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

Markus Neumann* Fridolin Linder† Bruce Desmarais‡

August 2, 2019

Abstract

A local government's website is an important source of information about policies and procedures for residents and community stakeholders. Existing research in political science and related fields has relied on manual methods of website content collection, limiting the scale and scope of website content analysis. Our objectives and corresponding contributions are two-fold. First, we develop a methodological pipeline to gather, process, and analyze website content, along with an R package implementation of the pipeline. Second, we illustrate the pipeline through the collection and analysis of 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 posted on the government's website.

1 Introduction

Government websites convey voluminous information about all aspects of government policymaking, policy implementation, and public deliberation. 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, focusing in particular on textual contents (e.g., Grimmelikhuijsen 2010; Wang et al. 2005; Osman et al. 2014; Eschenfelder et al. 1997). The conventional approach to data collection in projects focused on government websites involves manual content extraction from each website in the dataset. Though accurate, the manual approach to data collection is costly

*Department of Political Science, The Pennsylvania State University, University Park, PA 16802, USA. Email: mvn5218@psu.edu. Corresponding author.

†Department of Political Science, Social Media and Political Participation Lab, New York University, New York, NY 10012, USA. Email: fridolin.linder@nyu.edu

‡Department of Political Science, The Pennsylvania State University, University Park, PA 16802, USA. Email: bdesmarais@psu.edu

for large-scale analysis. 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—describing challenges in data collection and processing, as well as the solutions we adopt. We provide an illustrative application in which we explore the ways in which the textual contents on city government websites in six American states correlate with the partisanship of the city mayor.

Our research objectives and corresponding contributions are two-fold. First, we present a set of tools that can be effectively used in combination to automatically gather sections of substantive plain text from government websites—covering entire contents including the plain HTML files, and linked files in various formats (e.g., DOC, PDF, and TXT). Our pipeline provides researchers with a highly scalable approach to constructing comprehensive samples of textual data associated with government entities.[1] Second, we gather and analyze a dataset that covers the textual contents of websites from over two hundred municipal governments in the United States. By studying the covariation of topical contents on these websites with the partisanship of the city mayors, we validate the utility of both the pipeline, and this specific dataset. In summary, our contributions offer scholars of government a roadmap for collecting, and an example of, large scale textual data extracted from government websites.

2 Politics and Government Website Content

Though government websites serve largely instrumental service-delivery purposes, they also offer officials a prime venue via which to communicate policy goals and accomplishments, which inevitably reflect officials' politics. In the current paper, we focus on the running example of the reflection of mayoral partisanship on municipal government websites. A substantial body of

[1] It is useful to note a couple of features that we do not intend to provide with this pipeline. First, this is not intended to be a pipeline for scraping and analyzing any and all websites. Our focus is on government websites, as we assume that the contents are in the public domain, and not subject to limitations on the application of scraping technology. Second, though such extensions would be valuable, we do not address the collection and analysis of image, video, sound, and tabular data from government websites.

