The effects of transitions of power on the contents of municipal government websites

Markus Neumann
Fridolin Linder
Bruce Desmarais

The Pennsylvania State University

January 6, 2018

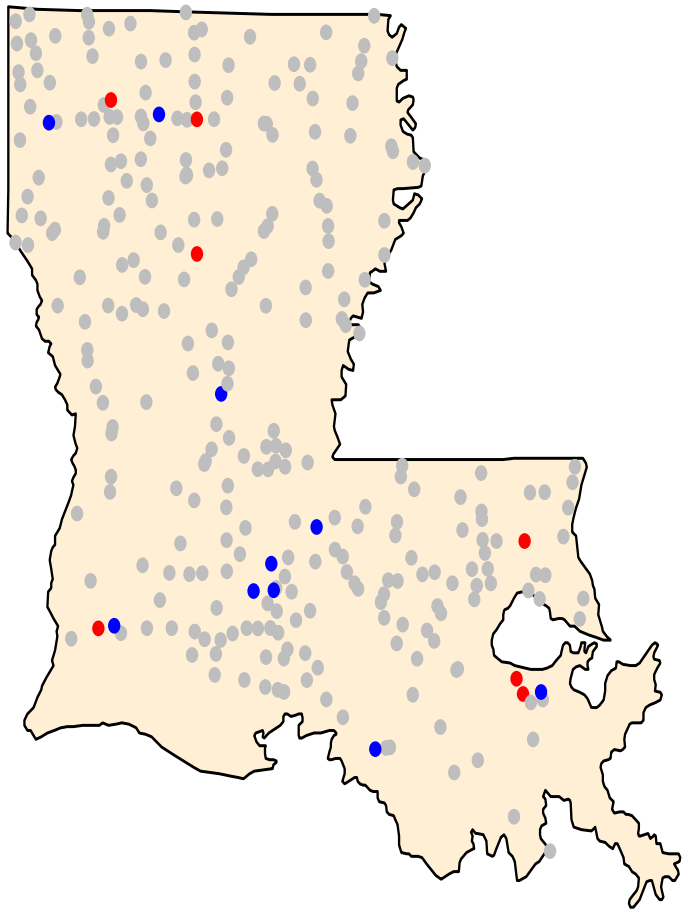
Government Websites

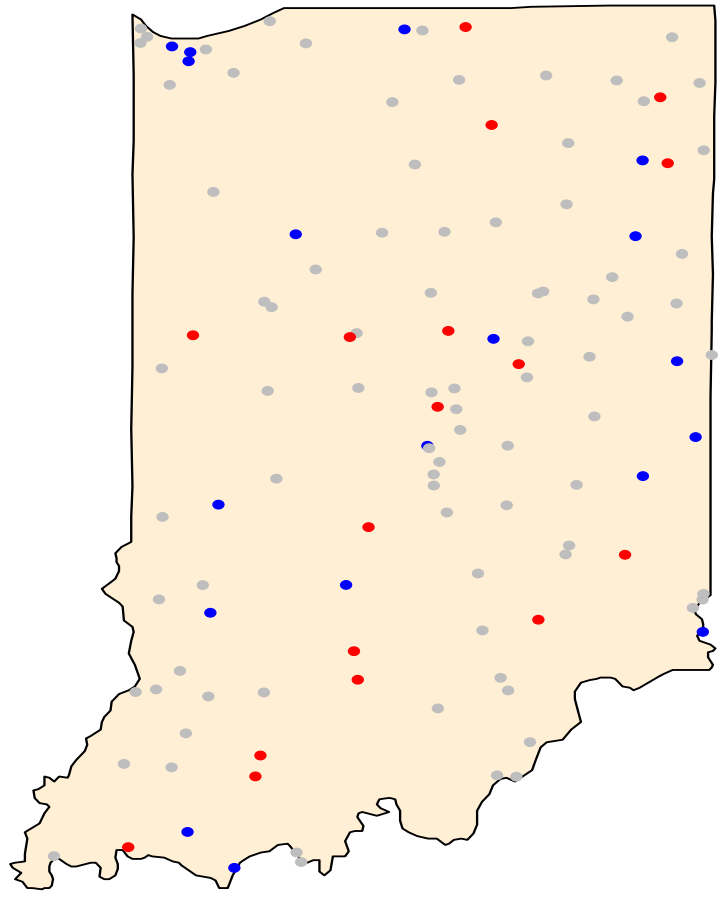
- Content of government websites is an important source of information & transparency
- After coming into power, the Trump administration has made some controversial changes to the websites of federal agencies
- Website content is political
- Partisanship of city government is expected to have an effect

