Online Appendix

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

January 16, 2020

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.

1 Overview

In this online appendix we include supporting information about our data, the data collection process, the data collection and cleaning pipeline, and some additional analyses. In the first section, we present additional details on data collection, along with some additional descriptive statistics. In the second section, we provide additional details on, and results from, our topic modeling analysis.

2 Data collection methods and sources

We acquired the municipal 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.[1] 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 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

[1] 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.

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

The partisanship of the mayor of each city is coded in different ways, depending on the state.

[2]An alternative source of web data which has attained some popularity within the digital humanities is the Internet Archive's Wayback Machine (https://archive.org/), which preserves snapshots of websites over time. In theory, this would be very useful to projects such as ours, since it would allow us to measure the evolution of websites over time, in response to changes in city executives. However, the Wayback Machine suffers from some limitations that, in our eyes, inhibit its usefulness for scientific research in general, and our project in particular. First of all, the Internet Archive does not conduct all of its web crawling itself. Rather, third parties donate their crawling data to the Internet Archive. This means that the source of the data varies and the crawling process often remains opaque, which is a problem with respect to scientific transparency. Second, this also means that the crawling isn't done in regular intervals unless the website in question is extremely prominent. Our research is focused on municipal websites, many of which are fairly obscure, preventing the Wayback Machine from being useful to us. For example, as of the time of writing, the website of Attica, IN, has only been crawled 25 times since the Wayback Machine's inception in 1996 (and in this case, all of these crawls are from 2014 on). By contrast, the website of New York City has been crawled over 7,000 times. Third, webcrawls are not instant and are not always done on an entire website at the same time. This means that for example, a website's frontpage might have been scraped one day, and its page on the city's mayor only a month later – within the same crawl. Once again, this problem is exacerbated for small, less prominent websites. While we appreciate the Internet Archive's efforts to preserve snapshots of the web and recommend its API to practitioners who can work around the limitations outlined here, these problems preclude its usefulness for our purposes.

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

For Indiana, where elections are nominally partisan, this information is accessible through the state government's website[4]. For Louisiana, we received data on the outcomes of mayoral elections from the Local Elections in America Project (LEAP) (Marschall and Shah 2013). For the other states, where mayoral elections are not nominally partisan (but the partisanship of the mayor is still well-known), we employed different means: For New York and Washington, we searched the state campaign finance websites, and coded the parties of the candidates based on the party committees from which they received donations. For California and Texas, where our data consists of highly populated cities, partisanship information was acquired from Ballotpedia[5]. Finally, we also scraped mayoral partisanship from the cities' Wikipedia pages. When compared to the other data sources above, (and manual searches in case of conflicts) Wikipedia proved to be very reliable and added additional cases to our dataset even for Indiana and Louisiana. Generally speaking, we found data scraped from Wikipedia, aided by manual corrections in case of missing or conflicting data, to be more reliable than data from governmental sources.[6]

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[7].

Tables 1 and 2 provide additional information about the data collected for this project. In Table 1, we present the state-by-state breakdown of the mayoral partisanship of the cities collected in the respective state. In Table 2, we present the distribution of file extensions before and after processing.

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

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

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

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

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.

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

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

3 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.[8] 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,

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

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

3.1 Site to Text Conversion

3.1.1 File Type Detection

The format of a file has a major impact on whether and how textual data can be extracted from a document. For the most part, the file type of a document can be correctly determined through the filename ending—its extension. However, there are exceptions to this, which, if ignored, can lead to large amounts of improperly formatted text. For example, we found thousands of documents that ended in .html, when they were actually PDFs. A more accurate test for file type relies on the use of magic numbers, a short sequence of bytes at the start (and sometimes end) of files that is unique for each file type and therefore allows its correct identification. We implement this method using the R package `wand` (Rudis et al. 2016).[9]

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

3.1.2 Extracting Text from HTML

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

A subfield of the information retrieval literature, dealing with boilerplate extraction, can offer a solution to this problem. The goal of this branch of research is to develop algorithms with the ability to estimate whether a given portion of an HTML file is substantive. To this end, structural features, such as HTML tags (which are not sufficiently informative on their own), text statistics such as word and sentence length, as well as other heuristics are used. We rely on the `boilerpipe` classifier described in Kohlschütter et al. (2010), which is implemented through the R package `boilerpipeR`. The `boilerpipe` algorithm has been widely used in the computer science and natural language processing literatures, but to our knowledge has not been previously used in the social sciences. The complete text extracted from the Gary, IN homepage using boilerpipe is depicted in the screenshot in the bottom row of Figure 1. We see that only the Mayor's message is extracted, leaving the rest of the text as boilerplate.[10]