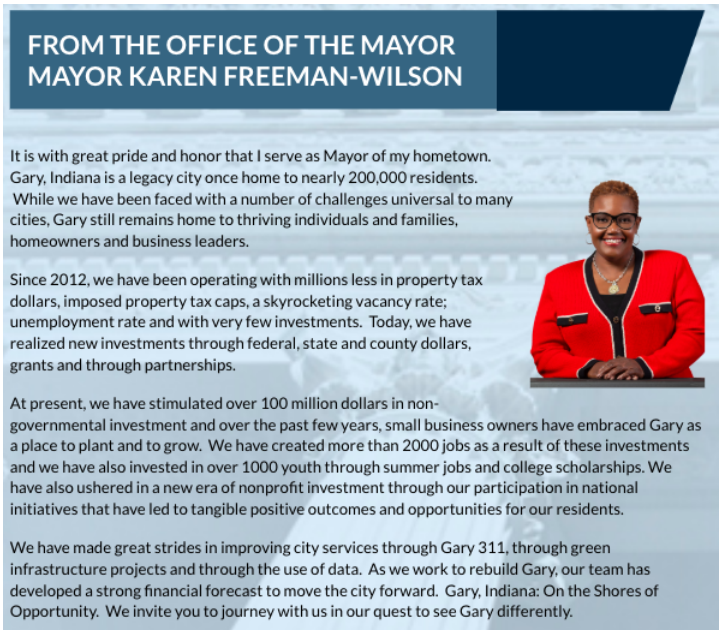


Figure 1: Screenshot from the homepage at https://garyin.us/, accessed on 05/22/2019. Image depicts Democratic mayor of Gary, IN, Karen Freeman-Wilson.

research has found that the partisanship of the mayor affects city governance along multiple dimensions of spending and policy attention (Gerber and Hopkins 2011; de Benedictis-Kessner and Warshaw 2016; Einstein and Glick 2016; Marion and Oliver 2013). Official city websites allow mayors to present their views and policy priorities to the public. In local politics, where campaign funds are low, this lends incumbents a crucial advantage in becoming more well-known among their constituencies (Stanyer 2008). Local government websites are frequently visited by the public (Thomas and Streib 2003). City websites can be used to communicate the stance of a mayor on social or economic programs. Consider the example of the Gary, Indiana homepage, depicted in Figure 1. This screenshot provides a clear example of the utility of a city website for communicating the mayor's policy priorities and accomplishments.

The existing research that uses scraped websites provides an indication of the theoretical value of empirical analysis of web contents. Research on 'e-governance' evaluates government websites

in terms of accessibility, ease-of-use, and function (e.g., Urban 2002; McNutt 2010; Armstrong 2011; Feeney and Brown 2017). As an example, Grimmelikhuijsen and Welch (2012) study local government websites of Dutch municipalities to measure government transparency regarding air quality in the municipalities. The websites of politicians and their parties have also been the object of research (Druckman et al. 2009, 2010; Cryer 2017; Esterling et al. 2011; Esterling and Neblo 2011; Norris 2003; Therriault 2010). For example, Druckman et al. (2010) analyze the issues engaged on websites for candidates in U.S. Congressional elections, and find that candidates strategically engage just a few issues based on the priorities in their districts and the characteristics of their opponents.

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 websites were scraped in March 2018. 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, so our sample is primarily focused on these two states. For Indiana and Louisiana, 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 appendix for more detail). California and Texas contain a number of large cities whose mayors are sufficiently well-known for their partisanship to be available. Our sample is well-balanced on a number of theoretically important dimensions. One, each of the four Census regions are represented with at least one state. Two, we have a fairly well-balanced sample with respect to the urban/rural cleavage. Furthermore, the sample is politically balanced—we have three blue states, and three red states. The partisan breakdown of city websites by state is provided in the appendix. This dataset of city web-

site contents represents a contribution in the growing area of cross-municipality datasets covering local governements (e.g., Marschall and Shah 2013; Sumner, Farris, and Holman Sumner et al.). Details on the sources and methods of raw data collection can be found in the online appendix.

4 The Web to Text Pipeline

In this section, we describe our methodological pipeline, with which we take an archive of website files, and output a corpus of formatted plain text documents. We address three 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 is aimed at extracting clean plain text from this heterogeneous file base. The second step in our pipeline is to process the text to remove language that is effective at differentiating one website from another but is uninformative regarding policy or political differences between governments. Finally, these tools need to work consistently across all of the websites in our corpus, in spite of the fact that relevant information is stored and structured in different ways. We make a software recommendation for each of these steps and gather most of them our R package, `gov2text`. All of the recommended software is either well-established in the natural language processing community, or part of the Unix ecosystem. As such, all of it is free, open source, well-developed and will continue to be supported by a dedicated community. Some of the steps we take in this processing pipeline are universally applicable in the analysis of textual data, and some of them are most appropriate for the particular type of text analysis that we apply to this data—statistical topic modeling. We will clarify this distinction as we describe steps in our pipeline.