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

Maps





The Pipeline

Data Collection –> **Preprocessing** –> **Analysis**

- Scraping URLs
- Browser automation to verify URLs
- Download sites with wget
- Determine file type
- Convert to txt

- Remove punctuation, dates, etc.
- To lowercase
- Duplicate content removal
- Spellchecking
- Lemmatization

- Hierarchical clustering
- Fightin' Words
- Topic models
  - LDA
  - STM

The Pipeline

Data Collection –> **Preprocessing** –> **Analysis**

- Scraping URLs
- Browser automation to verify URLs
- Download sites with wget
- Determine file type
- Convert to txt

- Remove punctuation, dates, etc.
- To lowercase
- **Duplicate content removal**
- Spellchecking
- Lemmatization

- Hierarchical clustering
- Fightin' Words
- Topic models
  - LDA
  - STM

Duplicate Content 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.

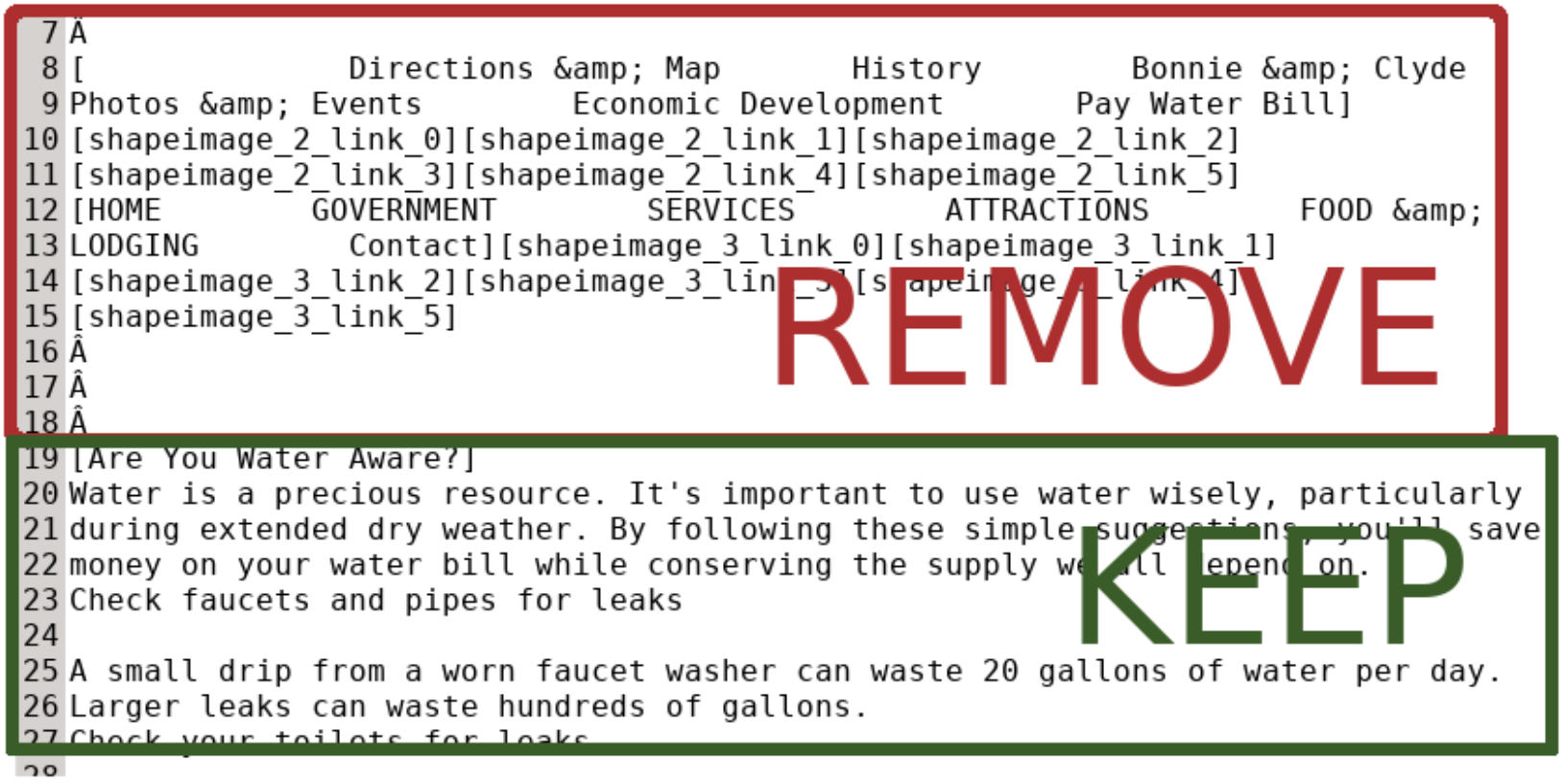
Check your toilets for leaks

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.

