Government websites as data: A methodological pipeline for collection, processing, and text analysis

Markus Neumann
Fridolin Linder
Bruce Desmarais

The Pennsylvania State University

January 6, 2018

Government Websites

- Content of government websites is an important source of information & transparency
- After coming into power, the Trump administration has made some controversial changes to the websites of federal agencies
- Website content is political
- Partisanship of city government is expected to have an effect

Example Website

THE TOWN OF Arcadia, Louisiana

HOME GOVERNMENT SERVICES ATTRACTIONS FOOD & LODGING CONTACT

Directions & Map History Bonnie & Clyde Photos & Events Economic Development Pay Water Bill

ARE YOU WATER AWARE?

Water is a precious resource. It's important to use water wisely, particularly during extended dry weather. By following these simple suggestions, you'll save money on your water bill while conserving the supply we all depend on.

Check faucets and pipes for leaks

A small drip from a worn faucet washer can waste 20 gallons of water per day. Larger leaks can waste hundreds of gallons.

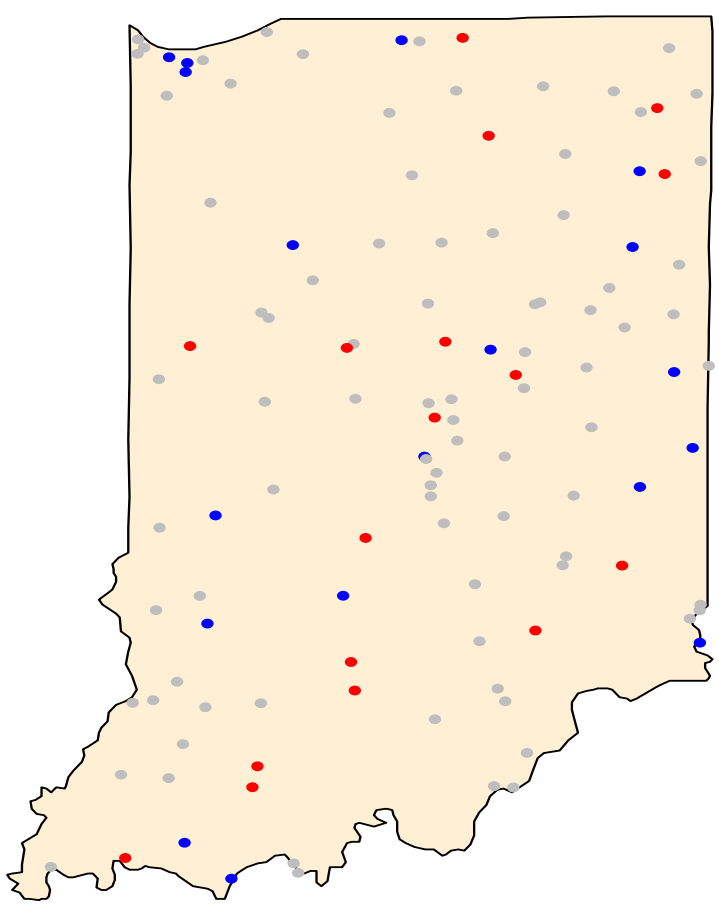
Check your toilets for leaks

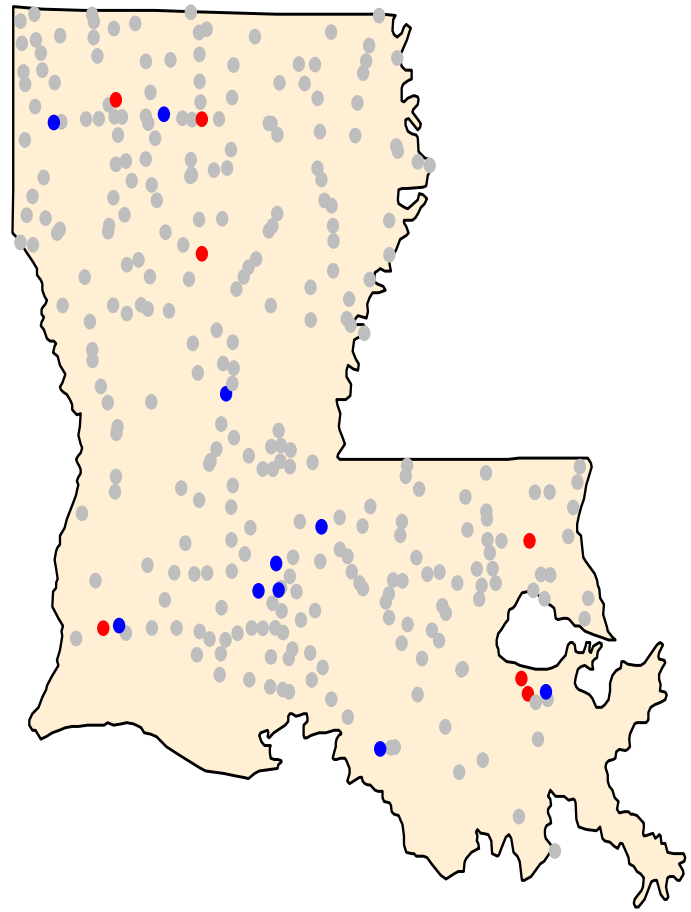
Put a little food coloring in your toilet tank. If, without flushing, the color begins to appear in the bowl within 30 minutes, you have a leak that should be repaired immediately. Most replacement parts are inexpensive and easy to install.

Local Government Websites

