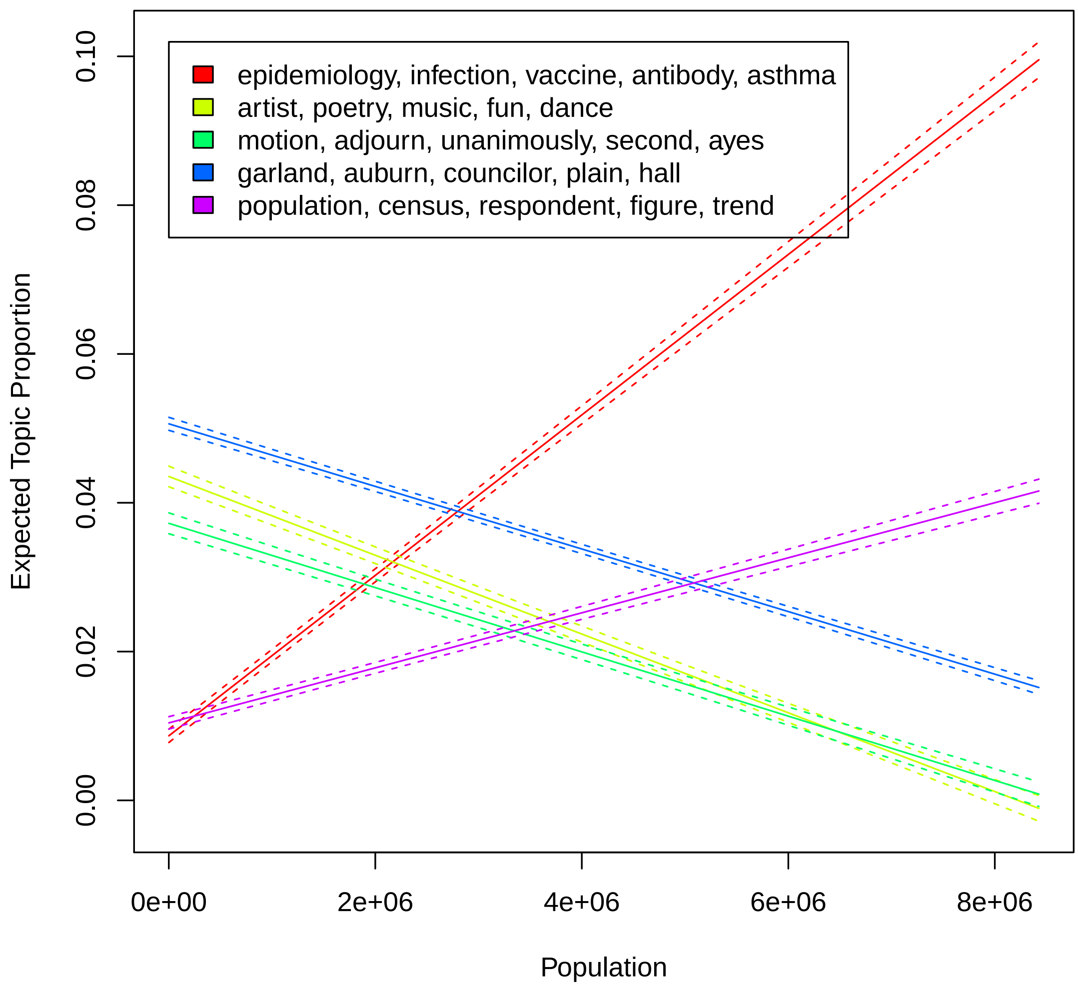


Figure 8: Five largest topic effects for the median income covariate. The fact that the crime topic is most prevalent in poorer cities, good governance is the most positively correlated with income is indicative of the model's validity.