

Online Appendix

Government websites as data:

A methodological pipeline with application to the websites of municipalities in the United States

Markus Neumann* Fridolin Linder† Bruce Desmarais‡

May 24, 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.

1 Overview

In this online appendix we include supporting information about our data, the data collection process, and some additional analyses. In the first section, we present additional details on data

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

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

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

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

¹The 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.

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 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.²

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

²There 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.

government's website³. 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 Ballotpedia⁴. 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.⁵

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 Bureau⁶.

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.

³<http://www.in.gov/apps/sos/election/general/general2015?page=office&countyID=1&officeID=32&districtID=-1&candidate=>

⁴https://ballotpedia.org/List_of_current_mayors_of_the_top_100_cities_in_the_United_States

⁵In Indiana, the data includes only cities - incorporated municipalities with at least 2,000 inhabitants - as opposed to towns.

⁶<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 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.

3 Supplemental Information 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 authors⁷ for medium- to large-sized corpora.⁸ 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

⁷For this recommendation, see the documentation for the function `stm()` in version 1.3.0 of the R package *stm* (Roberts et al. 2018).

⁸Since our corpus is at the larger end of that spectrum, we also estimated a model with 120 topics, but found no notable differences.

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.⁹

3.1 Supplemental results

In Tables 3 and 4 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 6 and 5 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).

⁹The “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.

Funding

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

References

- DeSantis, V. S. and T. Renner (2002). City government structures: An attempt at clarification. *State and Local Government Review* 34(2), 95–104.
- Gerber, E. R. and D. J. Hopkins (2011). When mayors matter: estimating the impact of mayoral partisanship on city policy. *American Journal of Political Science* 55(2), 326–339.
- 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.
- 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)*.
- 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.
- Morgan, D. R. and S. S. Watson (1992). Policy leadership in council-manager cities: Comparing mayor and manager. *Public Administration Review*, 438–446.
- Roberts, M. E., B. M. Stewart, and D. Tingley (2018). *stm: R Package for Structural Topic Models*. R package version 1.3.3.

#	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	dwel	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 3: 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 4: 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 5: 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 6: 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.