

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

Markus Neumann\*

Fridolin Linder<sup>†</sup>

Bruce Desmarais<sup>‡</sup>

May 23, 2019

## Abstract

A local government's website is an important source of information about policies and procedures for residents, community stakeholders and scholars. Existing research in public administration, public policy, and political science has relied on manual methods of website content collection and processing, limiting the scale and scope of website content analysis. We develop a methodological pipeline that researchers can follow in order to gather, process, and analyze website content. Our approach, which represents a considerable improvement in scalability, involves downloading the entire contents of a website, extracting the text and discarding redundant information. We provide an R package that can be used to apply our proposed pipeline. 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, we find that cities with Democratic mayors provide more information on policy deliberation and crime control, whereas Republicans prioritize basic utilities and services such as water, electricity and fire safety.

PA Letter Requirements:

2-4 pages

no longer than 1500-3000 words

1-3 small display items (figures, tables, or equations)

200-300 word abstract

---

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

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

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

# **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, Balci, Medeni and Weerakkody 2014; Eschenfelder, Beachboard, McClure and Wyman 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 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.

# **2 Mayoral Politics and City Government Website Content**

A substantial body of 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



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.

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, 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). 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 is depicted in Table 1. Details on the sources and methods of raw data collection can be found in the Appendix.

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

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

One of the more subtle aspects of local government is the presence of different types of government structures. Between council-manager governments and mayor-council governments (Morgan and Watson 1992)—either in the weak or strong mayor variant (DeSantis and Renner 2002)—there is variance in where a city’s executive authority lies. We do not have access to information about the type of governments across the breadth of our dataset. Given the prominent place that mayors tend to have on their cities’ websites, we feel that any bias arising from this nuance should be minor. Gerber and Hopkins (2011), whose theory is somewhat comparable to ours, find that the inclusion of this potential confounder does not affect the results.

## **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 that are suitable for comparative analysis with text as data methods. Here, 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.

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

### **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. In theory, HTML tags intended to mark text content exist to prevent this problem. In practice, not every web developer adheres to this standard, and even when they do, there is a large degree of variation in the way it is implemented. Parsing any one website is fairly straightforward, but parsing the 283 different websites in our sample in a consistent and comparable manner makes a rule-based approach impossible. The side-by-side screenshots presented in Figure 2 convey the

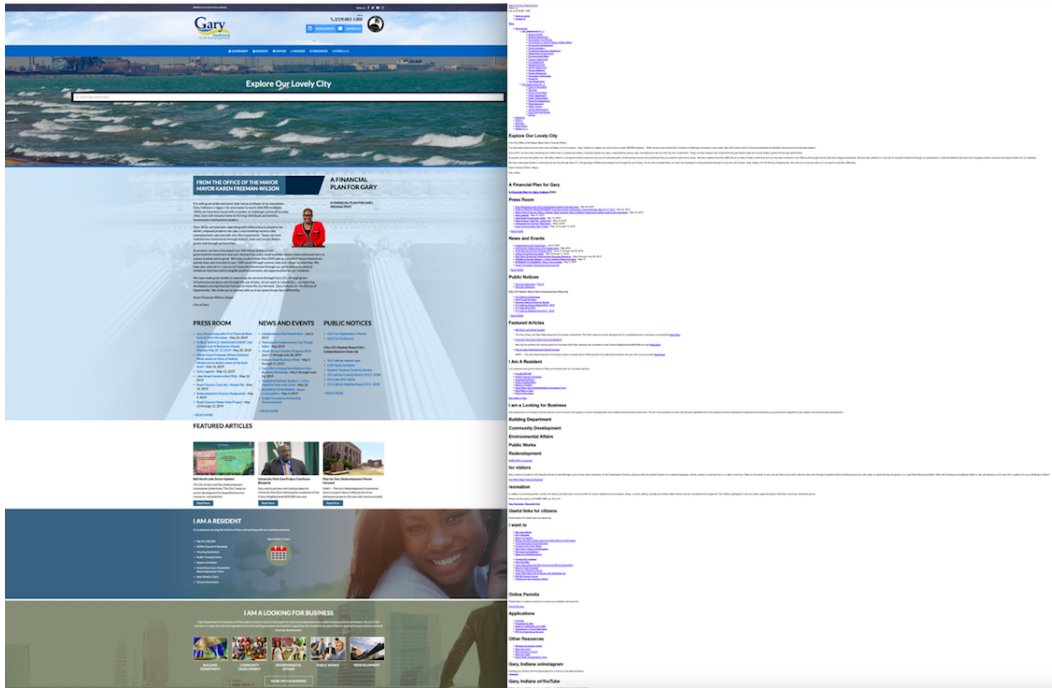


Figure 2: Side-by-side depiction of the entire homepage of <https://garyin.us/>, accessed on 05/22/2019, and complete/naive extraction of all of the text on the site.