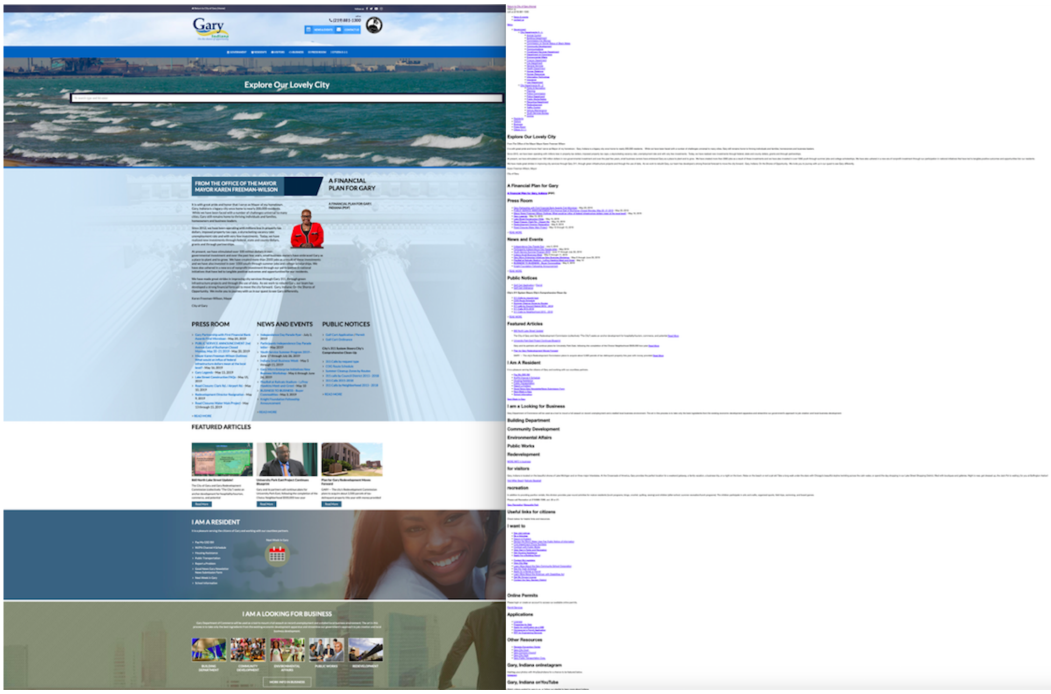
3.1.3 Extracting Text from PDF, DOC, DOCX and TXT

The extraction of information from other text-based file formats is more straightforward.[11] To this end, we rely on `readtext` R package (Benoit and Obeng 2019), which is a wrapper for a

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

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

(a) Naive Parsing



(b) Boilerpipe

From The Office of the Mayor Mayor Karen Freeman-Wilson

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

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

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

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

Karen Freeman-Wilson, Mayor

Figure 1: The top image provides a side-by-side depiction of the entire homepage of https://garyin.us/, accessed on 05/22/2019, and complete/naive extraction of all of the text on the site. Bottom image provides the result of running https://garyin.us/ through the `boilerpipe` algorithm at https://boilerpipe-web.appspot.com/.

set of parsers.[12,13] The breakdown of all files by type is given in the online appendix. The most frequent file type besides HTML is PDF, from which we are able to extract a substantial amount of usable text. Files of type DOC, TXT, and DOCX, also occur regularly in our corpus and offer a considerable volume of textual data.

3.2 Preprocessing

Preprocessing is an important part of text-as-data research and choices made therein can have significant effects on the outcomes of an analysis (Denny and Spirling 2018). As such, our advice given in this section, more than in any other, is specific to the problem of extracting meaningful textual information from municipal government websites, with the end goal of its use in a bag-of-words-based model. The techniques we employ might also be of use in other types of applications, but by no means should this section be regarded as a general-purpose manual for preprocessing. The challenge in conducting preprocessing for a comparative analysis of websites lies in the considerable variance between websites. Some of it is substantively informative and some of it is completely irrelevant. As an example of the latter, names of city officials and citizen petitioners feature frequently in city documents. The same is true for streets, locations and not least of all, the city itself. Since individual names recur at a much higher rate within a city than across the entire corpus, this would cause a topic model to cluster its topics by city. Consequently we require a tool which detects the signal in the noise and does so consistently for a discordant set of sources.

