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.

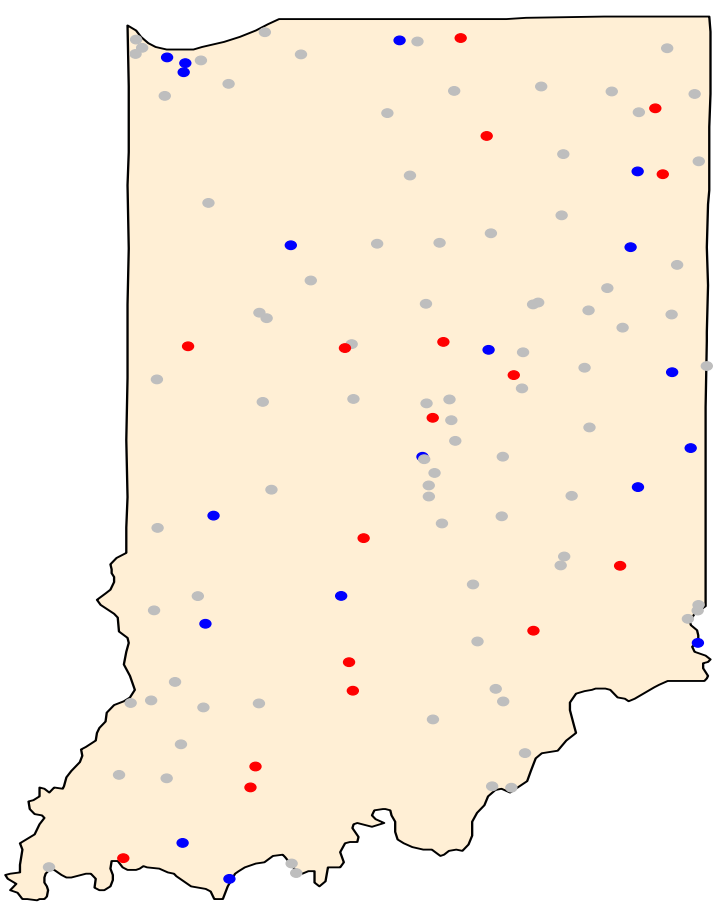
Check your toilets for leaks

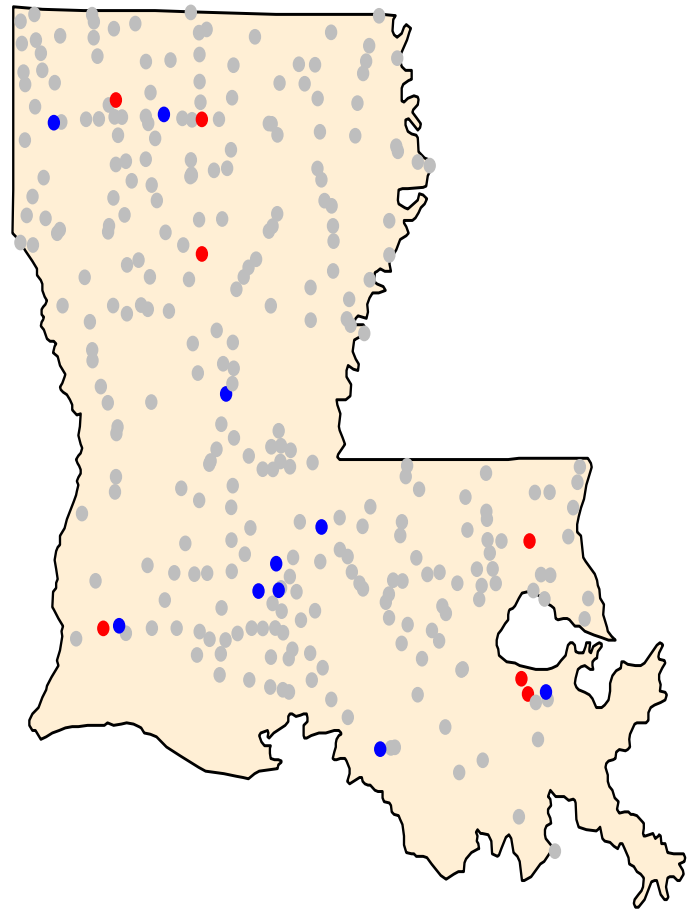
Put a little food coloring in your toilet tank. If, without flushing, the color begins to appear in the bowl within 30 minutes, you have a leak that should be repaired immediately. Most replacement parts are inexpensive and easy to install.

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.