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 ??.

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.

We leverage methods developed in the information retrieval literature to deal with this problem. These boilerplate detection tools are classifiers which rely on both structural features, such as HTML tags, as well as text statistics such as word and sentence length to estimate whether a given portion of an HTML file is useful. We rely on the ‘boilerpipe’ method described in Kohlschütter, Fankhauser and Nejdl (2010), which has been published in a reputable conference, is widely cited, and can easily be implemented through the R package `boilerpipeR`. The complete text

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 3: Result of running <https://garyin.us/>, accessed on 05/22/2019, through the default boilerpipe algorithm at <https://boilerpipe-web.appspot.com/>.

extracted from the Gary, IN homepage using boilerpipe is depicted in the screenshot in Figure ??.

We see that only the Mayor’s message is extracted, leaving the rest of the text as boilerplate.

#### 4.1.3 Extracting Text from PDF, XML, DOC, DOCX and TXT

Other files are read in through the `readtext` R package (Benoit and Obeng 2018), which is a wrapper for a set of parsers.<sup>1,2,3</sup> The breakdown of all files by type is given in Table 2. 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 XML, 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). The challenge in conducting preprocessing for a comparative analysis of websites lies in the considerable variance

<sup>1</sup>`readtext` determines a document’s type solely through its ending – so the conversion described above is strictly necessary for the package to work correctly.

<sup>2</sup>We have also experimented with several Unix-based alternatives, but found that they largely led to the same results as `readtext`.

<sup>3</sup>`readtext` also contains an HTML parser, but this tool simply extracts all text and is therefore inferior to boilerpipe.



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

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

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). While not their original purpose, these procedures perform very well in separating the wheat from the chaff. As names convey no substantive information, NER is used to detect and remove them.<sup>4</sup> Furthermore, we select words on the basis of their POS-tags, retaining only nouns (by far the most informative category), verbs, and adjectives. 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.

---

<sup>4</sup>We retain laws, nationalities or religious or political groups (which are politically salient with respect to immigration and identity) as well as works of art (which frequently feature statutes, plans, etc.)

Finally, we also conduct lemmatization to reduce words to their basic form.<sup>5</sup> POS-tagging, NER and lemmatization are all implemented through `spacyR`. This parsing-based approach is very effective in distilling a comparable corpus from a varied set of sources, as the grammatical rules of the English language remain their common denominator. 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`)<sup>6</sup>.

A final and crucial step is the removal of duplicate documents, which occur very 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. The removal of duplicate information also solves a problem posed by a recurring feature of municipal websites – city calendars. For most days in most cities, these calendars tend to be empty, but nevertheless contain a template, which would lead to the creation of a document in our corpus. These short and repetitive – but numerous – texts would pose severe problems to a topic model. Our approach preserves the valuable information contained within these calendars when cities do make use of them, but removes them when they offer no substantive data.

After all the preprocessing is set and done, our corpus consists of 259,099 documents. In Table 3 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`.

---

<sup>5</sup>Lemmatization is similar to stemming, but works in a somewhat more sophisticated manner by taking grammar and surrounding words into account to identify the dictionary form of a word. For example, the lemma of the word “lemmatization” would be “lemmatize”, whereas most stemmers would simply chop off the ending, which would yield “lemmatiz”. Thus, lemmatization makes the results more easily comprehensible.

<sup>6</sup>As we are using `spacyR` with an English parser, non-English words tend to be erroneously categorized as (proper) nouns.

| Process                                  | Software dependency | gov2text inclusion |
|--|---------------------|--------------------|
| 1. Assemble url list.                    | N/A                 | no                 |
| 2. Collect website files.                | wget & Selenium     | no                 |
| 3. Lemmatize text.                       | spacyR              | yes                |
| 4. Remove names.                         | spacyR              | yes                |
| 5. Retain nouns, verbs, adjectives.      | spacyR              | yes                |
| 6. Retain tokens that are English words. | Hunspell            | yes                |

Table 3: Data collection and processing pipeline. Steps to collect and prepare text for topic modeling.