To this end, we turn to a common method in natural language processing—part-of-speech (POS) tagging and named entity recognition (NER). In our case, names are the source of substantively uninteresting heterogeneity between cities, so NER is used to detect and remove them.[14] However, we caution here that for many other applications, where the names of political actors

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

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

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

might be of interest, this step is not recommended. Furthermore, we select words on the basis of their POS-tags, retaining only nouns (the modal category), verbs, and adjectives.[15] Furthermore, we keep proper nouns that also occur as nouns—this removes names, but retains titles such as "Police Chief" which can appear as proper nouns if they are followed by a name. Finally, we also conduct lemmatization to reduce words to their basic form.[16] POS-tagging, NER and lemmatization are all implemented through `spacyr`. To deal with any leftover issues, we remove words with less than three characters (these are usually artifacts from improperly encoded documents and faulty or impartial optical character recognition), stopwords and non-English words (using the R package `Hunspell`). A final and crucial step is the removal of duplicate documents, which occur very frequently on websites. In addition to their primary purpose, the previous preprocessing steps also help in stripping otherwise identical documents of information that makes them unique – such as names and dates – thus facilitating their deletion.

After preprocessing, our corpus consists of 356,911 documents. In Table 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`.

The biggest limitation in our pipeline, and an open area for future research, is the reliance on `wget` to gather the initial website files. By using `wget`, we miss content that is displayed dynamically on websites using JavaScript. For any one website, it would be possible to customize a routine with `Selenium` to access dynamic elements, but the process would need to be customized for each website.[17]

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

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

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

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

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

4 Supplemental Informational on Topic Modeling Application

The structural topic model is implemented in the R package STM (Roberts et al. 2018). We use 60 topics—the number recommended by the authors[18] for medium- to large-sized corpora.[19] We use four covariates: First, *party*, to estimate the difference in topic prevalence based on whether mayors are Republican or Democratic. Second, *city population*, which the literature frequently emphasizes as a determinant of the issues a city faces (see, for example, Guillamón et al. (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.[20]

[18]For this recommendation, see the documentation for the function stm() in version 1.3.0 of the R package stm (Roberts et al. 2018).

[19]Since our corpus is at the larger end of that spectrum, we also estimated a model with 120 topics, but found no notable differences.

[20]The "Fightin' Words" methodology developed by Monroe et al. (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.

4.1 Supplemental results

In Table 4, we present topic results from an STM, and mayoral party effect, estimated on data processed using a minimal HTML parser, implemented as the 'KeepEvertything' algorithm in the boilerpipeR package. An illustration of the content removed by the full boilerpipe procedure is given by Topic 56. This is the fourth-most Republican topic, and is a website boilerplate topic, with top words 'click', 'reserved', 'trademark', 'password', 'search', and 'connected'. As further illustration, the second-most Democratic topic is another website boilerplate topic, with top words 'sitemap', 'clerk', 'calendar', 'online', 'bureau', and 'alert'. We do not see any clear boilerplate topics when using data processed using boilerpipe.

In Tables 5 and 6, we present results of the STM with 120 topics organized according to the effect of mayoral partisanship on topic prevalence. The partisan themes in the 120-topic STM mirror those in the 60-topic model, with cities led by Democratic mayors focused disproportionately on finances (e.g., Topics 75, 71, 61) and social problems (e.g., Topics 93, 101, 39, 52), and cities led by Republican mayors focused disproportionately on basic services and utilities (e.g., Topics 114, 11, 86, 116).

In Tables 7 and 8, we present the topics ordered according to the effects of median income and population, respectively. Considering income-based variation in topics, there are several topics prevalent in more wealthy cities that focus on initiatives that go well-beyond standard city services—downtown and building revitalization (Topics 39 and 12), renewable energy (Topic 20), bike/pedestrian-oriented development (Topic 60), wildlife conservation (Topic 3). Such topics cannot be found among those that are more prevalent in less wealthy cities. When it comes to population, more populous cities deal disproportionately with issues related to public health (Topics 2 and 11), crime (Topic 59), homelessness (Topic 52), and diversity (Topic 6).