4.1 Site to Text Conversion

4.1.1 File Type Detection

The format of a file has a major impact on whether and how textual data can be extracted from a document. 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. For example, we found thousands of documents that ended in .html, when they were actually PDFs. A more accurate test for file type relies on the use of magic numbers, a short sequence of bytes at the start (and sometimes end) of files that is unique for each file type and therefore allows its correct identification. We implement this method using the R package `wand` (Rudis et al. 2016).[2]

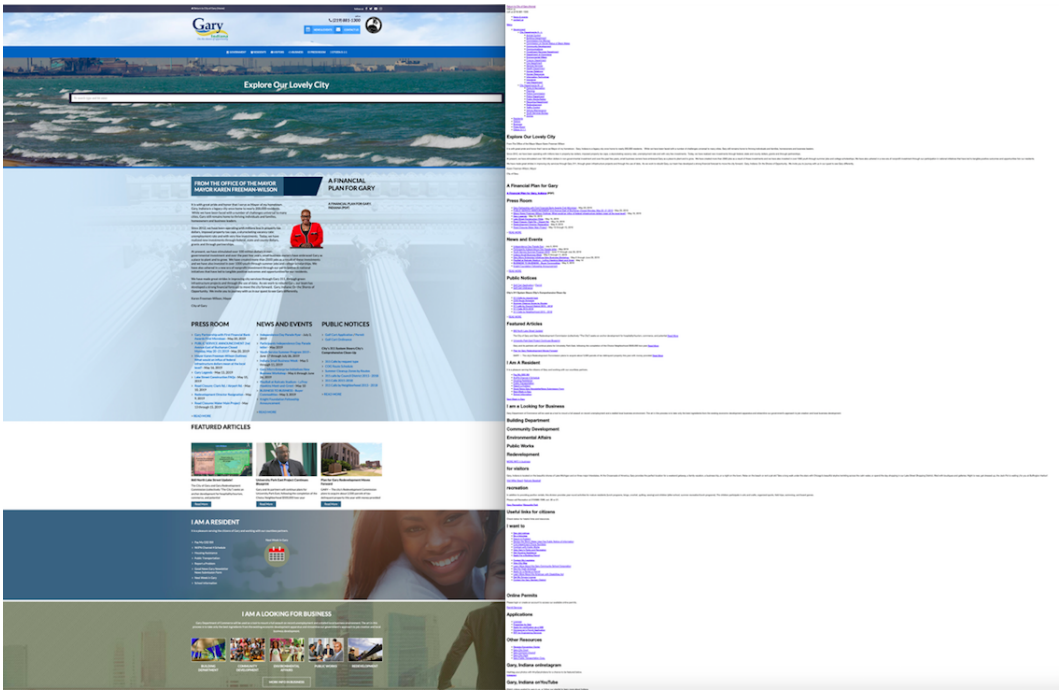
4.1.2 Extracting Text from HTML

The HTML files that websites are comprised of contain a large amount of useful information, but also completely irrelevant text such as menus, navigational elements and other boilerplate. The side-by-side screenshots presented in Figure 2 convey the challenges presented by extracting content for text analysis for websites. The textual content that is substantive and unique to the Gary, IN homepage is the Mayor's message depicted in Figure 1. The top row of Figure 2 presents the complete homepage, along with all of the text that can be naively extracted from the site. The Mayor's message represents a relatively small fraction of the total text on the page.

A subfield of the information retrieval literature, dealing with boilerplate extraction, can offer a solution to this problem. The goal of this branch of research is to develop algorithms with the ability to estimate whether a given portion of an HTML file is substantive. To this end, structural features, such as HTML tags (which are not sufficiently informative on their own), text statistics such as

[2] `Wand` is an R interface to the Unix library `libmagic` (Darwin 2008), which is included in all Linux distributions (which use this library to determine file types by default), Mac OS X, and has also been ported to Windows.