## 5 Partisan Language on Municipal Websites

We illustrate the analysis of municipal website content by studying differences in website content based on the party of the mayor. As we reviewed above, the partisanship of the mayor has been found in past research to affect several features of city governance. However, Gerber and Hopkins (2011) note that, due to the constraints of state and national policies, municipalities lack discretion in many domains of governance. 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.

In order to analyze 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, Stewart, Tingley, Lucas, Leder-Luis, Gadarian, Albertson and Rand (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 coeffi-

cient 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 with a tool with which to estimate the relationship between the party of the city’s mayor and the prevalence of each topic we estimate. Further details on our STM specification can be found in the appendix.

### 5.0.1 Structural topic model results

The results are shown in Table 5. 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. Similarly, employee rights are represented in topics **10** and **29** [DEBATABLE]. 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. 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’) depicts Democrats’ complicated relationship with law enforcement [**This used to be about police as well, but now it’s more of a crime topic**]. On the one hand, Democratic partisans have a more negative perception of the police, rating it considerably more negatively on the appropriate use of force and the equal treatment of minorities (Brown 2017). On the other hand, the literature has also shown that cities with a higher Democratic vote share spend more on the police, even after controlling for crime (Einstein and Kogan 2015). Finally, Democrats also focus more on the deliberative process of policymaking, as topics 31 (‘agenda’, ‘committee’) [**This is now the second-most REPUBLICAN topic 46**], 34 (‘comment’, ‘draft’, ‘feedback’) [**This is now REPUBLICAN topic 5**], 1 (‘absent’, ‘preside’,

‘authorized’), and **4** (‘audit’, ‘procedure’, ‘oversight’) attest to. **[Republicans also have topic 16]** This openness regarding the policy process on behalf of cities with Democratic mayors fits with the findings of Grimmelikhuijsen and Welch (2012), which are that left-wing local governments exhibit greater transparency via website content.

City websites with Republican mayors, meanwhile, exhibit a pronounced focus on the essential functions of government. Basic utilities such as energy (Topic **20**), fire protection (Topic **51 [This topic has changed a lot.]**), drinking water (53) [topic missing], and garbage removal (Topic 49) [topic missing] are included among those topics that are more prevalent in cities with Democratic mayors. Similarly, protecting citizens from natural disasters is a focus in topics 1 (‘storm’, ‘runoff’, ‘drainage’) [topic missing] and **2 (‘influenza’, ‘infection’, ‘vaccine’ [This is not about Zika anymore, but infectious diseases in general.]**), which may reflect the greater prevalence of Republican mayors in the southeast, a region which is more often affected by hurricanes and tropical diseases.

## 6 Conclusion

We have developed a methodological pipeline for automatically gathering and preparing government websites for comparative content analysis. 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.

## **Funding**

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

## References

- Armstrong, Cory L. 2011. "Providing a clearer view: An examination of transparency on local government websites." *Government Information Quarterly* 28(1):11–16.
- Benoit, Kenneth and Adam Obeng. 2018. *readtext: Import and Handling for Plain and Formatted Text Files*. R package version 0.71.  
**URL:** <https://CRAN.R-project.org/package=readtext>
- Brown, Anna. 2017. "Republicans more likely than Democrats to have confidence in police.".  
**URL:** <http://www.pewresearch.org/fact-tank/2017/01/13/republicans-more-likely-than-democrats-to-have-confidence-in-police/>
- Cryer, J. E. 2017. "Candidate Identity and Strategic Communication." pp. 1–42.
- Darwin, IF. 2008. "Libmagic."
- de Benedictis-Kessner, Justin and Christopher Warshaw. 2016. "Mayoral partisanship and municipal fiscal policy." *The Journal of Politics* 78(4):1124–1138.
- Denny, Matthew J and Arthur Spirling. 2018. "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it." *Political Analysis* 26(2):168–189.
- DeSantis, Victor S and Tari Renner. 2002. "City government structures: An attempt at clarification." *State and Local Government Review* 34(2):95–104.
- Druckman, James N., Cari Lynn Hennessy, Martin J. Kifer and Michael Parkin. 2010. "Issue Engagement on Congressional Candidate Web Sites, 2002—2006." *Social Science Computer Review* 28(1):3–23.  
**URL:** <http://journals.sagepub.com/doi/10.1177/0894439309335485>