- Most local (i.e. mayoral) elections are non-partisan
- Few states have exclusively partisan local elections
- Data for these elections can be difficult to find
- We selected Indiana and Louisiana

Data Overview





The Pipeline

Data Collection → **Preprocessing** → **Analysis**

- Identify URLs
- Verify URLs (browser automation)
- Download websites
- Determine file type
- Convert to txt

- Remove punctuation, dates, etc.
- To lowercase
- Boilerplate removal
- Spellchecking
- Lemmatization (city & cities = city)

- Fightin' Words
- Structural topic model

The Pipeline

Data Collection → **Preprocessing** → **Analysis**

- Identify URLs
- Verify URLs (browser automation)
- Download websites
- Determine file type
- Convert to txt

- Remove punctuation, dates, etc.
- To lowercase
- **Boilerplate removal**
- Spellchecking
- Lemmatization (city & cities = city)

- Fightin' Words
- Structural topic model

Boilerplate Removal

THE TOWN OF Arcadia, Louisiana

HOME GOVERNMENT SERVICES ATTRACTIONS FOOD & LODGING CONTACT

Directions & Map History Bonnie & Clyde Photos & Events Economic Development Pay Water Bill

ARE YOU WATER AWARE?

Water is a precious resource. It's important to use water wisely, particularly during extended dry weather. By following these simple suggestions, you'll save money on your water bill while conserving the supply we all depend on.

Check faucets and pipes for leaks

A small drip from a worn faucet washer can waste 20 gallons of water per day. Larger leaks can waste hundreds of gallons.

