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

Markus Neumann, Fridolin Linder, Bruce Desmarais
Department of Political Science - Penn State University

Overview

- Websites of local governments are an important source of information for citizens
- Extant research of government websites has largely relied on manual coding
- We develop a methodological pipeline for the automated analysis of government websites
- We demonstrate the use of this pipeline on the websites of 234 municipal websites in CA, IN, LA, NY, TX and WA.
- We show that content varies with the partisanship of the mayor.

Data Collection

- Scraping city URLs from Wikipedia and the General Services Administration, which keeps a list of all .gov addresses
- Verification of which websites actually work through an automated browser
- Downloading the websites through `wget`
- Information on mayoral partisanship from state websites, the LEAP project, Ballotpedia and Wikipedia

Site to Text Conversion

- The file endings from open data sources such as governmental websites are sometimes wrong. This causes problems when converting to text.
- We use file signatures to identify the correct type before conversion with `readtext`.
- Used file types: Html, xml, pdf, doc, docx, txt
- Everything is converted to .txt
- Preprocessing: Lowercase; removal of punctuation, numbers, dates, etc.; spellchecking/Removal of non-English words; Lemmatization

Boilerplate Removal

- Websites contain a lot of text that is not substantively interesting
- This text is often repeated on each page of a site - for example website elements such as “You are here”, ”Home”, or the names of city officials and offices.
- If this content is not removed, the tool of analysis will associate specific patterns of boilerplate with the respective cities
- Problem: how to remove boilerplate content without dropping useful information?

Boilerplate Classifier

- Solution: Train a classifier (random forest) on manually annotated lines of website content. A line can be flagged as either substantively useful or boilerplate.
- The classifier relies on the following features:
  - Line length (number of words characters)
  - Number of line duplications within a website
  - Line position in the document (median distance of a line to the document midpoint)

Analysis

- Structural topic model with 60 topics
- Covariates: Party, state, population, median income

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

Conclusion

- Cities with Republican mayors provide more information about basic utilities such as water, energy, fire safety, or natural disaster protection.
- Cities with Democratic mayors provide more information about policy deliberation, crime control, or public housing.
- These findings call into question the commonly held notion that politics at the municipal level is largely non-partisan.
- We plan to implement the pipeline in an R package