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

**URL:** [http://www.journals.cambridge.org/abstract\\_S0003055409990037](http://www.journals.cambridge.org/abstract_S0003055409990037)

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

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

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

**URL:** [http://www.sciencedirect.com/science/article/pii/S0740624X97900186%5Cnpapers2://publication/doi/10.1016/0740-624X\(97\)90018-6](http://www.sciencedirect.com/science/article/pii/S0740624X97900186%5Cnpapers2://publication/doi/10.1016/0740-624X(97)90018-6)

Esterling, Kevin M, David MJ Lazer and Michael A Neblo. 2011. "Representative communication: Web site interactivity and distributional path dependence in the US Congress." *Political Communication* 28(4):409–439.

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

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

**URL:** <http://dx.doi.org/10.1016/j.giq.2016.10.005>

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



- Grimmelikhuijsen, Stephan G. 2010. "Transparency of Public Decision-Making: Towards Trust in Local Government?" *Policy & Internet* 2(1):5–35.
- Grimmelikhuijsen, Stephan G and Eric W Welch. 2012. "Developing and testing a theoretical framework for computer-mediated transparency of local governments." *Public administration review* 72(4):562–571.
- Guillamón, Ma Dolores, Francisco Bastida and Bernardino Benito. 2013. "The electoral budget cycle on municipal police expenditure." *European Journal of Law and Economics* 36(3):447–469.
- Kohlschütter, Christian, Peter Fankhauser and Wolfgang Nejdl. 2010. Boilerplate Detection using Shallow Text Features. In *Web Search and Data Mining*.
- Marion, Nancy E and Willard M Oliver. 2013. "When the Mayor Speaks... Mayoral Crime Control Rhetoric in the Top US Cities: Symbolic or Tangible?" *Criminal justice policy review* 24(4):473–491.
- Marschall, Melissa and Paru Shah. 2013. "Local Elections in America Project." *Center for Local Elections in American Politics. Kinder Institute for Urban Research, Rice University.(Database)* .  
**URL:** <http://www.leap-elections.org/>
- McNutt, Kathleen. 2010. "Virtual policy networks: Where all roads lead to Rome." *Canadian Journal of Political Science/Revue canadienne de science politique* 43(4):915–935.
- Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2008. "Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16(4 SPEC. ISS.):372–403.

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

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

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

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

**URL:** <http://www.structuraltopicmodel.com>

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand. 2014. "Structural topic models for open-ended survey responses." *American Journal of Political Science* 58(4):1064–1082.

Rudis, Bob, Christos Zoulas, Mans Rullgard and Jonathan Ong. 2016. *wand: Retrieve 'Magic' Attributes from Files and Directories*. R package version 0.2.0.

**URL:** <https://CRAN.R-project.org/package=wand>

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

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

**URL:** <http://poseidon01.ssrn.com/delivery.php?ID=5881250961130801011070071091041011210350310770540170>

- 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.
- 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.

## Appendix

### Raw data collection methods and sources

We acquired the website URLs from two sources: One, we scraped the URLs of city websites from their respective Wikipedia pages, which we found from lists of cities contained within each state. Two, the General Services Administration (GSA) maintains all ‘.gov’ addresses, and provides a complete list of all such domains to the public.<sup>7</sup> The data from the GSA contains the following variables: (1) domain name, specifically, the all-uppercase version of domain and top-level domain (for example, ‘ABERDEENMD.GOV’); (2) the type of government entity to which the domain is registered, such as city, county, federal agency, etc; (3) for federal agencies, the name is specified; (4) the city in which the domain is registered. Naturally, the GSA’s list does not contain cities 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 latter and only supplement them with the GSA data if a specific city doesn’t have a URL recorded on Wikipedia, or our tests (see below) find it to be non-functional.