(a) Naive Parsing



(b) Boilerpipe

From The Office of the Mayor Mayor Karen Freeman-Wilson

It is with great pride and honor that I serve as Mayor of my hometown. Gary, Indiana is a legacy city once home to nearly 200,000 residents. While we have been faced with a number of challenges universal to many cities, Gary still remains home to thriving individuals and families, homeowners and business leaders.

Since 2012, we have been operating with millions less in property tax dollars, imposed property tax caps, a skyrocketing vacancy rate; unemployment rate and with very few investments. Today, we have realized new investments through federal, state and county dollars, grants and through partnerships.

At present, we have stimulated over 100 million dollars in non-governmental investment and over the past few years, small business owners have embraced Gary as a place to plant and to grow. We have created more than 2000 jobs as a result of these investments and we have also invested in over 1000 youth through summer jobs and college scholarships. We have also ushered in a new era of nonprofit investment through our participation in national initiatives that have led to tangible positive outcomes and opportunities for our residents.

We have made great strides in improving city services through Gary 311, through green infrastructure projects and through the use of data. As we work to rebuild Gary, our team has developed a strong financial forecast to move the city forward. Gary, Indiana: On the Shores of Opportunity. We invite you to journey with us in our quest to see Gary differently.

Karen Freeman-Wilson, Mayor

Figure 2: The top image provides a side-by-side depiction of the entire homepage of https://garyin.us/, accessed on 05/22/2019, and complete/naive extraction of all of the text on the site. Bottom image provides the result of running https://garyin.us/ through the `boilerpipe` algorithm at https://boilerpipe-web.appspot.com/.

word and sentence length, as well as other heuristics are used. We rely on the `boilerpipe` classifier described in Kohlschütter et al. (2010), which is implemented through the R package `boilerpipeR`. The `boilerpipe` algorithm has been widely used in the computer science and natural language processing literatures, but to our knowledge has not been previously used in political science. The complete text extracted from the Gary, IN homepage using boilerpipe is depicted in the screenshot in the bottom row of Figure 2. We see that only the Mayor's message is extracted, leaving the rest of the text as boilerplate.[3]

4.1.3 Extracting Text from PDF, DOC, DOCX and TXT

The extraction of information from other text-based file formats is more straightforward.[4] To this end, we rely on `readtext` R package (Benoit and Obeng 2019), which is a wrapper for a set of parsers.[5,6] The breakdown of all files by type is given in the online appendix. The most frequent file type besides HTML is PDF, from which we are able to extract a substantial amount of usable text. Files of type DOC, TXT, and DOCX, also occur regularly in our corpus and offer a considerable volume of textual data.

4.2 Preprocessing

Preprocessing is an important part of text-as-data research and choices made therein can have significant effects on the outcomes of an analysis (Denny and Spirling 2018). As such, our advice given in this section, more than in any other, is specific to the problem of extracting meaningful textual information from municipal government websites, with the end goal of its use in a bag-of-

[3] In the online appendix we present a replication of the topic modeling presented in the main text below in which we use a minimal HTML parser rather than `boilerpipe` to process the data. We show that without `boilerpipe`, some of the most partisan 'topics' are simply website boilerplate text.

[4] See Berg et al. (2012) for a discussion of why extracting text from PDFs is nevertheless nontrivial.

[5] `readtext` determines a document's type solely through its ending—so the conversion described above is necessary.

[6] `readtext` also contains an HTML parser, but it does not eliminate boilerplate like boilerpipe.

words-based model. The techniques we employ might also be of use in other types of applications, but by no means should this section be regarded as a general-purpose manual for preprocessing. The challenge in conducting preprocessing for a comparative analysis of websites lies in the considerable variance between websites. Some of it is substantively informative and some of it is completely irrelevant. As an example of the latter, names of city officials and citizen petitioners feature frequently in city documents. The same is true for streets, locations and not least of all, the city itself. Since individual names recur at a much higher rate within a city than across the entire corpus, this would cause a topic model to cluster its topics by city. Consequently we require a tool which detects the signal in the noise and does so consistently for a discordant set of sources.