In Figure 2, we present a quantitative assessment of relative topic quality for our main model, using two metrics, semantic coherence and exclusivity. Semantic coherence (Mimno et al. 2011)

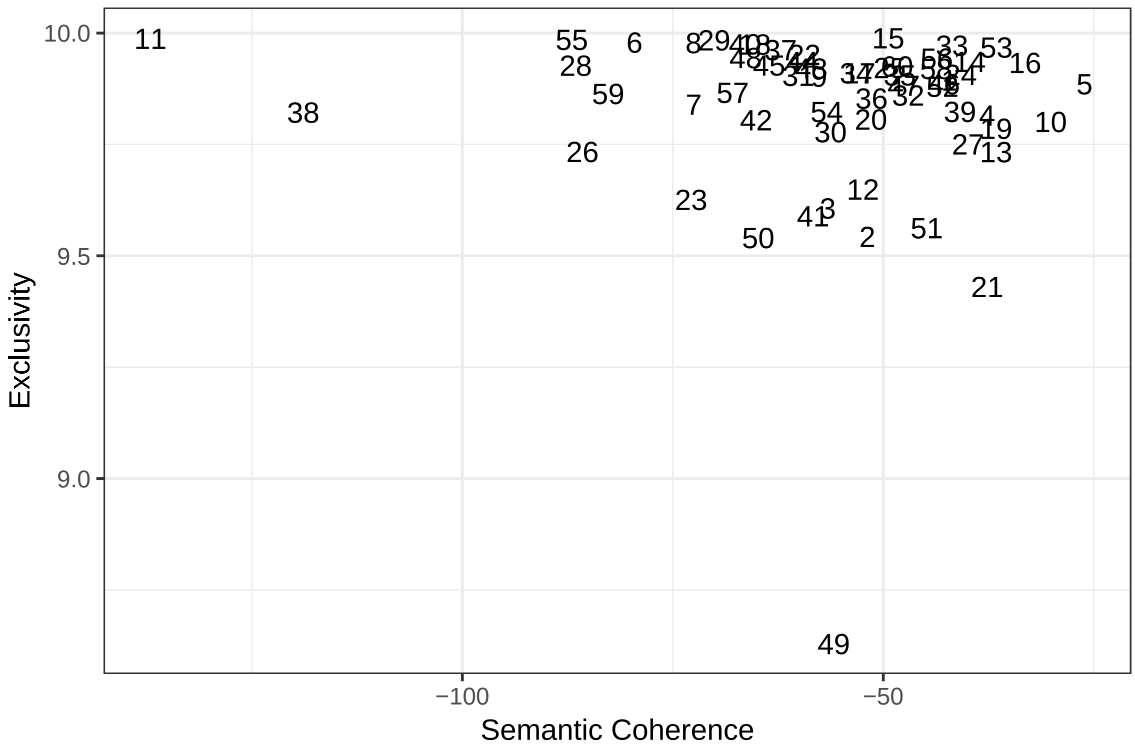


Figure 2: Semantic coherence and exclusivity for each of the 60 topics in our main structural topic model. Coherence describes the extent to which the 10 top words in a topic model belong to the same underlying concept. Exclusivity measures whether the top words in one topic feature primarily in this topic, rather than being dispersed across a range of topics. The model performs well in this trade-off, as both coherence and exclusivity are high for most topics.

describes the extent to which the 10 top words in a topic model belong to the same underlying concept. Exclusivity (Bischof and Airoldi 2012; Roberts et al. 2014) measures whether the top 10 words in one topic feature primarily in this topic, rather than being dispersed across a range of topics. As coherence tends to be outright better in models with fewer topics, the comparison with exclusivity creates a trade-off. Figure 2 shows that the model performs well in this regard, as most topics have both high coherence and exclusivity. Even topic 11, which does worst in terms of coherence, looks like a good topic upon manual inspection, as its top words all pertain to obesity.

References

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

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

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

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

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

Bischof, J. and E. Airoldi (2012). Summarizing topical content with word frequency and exclusivity. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012 1*.

Darwin, I. (2008). Libmagic.

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

Guillamón, M. D., F. Bastida, and B. Benito (2013). The electoral budget cycle on municipal police expenditure. *European Journal of Law and Economics 36*(3), 447–469.

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

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

Mimno, D., H. M. Wallach, E. Talley, M. Leenders, and A. McCallum (2011). Optimizing semantic coherence in topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (2), 262–272.

Monroe, B. L., M. P. Colaresi, and K. 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.

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

Roberts, M. E., B. M. Stewart, and D. Tingley (2018). *stm: R Package for Structural Topic Models*. R package version 1.3.3.