Not all of the URLs contained in these archives are functional. To test the URLs’ functionality, we use a web driver-controlled browser - a browser that is automatically controlled by a program rather than a human user. We use the Python bindings for the program `Selenium`, which we use to control `Firefox` through the web driver `Geckodriver`. This is advantageous compared to conventional scraping tools such as `Beautiful Soup` or `Rvest` because most websites are designed to be explored by browsers. Modern browsers perform a lot of actions behind the scenes, such as URL resolution and redirection. The use of a web driver-controlled browser is necessary

---

<sup>7</sup>The dataset is made available at <https://github.com/GSA/data/tree/gh-pages/dotgov-domains>. This list is updated once per month—we rely on the version released on January 16, 2017.

in our case because a) some city websites simply don't work, but they don't always output an error code correctly (this can fail, for example, if a webmaster simply stops maintaining a site without removing it entirely) which would throw off an automatic scraper, and more often, b) cities sometimes change their websites' URLs, in which case they redirect from the old to the new URL. A web driver-controlled browser, unlike the more rigid conventional scraping tools, will simply follow this redirection. This allows us to subsequently record and use the new URL for the actual website scraping. Consequently, an automated browser allows us to robustly answer the following questions: Is the website actually there? Does it work? If not, is it somewhere else or is it broken? We record this information and construct a list of verified URLs.

To download the websites, we rely on the Unix command line tool `wget`. This program is used to download files from the Internet, and with the use of a recursive option, acts like a web crawler and scraper. This means that `wget` downloads HTML files, parses them and then follows the links contained therein. Then it follows those links and repeats the process until it has constructed a complete tree of the website (note that the program is instructed to stay on the same domain, i.e. it does not follow external links). This way, all the files that make up a website are downloaded. For some cities, whose websites make heavy use of JavaScript to serve content dynamically, such content is not reachable with our methodology and would require additional steps to obtain. For this paper, we ignore such sites and restricted our corpus to cities with at least three successfully downloaded pages.<sup>8</sup>

The partisanship of the mayor of each city is coded in different ways, depending on the state. For Indiana, where elections are nominally partisan, this information is accessible through the state government's website<sup>9</sup>. For Louisiana, we received data on the outcomes of mayoral elections

---

<sup>8</sup>There is a possibility that this leads to a small bias in selecting against cities with the resources to build more elaborate websites. However, given that our sample is generally more on the wealthy side, this, if anything, should lead to a more balanced sample.

<sup>9</sup><http://www.in.gov/apps/sos/election/general/general2015?page=office&countyID=1&officeID=32&districtID=-1&candidate=>

from the Local Elections in America Project (LEAP) (Marschall and Shah 2013). For the other states, where mayoral elections are not nominally partisan (but the partisanship of the mayor is still well-known), we employed different means: For New York and Washington, we searched the state campaign finance websites, and coded the parties of the candidates based on the party committees from which they received donations. For California and Texas, where our data consists of highly populated cities, partisanship information was acquired from Ballotpedia<sup>10</sup>. Finally, we also scraped mayoral partisanship from the cities' Wikipedia pages. When compared to the other data sources above, (and manual searches in case of conflicts) Wikipedia proved to be very reliable and added additional cases to our dataset even for Indiana and Louisiana. Generally speaking, we found data scraped from Wikipedia, aided by manual corrections in case of missing or conflicting data, to be more reliable than data from governmental sources.<sup>11</sup>

Information on other covariates (population and median household income - from the American Community Survey 5-Year Data (2015)) was acquired through the API of the U.S. Census Bureau<sup>12</sup>.

## Details on STM specification

The structural topic model is implemented in the R package *STM* (Roberts, Stewart and Tingley 2018). We use 60 topics—the number recommended by the authors<sup>13</sup> for medium- to large-sized corpora.<sup>14</sup> We use four covariates: First, *party*, to estimate the difference in topic prevalence based on whether mayors are Republican or Democratic. Second, *city population*, which the literature

<sup>10</sup>[https://ballotpedia.org/List\\_of\\_current\\_mayors\\_of\\_the\\_top\\_100\\_cities\\_in\\_the\\_United\\_States](https://ballotpedia.org/List_of_current_mayors_of_the_top_100_cities_in_the_United_States)

<sup>11</sup>In Indiana, the data includes only cities - incorporated municipalities with at least 2,000 inhabitants - as opposed to towns.

<sup>12</sup><https://www.census.gov/data/developers/data-sets.html>

<sup>13</sup>For this recommendation, see the documentation for the function `stm()` in version 1.3.0 of the R package *stm* (Roberts, Stewart and Tingley 2018).