To this end, we turn to a common method in natural language processing—part-of-speech (POS) tagging and named entity recognition (NER). In our case, names are the source of substantively uninteresting heterogeneity between cities, so NER is used to detect and remove them.[7] However, we caution here that for many other applications, where the names of political actors might be of interest, this step is not recommended. Furthermore, we select words on the basis of their POS-tags, retaining only nouns (the modal category), verbs, and adjectives.[8] Furthermore, we keep proper nouns that also occur as nouns—this removes names, but retains titles such as "Police Chief" which can appear as proper nouns if they are followed by a name. Finally, we also conduct lemmatization to reduce words to their basic form.[9] POS-tagging, NER and lemmatization are all implemented through `spacyr`. To deal with any leftover issues, we remove words with less than three characters (these are usually artifacts from improperly encoded documents and faulty or impartial optical character recognition), stopwords and non-English words (using the R package `Hunspell`). A final and crucial step is the removal of duplicate documents, which occur very

[7] We retain laws, nationalities or religious or political groups, as well as works of art (e.g., statutes).

[8] For applications outside of bag-of-words models, where the grammatical structure remains of interest, users might also want to retain other parts of speech.

[9] Lemmatization is similar to stemming, but works differently by taking grammar and surrounding words into account to identify the dictionary form of a word.

frequently on websites. In addition to their primary purpose, the previous preprocessing steps also help in stripping otherwise identical documents of information that makes them unique – such as names and dates – thus facilitating their deletion.

After preprocessing, our corpus consists of 356,911 documents. In Table 1 we summarize all of the steps we take in gathering and processing our data. The summary includes a brief description of the step, the software packages used, and an indicator of whether the method is implemented in our R package, `gov2text`.

Process	Software dependency	in `gov2text`
1. Assemble url list.	`Selenium`	no
2. Collect website files.	`wget`	no
3. Correct file extensions.	`wand` (Rudis et al. 2016)	yes
4. Discard website boilerplate.	`boilerpipeR` (Annau et al. 2015)	yes
5. Convert non-HTML files to text.	`readtext` (Benoit and Obeng 2019)	yes
6. Lemmatize text.	`spacyr` (Benoit and Matsuo 2018)	yes
7. Remove names.	`spacyr`	yes
8. Retain nouns, verbs, adjectives.	`spacyr`	yes
9. Stopword/number removal.	`quanteda` (Benoit et al. 2018)	yes
10. Retain only English words.	`Hunspell` (Ooms 2018)	yes
11. Removal of duplicate documents.	`gov2text`	yes

Table 1: Data collection and processing pipeline. Steps to collect and prepare text for topic modeling.

5 Partisan Language on Municipal Websites

City mayors use government websites to present their policy priorities to the public. Consider an example; (Formicola et al. 2003, p.55) document a significant website content change during a transition in mayoral administrations in the city of Indianapolis. Under the Republican mayor Stephen Goldsmith, voluminous content was added to the city website in connection with the Front Porch Alliance—a faith-based initiative to create partnerships with religious organizations for the use and administration of city resources. Faith-based initiatives represent a type of public-private

partnership that is popular with Republicans (Saperstein 2003). When Democrat Bart Peterson took office in 2000, the material related to the Front Porch Alliance was removed from the website. We consider whether the partisan manipulation of city website contents documented in this example holds in a large-scale and more recent sample of city websites. We illustrate the analysis of municipal website content by studying the ways in which differences in website content correlate with the partisanship of the city's mayor 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. These constraints do not apply to website contents. City governments have great discretion in composing their websites, modifying website content is low cost relative to other policy changes, and, as reviewed above, city websites provide an effective and often-used means of communication with city residents.