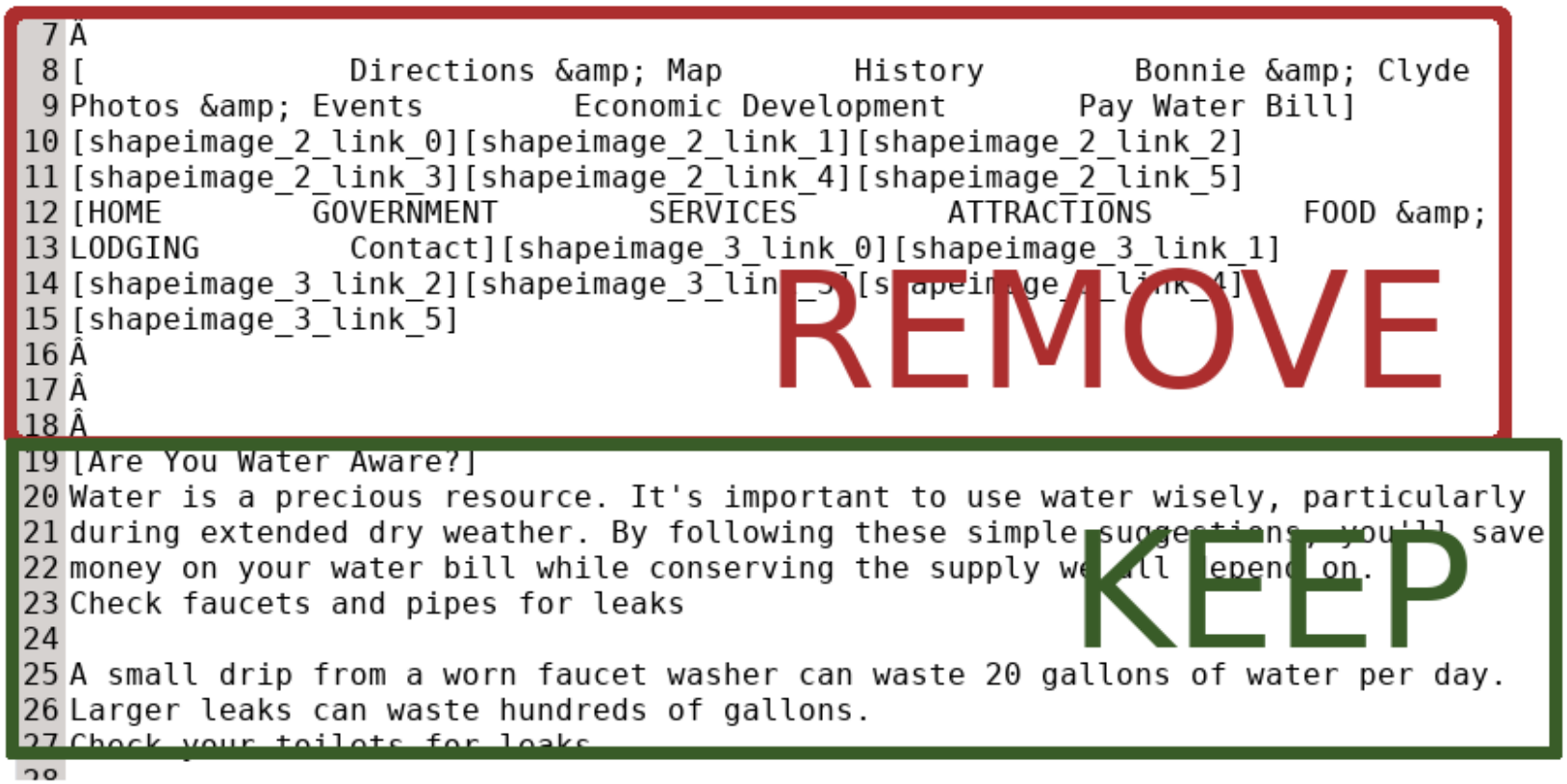
Check your toilets for leaks

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

Data Overview

	Democratic	Republican	Total
Cities	16	17	33
Documents	10868	6438	17306
Token types	20774	17947	21697
Token instances	6532383	2651876	9184259

Table: Indiana

	Democratic	Republican	Total
Cities	10	8	18
Documents	6636	1378	8014
Token types	16649	9234	16856
Token instances	3764877	355774	4120651

Table: Louisiana

Fightin' Words - Monroe et al. (2008)

- Goal: Split words based on one binary variable (party)
- Find the most Democratic and most Republican words
- Problem: Text mostly consists of meaningless words
- Solution: Set prior based on expected distribution of words in random text
- Method provides a z-Score indicating the degree to which a word is preferred by one party
- Very fast to compute

Fightin' Words (Indiana)

Word (D)	z-Score (D)	Word (R)	z-Score (R)
say	93.15	main	60.56
proposal	80.78	ave	58.11
fund	66.61	sewer	57.85
county	60.76	tree	53.82
budget	57.16	sign	52.42
ask	54.53	councilor	51.18
tax	52.95	utility	49.95
state	49.40	line	49.35
revenue	42.96	stream	49.03
division	42.25	street	47.47
grant	42.25	oral	46.87
million	40.21	member	45.96
contract	40.12	water	44.45
agency	38.15	motion	44.14
general	36.74	building	42.41
introduce	35.96	site	42.10
animal	34.54	flow	39.21
chair	34.19	lot	38.03
metropolitan	33.87	plat	37.84
support	33.78	zone	37.49
authorize	33.65	amp	37.24
federal	33.60	grease	37.21
cost	33.20	plan	36.98

Structural Topic Model - Roberts et al. (2014)

- Topic models find clusters of semantically related words in texts –> i.e. topics
- As a clustering method, LDA doesn't explicitly allow for covariates (although post-hoc comparison is possible)
- The structural topic model extends this approach to explicitly include (multiple) covariates
- Our covariates: Party + City Population

Structural Topic Model (Indiana)

-0.027	-0.022	-0.016	-0.011	-0.011	-0.01
city	school	downtown	city	trash	housing
ordinance	community	business	department	city	property
approve	program	project	mayor	waste	program
resolution	student	city	police	day	fund
property	education	development	officer	recycle	home
purchase	university	new	public	street	city
area	national	center	citizen	collection	project
department	award	economic	work	resident	neighborhood
contract	high	company	safety	recycling	grant
service	year	community	resident	snow	unit

Table: Top Democratic topics and words

Structural Topic Model (Indiana)

0.021	0.019	0.017	0.017	0.013	0.012
foot	team	ave	request	amp	building
sign	game	inc	board	traffic	historic
use	play	cross	member	stop	build
lot	league	creek	service	vehicle	material
building	camp	construction	street	block	preservation
zone	class	blvd	approve	sign	wall
area	age	park	city	airport	roof
district	must	lake	purchase	ave	window
parking	child	hill	move	theft	floor
residential	participant	ridge	good	signal	new

Table: Top Republican topics and words

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