<sup>14</sup>Since our corpus is at the larger end of that spectrum, we also estimated a model with 120 topics, but found no notable differences.

frequently emphasizes as a determinant of the issues a city faces (see, for example, Guillamón, Bastida and Benito (2013)). Third, we control for wealth by relying on *median income* as a covariate, which we use as a proxy for the tax base in a city. Fourth and finally, we include state dummy variables, which should account for language that is associated with state-specific issues, and general background variables that vary across states.<sup>15</sup>

---

<sup>15</sup>The “Fightin’ Words” methodology developed by Monroe, Colaresi and Quinn (2008) could also be used to analyze word-frequency differences between cities based on mayors’ partisanship, but we elected to use the structural topic model since, unlike “Fightin’ Words”, the structural topic model enables us to adjust for several other features through multiple regression.

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

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.



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

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

| #   | Top Word 1    | Top Word 2    | Top Word 3    | Top Word 4    | Top Word 5   | Top Word 6    | Tokens assigned |
|-----|---------------|---------------|---------------|---------------|--------------|---------------|-----------------|
| 96  | subcommittee  | agenda        | forum         | speaker       | item         | adjournment   | 217 ■           |
| 49  | prize         | celebration   | ceremony      | parade        | follower     | favorite      | 2043 ■■■■■      |
| 102 | motion        | second        | adjourn       | unanimous     | carry        | whiting       | 207 ■           |
| 73  | legislator    | player        | football      | leg           | town         | stadium       | 695 ■■          |
| 95  | online        | email         | website       | browser       | contact      | server        | 351 ■           |
| 70  | election      | ballot        | lobbyist      | voter         | candidate    | campaign      | 407 ■           |
| 74  | tentative     | conditional   | approval      | grading       | attachment   | deviation     | 177 ■           |
| 79  | snow          | remember      | plow          | lock          | scam         | sure          | 888 ■■          |
| 28  | craftsman     | revival       | historic      | gabled        | bungalow     | historical    | 882 ■■          |
| 114 | park          | playground    | recreation    | picnic        | mesa         | trail         | 235 ■           |
| 11  | tuberculosis  | infection     | hepatitis     | overdose      | influenza    | vaccine       | 1515 ■■■■       |
| 21  | think         | something     | want          | thing         | talk         | everybody     | 1155 ■■■        |
| 86  | sewer         | sanitary      | water         | pipeline      | drinking     | wastewater    | 176 ■           |
| 59  | fort          | worth         | plot          | tad           | falls        | demo          | 192 ■           |
| 20  | subsection    | licensee      | article       | chapter       | sec          | shall         | 214 ■           |
| 47  | inf           | micron        | effluent      | eff           | sludge       | isomer        | 591 ■■          |
| 62  | bid           | buyer         | seller        | bidder        | price        | quote         | 357 ■           |
| 48  | contributor   | instruction   | filer         | political     | officeholder | payee         | 79 ■            |
| 104 | provisions    | subcontractor | surety        | rev           | bidder       | supplementary | 232 ■           |
| 27  | breach        | franchisee    | hereunder     | agreement     | remedy       | agree         | 213 ■           |
| 112 | youth         | camp          | teach         | teen          | lesson       | yoga          | 722 ■■          |
| 23  | dog           | rabies        | euthanasia    | euthanized    | pet          | spay          | 1710 ■■■■■      |
| 35  | trust         | revocable     | mfr           | apportionment | living       | assn          | 285 ■           |
| 116 | emergency     | preparedness  | null          | dispatch      | rescue       | fire          | 340 ■           |