To study content differences between government websites based on mayoral partisanship, we draw upon a recently-developed class model for text, the structural topic model (STM), developed by Roberts et al. (2014). Building on the conception of "topics" in Latent Dirichlet Allocation, in the STM a topic is a multinomial distribution defined on the word types in the corpus dictionary. 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 STM provides a tool to estimate the relationship between the party of the city's mayor and the prevalence of each topic. We also include the municipality's population and median income as covariates. Further details on and results from our STM specification can be found in the online

appendix.

5.0.1 Structural topic model results

The results are shown in Table 2. First, it is notable that the 95% credible interval includes zero in only seven of the sixty topics, indicating that the topics discussed on city websites varies systematically with the partisanship of the mayor. Many of the topics associated with Democrats fit with what we understand to be national party priorities. Topic **52**, on affordable housing, clearly resonates with the Democratic party's appeal to low-income voters. Topic **6** ('race', 'islander', 'census, 'female') covers racial and gender identity issues. Similarly, employee rights and benefits are represented in topics **10** and **29**. Democrats also exhibit a strong preference for words related to public finances, such as Topic **58** ('budget', 'revenue', 'expenditure'), Topic **45** ('asset', 'actuarial', 'liability', 'financial'), Topic **35** ('bond', 'obligation', 'proceeds') as well as Topic **55** ('taxable', 'deed', 'value'). 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, and take credit for those efforts (e.g., the Gary, IN example in Figure 1). 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, **59** ('burglary', 'robbery', 'theft', 'homicide') is clearly focused on crime. 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 law enforcement, even after controlling for crime (Einstein and Kogan 2015).

City websites with Republican mayors, meanwhile, exhibit a pronounced focus on the essential functions of government. Basic utilities such as energy (Topic **20**), fire protection (Topic **51**), vaccination (Topic **2**), and sanitation (Topic **47**), are prevalent among cities with Republican mayors.

These basic service topics cannot be found among topics prevalent in cities with Democratic mayors. Similarly, zoning issues figure prominently in the set of republican topic (Topic **19**), which fits with the findings of Sorens (2018) that Republicans are more supportive of restrictive residential zoning rules. The Democratic topics also include one that is somewhat focused on zoning, Topic **39** ('downtown', 'mixed', 'density'), but emphasizes mixed-use zoning—a loosening of conventional single-use zoning rules.

6 Conclusion

We have developed a methodological pipeline for automatically gathering and preparing government websites for comparative content analysis. We have produced an R package `gov2text`, in which we have implemented and wrapped the core components of our pipeline. 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. 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 and past research on the effects of mayoral partisanship on city governments.

The biggest limitation in our pipeline, and an open area for future research, is the reliance on `wget` to gather the initial website files. By using `wget`, we miss content that is displayed dynamically on websites using JavaScript. For any one website, it would be possible to customize a routine with `Selenium` to access dynamic elements, but the process would need to be customized for each website.[10]

[10] We investigated whether the presence of JavaScript was related to the amount of text we gathered from the website. We calculated the correlation between the number of <script> HTML tags on a city's website, which indicate the use of JavaScript on a site, and the number of text tokens we scrape from the site. This correlation is -0.059, which indicates a

Funding

This work was supported by the National Science Foundation [1320219, 1637089, 1641047].

References

Annau, M., C. Kohlschuetter, and A. Clark (2015). *boilerpipeR: Interface to the Boilerpipe Java Library*. R package version 1.3.

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

Benoit, K. and A. Matsuo (2018). *spacyr: Wrapper to the 'spaCy' 'NLP' Library*. R package version 1.0.

Benoit, K. and A. Obeng (2019). *readtext: Import and Handling for Plain and Formatted Text Files*. R package version 0.74.

Benoit, K., K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo (2018). quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software 3*(30), 774.

Berg, Ø. R., S. Oepen, and J. Read (2012). Towards High-Quality Text Stream Extraction from PDF. In *ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pp. 98–103.