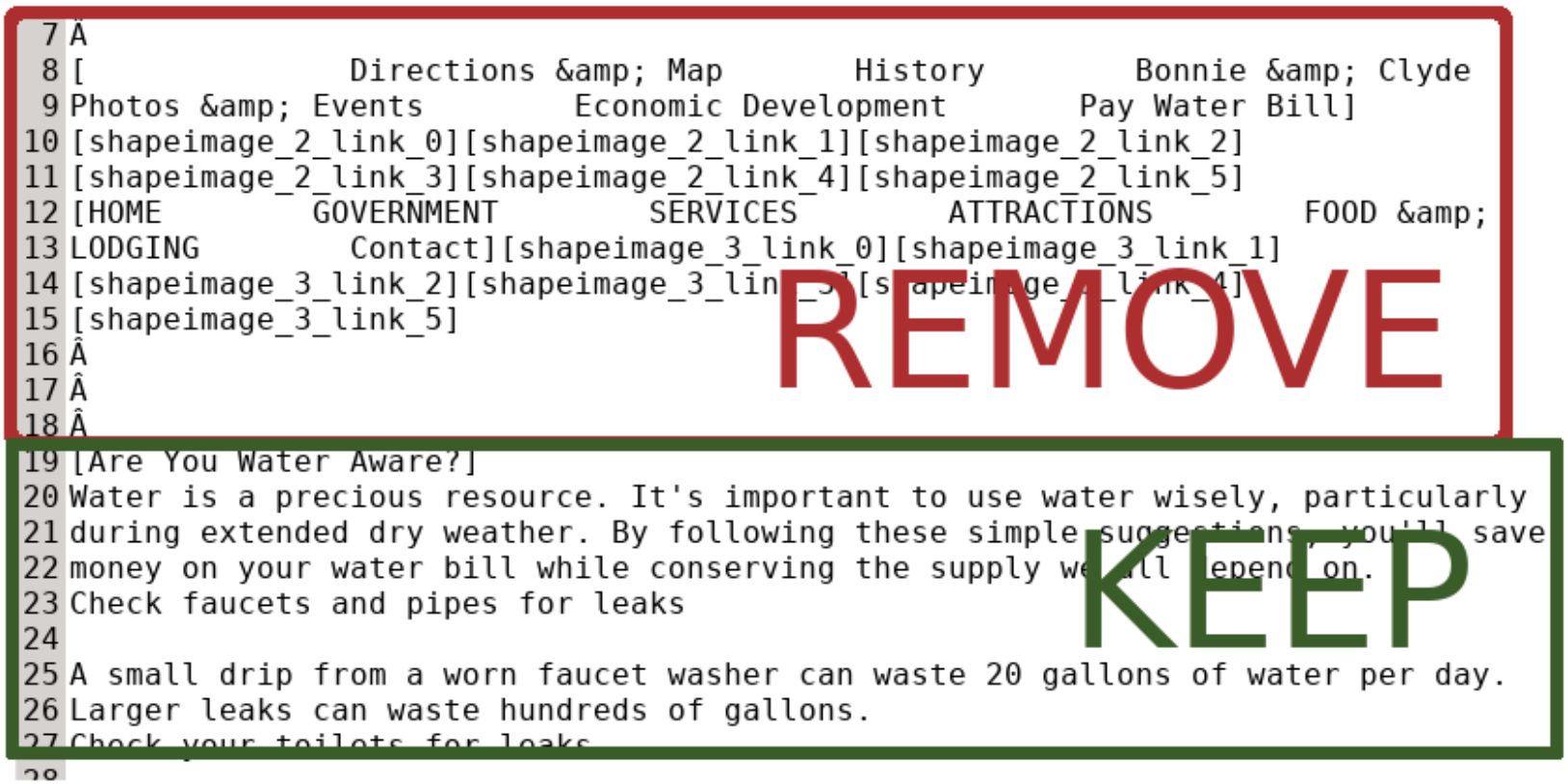
Check your toilets for leaks

Put a little food coloring in your toilet tank. If, without flushing, the color begins to appear in the bowl within 30 minutes, you have a leak that should be repaired immediately. Most replacement parts are inexpensive and easy to install.

Boilerplate Removal

```
 7 Â
 8 [              Directions &amp;amp; Map       History        Bonnie &amp;amp; Clyde
 9 Photos &amp;amp; Events        Economic Development       Pay Water Bill]
10 [shapeimage_2_link_0][shapeimage_2_link_1][shapeimage_2_link_2]
11 [shapeimage_2_link_3][shapeimage_2_link_4][shapeimage_2_link_5]
12 [HOME        GOVERNMENT        SERVICES        ATTRACTIONS        FOOD &amp;amp;
13 LODGING        Contact][shapeimage_3_link_0][shapeimage_3_link_1]
14 [shapeimage_3_link_2][shapeimage_3_link_3][shapeimage_3_link_4]
15 [shapeimage_3_link_5]
16 Â
17 Â
18 Â
19 [Are You Water Aware?]
20 Water is a precious resource. It's important to use water wisely, particularly
21 during extended dry weather. By following these simple suggestions, you'll save
22 money on your water bill while conserving the supply we all depend on.
23 Check faucets and pipes for leaks
24
25 A small drip from a worn faucet washer can waste 20 gallons of water per day.
26 Larger leaks can waste hundreds of gallons.
27 Check your toilets for leaks
```