| 80  | energy        | efficiency    | customer      | saving        | rebate       | renewable     | 382 ■           |
| 113 | proud         | leadership    | honor         | pleased       | grateful     | passion       | 1168 ■■■        |
| 18  | garland       | invoice       | assoc         | check         | firefighter  | association   | 152 ■           |
| 81  | page          | last          | sub           | update        | prime        | award         | 17 ■            |
| 2   | mosquito      | insecticide   | spray         | bait          | repellent    | pesticide     | 997 ■■■         |
| 120 | project       | improvement   | funding       | justification | completion   | acquisition   | 47 ■            |
| 105 | thence        | plat          | easement      | annexation    | pud          | westerly      | 255 ■           |
| 118 | comment       | concern       | suggest       | clarify       | suggestion   | dear          | 307 ■           |
| 34  | library       | campus        | doe           | branch        | center       | arena         | 208 ■           |
| 40  | portfolio     | treasury      | investment    | maturity      | yield        | liquidity     | 250 ■           |
| 115 | masonry       | plaster       | joist         | stud          | sheathing    | ceiling       | 875 ■■          |
| 53  | department    | authority     | dpt           | correction    | citywide     | transit       | 109 ■           |
| 3   | vend          | utensil       | meat          | fat           | cheese       | salad         | 1325 ■■■■       |
| 8   | assessor      | taxpayer      | determination | informal      | petition     | notification  | 39 ■            |
| 58  | recycling     | bag           | garbage       | recycle       | recyclable   | recyclables   | 318 ■           |
| 87  | sign          | billboard     | pole          | speeding      | illuminate   | banner        | 472 ■■          |
| 31  | student       | elementary    | school        | college       | graduate     | academic      | 233 ■           |
| 32  | dwelling      | alteration    | plumbing      | plumb         | canceled     | mechanical    | 143 ■           |
| 51  | combustible   | vent          | piping        | conductor     | duct         | flammable     | 517 ■■          |
| 91  | app           | credit        | download      | post          | issued       | agent         | 57 ■            |
| 66  | wetland       | vernal        | riparian      | habitat       | specie       | species       | 1040 ■■■        |
| 44  | findings      | string        | tank          | carcinogen    | qty          | lust          | 128 ■           |
| 42  | contamination | spill         | remediation   | groundwater   | asbestos     | hazardous     | 343 ■           |
| 99  | prep          | batch         | qualifier     | analytical    | surrogate    | sample        | 313 ■           |
| 84  | airport       | facility      | aviation      | maintenance   | operation    | aircraft      | 150 ■           |
| 19  | accessory     | height        | dwell         | frontage      | setback      | subsection    | 218 ■           |
| 6   | householder   | poverty       | disability    | married       | husband      | universe      | 93 ■            |
| 98  | obesity       | sugary        | epidemic      | soda          | sensible     | drink         | 65 ■            |
| 33  | avenue        | street        | west          | east          | boulevard    | south         | 98 ■            |
| 10  | ductible      | copay         | prescription  | coinsurance   | outpatient   | inpatient     | 488 ■■          |
| 50  | ductile       | trench        | pipe          | manhole       | coupling     | compaction    | 705 ■■          |
| 17  | margin        | error         | occupied      | race          | occupy       | islander      | 79 ■            |
| 5   | earthquake    | flood         | floodplain    | flooding      | landslide    | fault         | 723 ■■          |
| 76  | variance      | setback       | yard          | exception     | fence        | front         | 94 ■            |
| 16  | business      | marijuana     | cannabis      | manufacturing | industry     | collective    | 319 ■           |
| 108 | fugitive      | bio           | exhaust       | unmitigated   | noise        | receptor      | 262 ■           |