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

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

#	Top Word 1	Top Word 2	Top Word 3	Top Word 4	Top Word 5	Top Word 6	Tokens assigned
30	riverfront	department	authority	transit	police	enjoy	212
55	posted	dream	celebration	broadcast	ballpark	football	544
5	trust	revocable	leisure	learn	mfr	living	327
56	click	reserved	trademark	password	search	connected	387
49	chair	agenda	subcommittee	briefing	presentation	committee	416
47	motion	second	adjourn	carry	whiting	unanimous	451
51	storm	drain	sanitary	water	sewer	infiltration	391
43	article	subsection	shall	provision	chapter	unlawful	467
2	virus	tuberculosis	infection	influenza	hepatitis	cannabis	2555
1	councilman	whereas	alderman	resolved	yea	resolution	576
59	subcontractor	proposer	bidder	bid	consultant	subcontract	508
33	eff	inf	effluent	ether	batch	isomer	1240
54	dwelling	alteration	plumbing	canceled	plumb	mechanical	311
12	think	something	somebody	appreciate	seem	everything	2911
21	artist	ceremony	jazz	celebrate	prize	yoga	3326
11	disaster	evacuation	marshal	apparatus	tornado	aircraft	1174
4	craftsman	architecture	facade	distinctive	architectural	historic	1634
34	contributor	filer	officeholder	political	payee	candidate	272
17	assessor	informal	taxpayer	doc	viewing	determination	442
18	setback	variance	plat	height	thence	frontage	472
19	findings	tank	carcinogen	string	qty	yon	247
3	vend	meat	utensil	towel	fat	cheese	2791
38	application	applicant	must	copy	tenant	mail	390
60	wetland	shoreline	vernal	riparian	habitat	marsh	1986
26	student	teacher	classroom	beech	academic	doe	832
7	obesity	sugary	epidemic	drink	sensible	ounce	98
20	garland	invoice	assoc	rouge	baton	vendor	524
31	credit	docket	bbl	agent	month	app	61
40	slideshow	arrow	printer	stumble	blogger	google	189
8	deductible	outpatient	prescription	coinsurance	copay	inpatient	809
41	playground	park	tennis	picnic	viewpoint	ravine	405
45	taxable	res	deed	value	homestead	star	106
36	thickness	fitting	conduit	conductor	ductile	trench	1756
22	noise	mitigation	impact	significant	adverse	sensitive	346
35	householder	universe	margin	poverty	race	census	251
15	imp	amt	micron	rend	land	sustain	117
16	dist	applied	col	occupancy	monoxide	valuation	123
58	pickup	bag	recyclable	bin	landfill	curbside	651
46	contracted	medicare	allocation	subtotal	unencumbered	payroll	236
14	perm	queue	delay	peak	adj	volume	232
39	successor	franchisee	covenant	redemption	bankruptcy	obligation	678
44	prune	circumference	tree	planting	shrub	root	1798
23	tax	increment	deduction	abatement	assessed	levy	397
6	fugitive	exhaust	renewable	bio	emission	coal	859
57	savings	costs	capital	ltd	improvement	excise	172
27	density	mixed	planned	infill	orient	retail	365
52	bicycle	pedestrian	bike	route	bus	curb	572
48	actuarial	asset	governmental	assets	investment	debt	342
37	supervisor	technician	aide	incumbent	employee	trainee	758
32	introduced	absent	councilor	preside	digest	legislator	615
28	persona	para	sin	ante	junta	combo	2376
24	audit	procedure	effectiveness	ensure	software	timely	497
42	budget	endorse	revenue	endorsed	balance	expenditure	232
13	homeless	affordable	affordability	housing	supportive	homelessness	380
10	impound	taxicab	license	cat	dog	neuter	700
9	strategy	stakeholder	focus	goal	engagement	outreach	732
25	profile	executive	bend	sustainability	cleanup	rates	110
53	theft	burglary	fwy	aggravated	aggravate	robbery	253
29	sitemap	clerk	calendar	online	bureau	alert	124
50	arrest	complainant	allegation	shooting	homicide	victim	1665

Table 4: Top words from a structural topic model with 60 topics and FREX scoring, with data processed using a minimal HTML parser. 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	deductible	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 5: Top words from a structural topic model with 120 topics (first 60 topics displayed here) and FREX scoring. Colors depict partisanship based on coefficient size. White cells are non-significant topics.

#	Top Word 1	Top Word 2	Top Word 3	Top Word 4	Top Word 5	Top Word 6	Tokens assigned
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 6: Top words from a structural topic model with 120 topics (second 60 topics displayed here) and FREX scoring. Colors depict partisanship based on coefficient size. White cells are non-significant topics.

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

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

Table 8: Top words from a structural topic model with 60 topics and FREX scoring. Colors depict city population based on coefficient size (larger cities are cyan, smaller cities are magenta). White cells are non-significant topics.