Brown, A. (2017). Republicans more likely than Democrats to have confidence in police.

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

Darwin, I. (2008). Libmagic.

very weak relationship between the use of JavaScript and the amount of text scraped from the site.

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

Denny, M. J. and A. 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.

Druckman, J. N., C. L. Hennessy, M. J. Kifer, and M. Parkin (2010). Issue Engagement on Congressional Candidate Web Sites, 2002—2006. *Social Science Computer Review 28*(1), 3–23.

Druckman, J. N., M. Kifer, and M. Parkin (2009). Campaign Communications in U.S. Congressional Elections. *American Political Science Review 103*(03), 343–366.

Einstein, K. L. and D. M. Glick (2016). Mayors, partisanship, and redistribution: Evidence directly from us mayors. *Urban Affairs Review*, 1078087416674829.

Einstein, K. L. and V. Kogan (2015). Pushing the City Limits: Policy Responsiveness in Municipal Government. *Urban Affairs Review*, 1–30.

Eschenfelder, K. R., J. C. Beachboard, C. R. McClure, and S. K. Wyman (1997). Assessing U.S. federal government websites. *Government Information Quarterly 14*(2), 173–189.

Esterling, K. M., D. M. Lazer, and M. A. Neblo (2011). Representative communication: Web site interactivity and distributional path dependence in the us congress. *Political Communication 28*(4), 409–439.

Esterling, K. M. and M. A. Neblo (2011). Explaining the Diffusion of Representation Practices among Congressional Websites. *Working Paper*, 1–42.

Feeney, M. K. and A. Brown (2017). Are small cities online? Content, ranking, and variation of U.S. municipal websites. *Government Information Quarterly 34*(1), 62–74.

Formicola, J., M. Segers, and P. Weber (2003). *Faith-based Initiatives and the Bush Administration: The Good, the Bad, and the Ugly*. Rowman & Littlefield.

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

Grimmelikhuijsen, S. G. (2010). Transparency of public decision-making: Towards trust in local government? *Policy & Internet 2*(1), 5–35.

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

Kohlschütter, C., P. Fankhauser, and W. Nejdl (2010). Boilerplate Detection using Shallow Text Features. In *Web Search and Data Mining*.

Marion, N. E. and W. 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, M. and P. Shah (2013). Local elections in america project. *Center for Local Elections in American Politics. Kinder Institute for Urban Research, Rice University.(Database)*.

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

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

Ooms, J. (2018). *hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker*. R package version 3.0.

Osman, I. H., A. L. Anouze, Z. Irani, B. Al-Ayoubi, H. Lee, A. Balcı, T. D. Medeni, and V. Weerakkody (2014). Cobra framework to evaluate e-government services: A citizen-centric perspective. *Government Information Quarterly 31*(2), 243–256.

Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science 58*(4), 1064–1082.

Rudis, B., C. Zoulas, M. Rullgard, and J. Ong (2016). *wand: Retrieve 'Magic' Attributes from Files and Directories*. R package version 0.2.0.

Saperstein, D. (2003). Public accountability and faith-based organizations: A problem best avoided. *Harvard Law Review 116*(5), 1353–1396.

Sorens, J. (2018). The effects of housing supply restrictions on partisan geography. *Political Geography 66*, 44–56.

Stanyer, J. (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.

Sumner, J. L., E. M. Farris, and M. R. Holman. Crowdsourcing reliable local data. *Political Analysis*.

Therriault, A. (2010). Taking Campaign Strategy Online: Using Candidate Websites to Advance the Study of Issue Emphases. pp. 1–23.

Thomas, J. C. and G. 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.

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

Wang, L., S. Bretschneider, and J. 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*, pp. 129b–129b. Ieee.