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

| #   | Top Word 1     | Top Word 2     | Top Word 3    | Top Word 4    | Top Word 5    | Top Word 6    | Tokens assigned |
|-----|----------------|----------------|---------------|---------------|---------------|---------------|-----------------|
| 29  | labor          | worker         | force         | unemployed    | earnings      | civilian      | 80              |
| 111 | discharge      | pollutant      | inspection    | inspect       | pollution     | inspector     | 109             |
| 68  | contractual    | parts          | duke          | outside       | postage       | receipts      | 274             |
| 77  | curb           | pavement       | sidewalk      | ramp          | gutter        | asphalt       | 390             |
| 65  | draft          | update         | process       | review        | staff         | progress      | 67              |
| 24  | landlord       | tenant         | renewal       | rent          | lease         | expired       | 255             |
| 106 | consultant     | proposer       | procurement   | contract      | firm          | subcontractor | 179             |
| 43  | blanket        | medicare       | payroll       | premium       | undistributed | refund        | 107             |
| 103 | urban          | mixed          | density       | redevelopment | development   | industrial    | 115             |
| 89  | taxable        | res            | deed          | value         | homestead     | star          | 41              |
| 83  | building       | demolition     | story         | demolish      | floor         | build         | 82              |
| 119 | cost           | estimate       | estimated     | initial       | costs         | change        | 52              |
| 109 | respondent     | satisfied      | dissatisfied  | survey        | satisfaction  | disagree      | 403             |
| 64  | must           | signature      | copy          | application   | applicant     | submission    | 139             |
| 26  | tax            | deduction      | amt           | assessed      | bill          | abatement     | 171             |
| 78  | yes            | worksheet      | text          | pic           | font          | button        | 476             |
| 7   | greenhouse     | emission       | coal          | climate       | ozone         | dioxide       | 334             |
| 54  | parking        | tow            | taxi          | vehicle       | shuttle       | passenger     | 236             |
| 41  | assistant      | analyst        | technician    | aide          | specialist    | asst          | 119             |
| 22  | allocation     | val            | cove          | acct          | glen          | subtotal      | 79              |
| 63  | fee            | charge         | license       | reservation   | surcharge     | refundable    | 143             |
| 117 | delay          | perm           | queue         | peak          | flt           | detector      | 113             |
| 4   | datum          | database       | copyright     | accuracy      | data          | compile       | 193             |
| 45  | audit          | auditor        | auditing      | internal      | implemented   | procedure     | 222             |
| 100 | mitigation     | impact         | significant   | adverse       | significance  | unavoidable   | 136             |
| 88  | gender         | discrimination | transgender   | immigrant     | immigration   | religion      | 859             |
| 9   | district       | zoning         | maker         | vacancy       | speaker       | planner       | 45              |
| 12  | artist         | artwork        | art           | arts          | mural         | sculpture     | 1055            |
| 94  | contracted     | encumbrance    | unencumbered  | exp           | expend        | bud           | 71              |
| 110 | rouge          | parish         | baton         | thereto       | sewerage      | adjudicate    | 464             |
| 46  | commissioner   | chair          | commission    | committee     | briefing      | advisory      | 187             |
| 85  | sch            | min            | tin           | hump          | carpool       | qua           | 390             |
| 15  | complainant    | allegation     | allege        | complaint     | doc           | misconduct    | 963             |
| 30  | incumbent      | examination    | supervision   | knowledge     | exam          | ability       | 410             |
| 107 | savings        | ltd            | village       | neighborhood  | excise        | costs         | 81              |
| 72  | imp            | burglary       | theft         | testify       | petitioner    | mischief      | 116             |
| 60  | bike           | bicycle        | bicyclist     | pedestrian    | route         | mobility      | 336             |
| 82  | accomplishment | narrative      | grantee       | outcome       | objective     | mod           | 101             |
| 36  | decline        | trend          | recession     | average       | rate          | percentage    | 265             |
| 52  | homeless       | homelessness   | supportive    | consolidated  | transitional  | counseling    | 193             |
| 1   | alderman       | resolved       | whereas       | resolution    | authorizing   | authorize     | 245             |
| 92  | concept        | design         | realm         | visual        | character     | conceptual    | 433             |
| 71  | bond           | obligation     | proceeds      | redemption    | debt          | series        | 174             |
| 67  | dist           | applied        | col           | occupancy     | valuation     | monoxide      | 62              |
| 25  | scenario       | figure         | appendix      | assume        | assumption    | model         | 162             |
| 38  | horas          | persona        | para          | yon           | sou           | ante          | 1350            |
| 14  | federal        | agency         | entity        | recipient     | grant         | eligible      | 90              |
| 56  | waterfront     | shoreline      | marina        | beach         | port          | boat          | 844             |
| 61  | revenue        | balance        | expenditure   | reserve       | forecast      | budget        | 101             |
| 75  | governmental   | asset          | liability     | assets        | statement     | pension       | 142             |
| 37  | endorse        | endorsed       | budget        | proposed      | adopted       | adopt         | 111             |
| 69  | tree           | planned        | circumference | gross         | density       | infill        | 211             |
| 90  | councilman     | introduced     | ordain        | ordinance     | digest        | yea           | 244             |
| 97  | actuarial      | grievance      | employee      | retirement    | bargaining    | actuary       | 250             |
| 39  | affordable     | housing        | affordability | homeowner     | income        | bedroom       | 150             |
| 55  | ave            | combo          | blossom       | pearl         | cir           | olive         | 1091            |
| 13  | strategy       | goal           | strategic     | stakeholder   | focus         | initiative    | 162             |
| 57  | absent         | int            | preside       | ordained      | tag           | numbers       | 194             |
| 101 | violent        | gang           | firearm       | offender      | crime         | patrol        | 511             |
| 93  | shooting       | suspect        | pronounce     | gunshot       | flee          | shoot         | 730             |

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

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

Table 8: OLD: Top words from a structural topic model with 120 topics (first 60 topics displayed here) and FREX scoring. Colors depict partisanship based on coefficient size. White cells are non-significant topics.

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

Table 9: OLD: 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.