Boilerplate Removal




```
 7 Â
 8 [              Directions &amp;amp; Map       History        Bonnie &amp;amp; Clyde
 9 Photos &amp;amp; Events        Economic Development       Pay Water Bill]
10 [shapeimage_2_link_0][shapeimage_2_link_1][shapeimage_2_link_2]
11 [shapeimage_2_link_3][shapeimage_2_link_4][shapeimage_2_link_5]
12 [HOME        GOVERNMENT        SERVICES        ATTRACTIONS        FOOD &amp;amp;
13 LODGING        Contact][shapeimage_3_link_0][shapeimage_3_link_1]
14 [shapeimage_3_link_2][shapeimage_3_link_3][shapeimage_3_link_4]
15 [shapeimage_3_link_5]
16 Â
17 Â
18 Â
19 [Are You Water Aware?]
20 Water is a precious resource. It's important to use water wisely, particularly
21 during extended dry weather. By following these simple suggestions, you'll save
22 money on your water bill while conserving the supply we all depend on.
23 Check faucets and pipes for leaks
24
25 A small drip from a worn faucet washer can waste 20 gallons of water per day.
26 Larger leaks can waste hundreds of gallons.
27 Check your toilets for leaks
28
```

Boilerplate Removal

- Within each city, there is a lot of shared text
- If not removed, the text clusters into cities

- Solution: Compare each line in each document to every other line in every document of that city
- Count duplicates
- Remove a line if it is duplicated within a city above some threshold
- Fast & efficient implementation

The Pipeline

Data Collection → **Preprocessing** → **Analysis**

- Identify URLs
- Verify URLs (browser automation)
- Download websites
- Determine file type
- Convert to txt

- Remove punctuation, dates, etc.
- To lowercase
- Boilerplate removal
- Spellchecking
- Lemmatization (city & cities = city)

- Fightin' Words
- Structural topic model

Fightin' Words

Word (D)	z-Score (D)	Word (R)	z-Score (R)
said	87.42	request	69.24
proposal	71.24	member	68.50
fund	59.31	main	55.05
county	56.36	motion	54.92
asked	52.92	street	49.59
budget	49.63	councilor	47.26
stated	43.34	utility	45.38
ms	42.57	water	44.84
tax	42.50	goods	44.03
fort	41.31	rd	41.35
million	39.03	tree	41.15
division	38.26	site	40.33
revenue	35.53	ave	39.64
grants	35.21	plan	39.59
contract	34.24	amp	39.01
general	34.16	sewer	38.91
introduced	32.92	downtown	38.75
chair	32.45	use	38.75
brown	32.41	line	38.12
federal	32.22	pm	37.95
metropolitan	31.97	sign	36.60

Structural Topic Model

-0.026	-0.019	-0.018	-0.018	-0.018	-0.013
fort	propos	said	prosecutor	digest	fund
citi	author	ask	charg	introduc	grant
ordin	district	state	feloni	author	budget
approv	street	will	counti	counti	counti
purchas	public	chair	case	appoint	state
depart	control	propos	crime	board	feder
properti	amend	year	crimin	approv	depart
will	intersect	move	offic	district	appropri
resolut	counti	need	victim	fund	increas
contract	committe	council	sentenc	street	agenc

Structural Topic Model

0.024	0.019	0.019	0.016	0.016	0.016
motion	plan	inc	request	council	traffic
second	zone	electr	board	citi	amp
made	applic	build	member	ordin	vehicl
approv	properti	construct	servic	common	stop
mayor	approv	home	street	councilor	sign
present	sign	street	approv	amend	road
state	site	meridian	purchas	resolut	block
will	locat	servic	citi	adopt	signal
citi	commiss	west	move	wherea	street
council	file	main	good	approv	driver

Conclusion

- Manual vs. automatic analysis of government websites
- Our pipeline facilities scalable comparative analysis

- Partisanship affects website content:
- Democrats focus on raising and spending money
- Republicans focus on infrastructure and utilities