#	Top Word 1	Top Word 2	Top Word 3	Top Word 4	Top Word 5	Top Word 6	Tokens assigned
49	artist	fun	music	beginner	player	prize	4565
46	chair	subcommittee	speaker	agenda	committee	commission	446
16	motion	second	adjourn	carry	unanimous	chairman	419
47	effluent	inf	eff	infiltration	discharge	sludge	751
21	everybody	think	something	thing	try	want	2609
2	influenza	infection	vaccine	patient	tuberculosis	hepatitis	2980
27	article	subsection	shall	franchisee	paragraph	meaning	658
30	subcontractor	bid	bidder	proposer	subcontract	bidding	512
12	craftsman	architecture	brick	distinctive	revival	storefront	1731
24	mail	fax	application	click	applicant	copy	367
34	playground	recreation	picnic	park	restroom	zoo	546
19	setback	variance	zoning	height	yard	accessory	453
26	mesa	canyon	via	odd	unidentified	paradise	1886
23	bag	recyclable	recyclables	reusable	vegetable	bait	2254
20	customer	renewable	efficiency	energy	saving	conservation	652
31	student	teacher	preschool	academic	kindergarten	youth	855
28	garland	assoc	association	firefighter	duke	xerox	480
50	trench	manhole	ductile	excavation	pipe	grout	1436
32	canceled	dwelling	suite	ave	tad	alteration	491
51	vent	combustible	flammable	egress	ceiling	extinguisher	1160
44	findings	tank	string	carcinogen	lust	sic	255
17	portfolio	micron	maturity	treasury	yield	investment	538
48	contributor	filer	officeholder	political	rouge	payee	293
5	draft	comment	review	revision	clarify	process	356
37	endorsed	endorse	rescue	assistant	analyst	technician	355
9	trust	revocable	planned	mfr	apportionment	exhibit	361
8	imp	assessor	taxpayer	petition	preliminary	determination	91
40	amt	invoice	acct	exp	unencumbered	encumbrance	116
57	councilman	introduced	alderman	whereas	resolved	councilwoman	615
11	obesity	sugary	epidemic	drink	calorie	sensible	96
15	credit	docket	app	post	download	month	61
3	wetland	specie	species	vernal	ecological	riparian	2293
29	margin	error	disability	speak	employed	language	180
43	medicare	payroll	blanket	contractual	undistributed	dept	322
42	incumbent	prep	batch	qualifier	analytical	examination	1091
55	taxable	deed	res	homestead	value	book	87
22	allocation	subtotal	admin	cost	yon	allocate	190
25	mitigation	impact	significant	adverse	environmental	measure	217
56	savings	neighborhood	village	excise	ltd	matrix	131
33	thence	east	south	corner	west	avenue	340
7	fugitive	bio	emission	coal	unmitigated	exhaust	773
18	perm	queue	delay	peak	adj	flt	187
54	license	licensee	citation	tow	fee	taxicab	710
6	race	householder	islander	census	occupied	female	160
60	bicycle	bike	pedestrian	route	sidewalk	bicyclist	561
14	accomplishment	grantee	narrative	outcome	grant	recipient	255
53	applied	col	dist	occupancy	monoxide	valuation	128
4	audit	auditor	procedure	timely	implemented	oversight	472
35	redemption	bond	increment	obligation	proceeds	lease	339
39	downtown	mixed	retail	waterfront	orient	density	419
10	grievance	deductible	coinsurance	dependent	employee	copay	583
38	para	persona	horas	bud	contracted	ante	1334
36	respondent	compare	figure	trend	appendix	satisfied	696
45	governmental	asset	actuarial	liability	financial	statement	235
41	complainant	allegation	defendant	offender	commander	complaint	1695
52	homeless	homelessness	affordable	supportive	housing	affordability	394
58	budget	revenue	adopted	balance	transfer	expenditure	176
13	initiative	outreach	strategy	leadership	engagement	focus	502
1	absent	preside	authorize	ordained	int	tag	377
59	burglary	robbery	theft	homicide	murder	gunshot	945

Table 2: 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.