Duplicate Content 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
```

Duplicate Content Removal



Duplicate Content Removal

- Within each city, there is a lot of shared text
  - Boilerplate
  - Website elements
- 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
- Hash tables for computational efficiency

The Pipeline

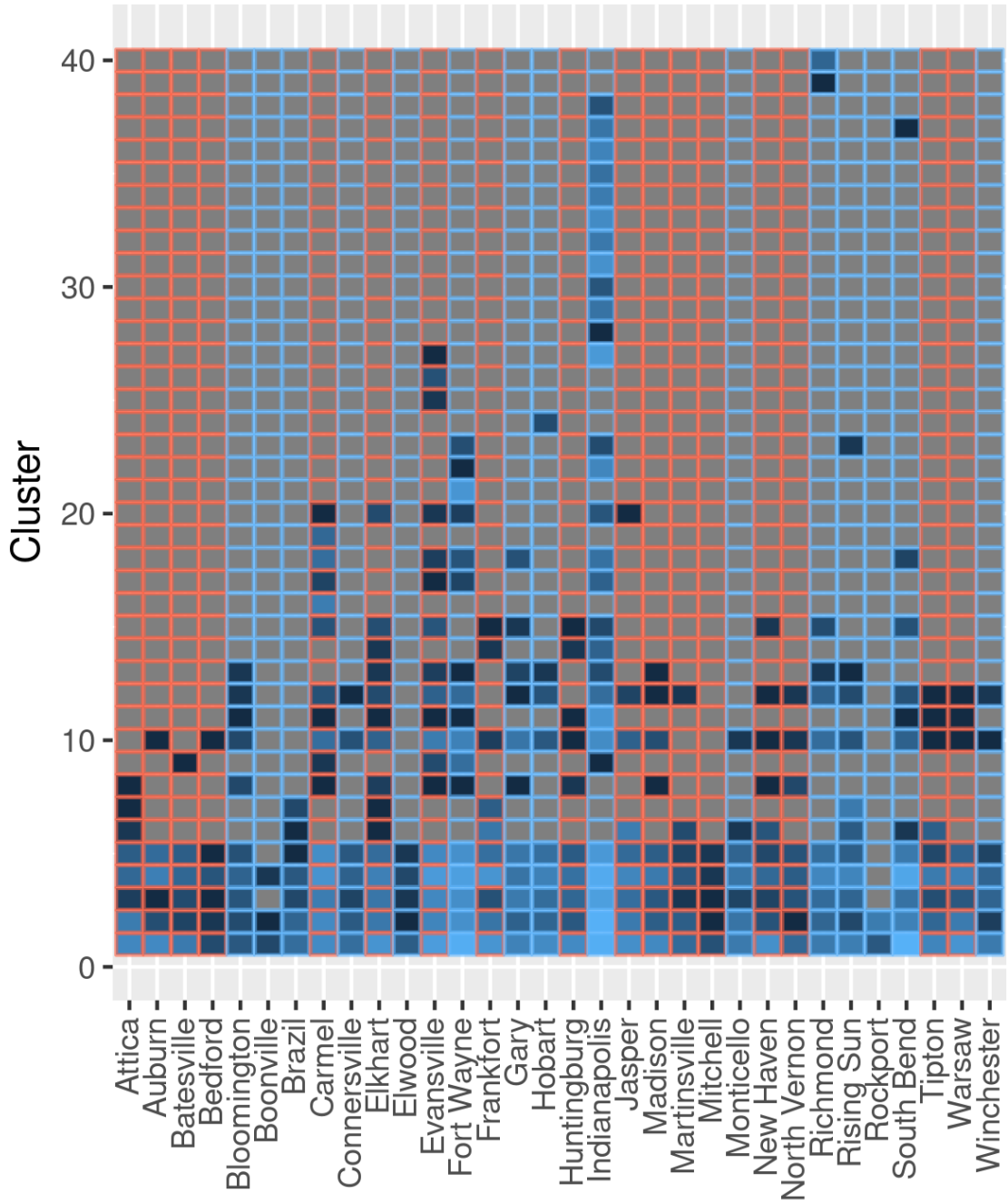
Data Collection –> **Preprocessing** –> **Analysis**

- Scraping URLs
- Browser automation to verify URLs
- Download sites with wget
- Determine file type
- Convert to txt

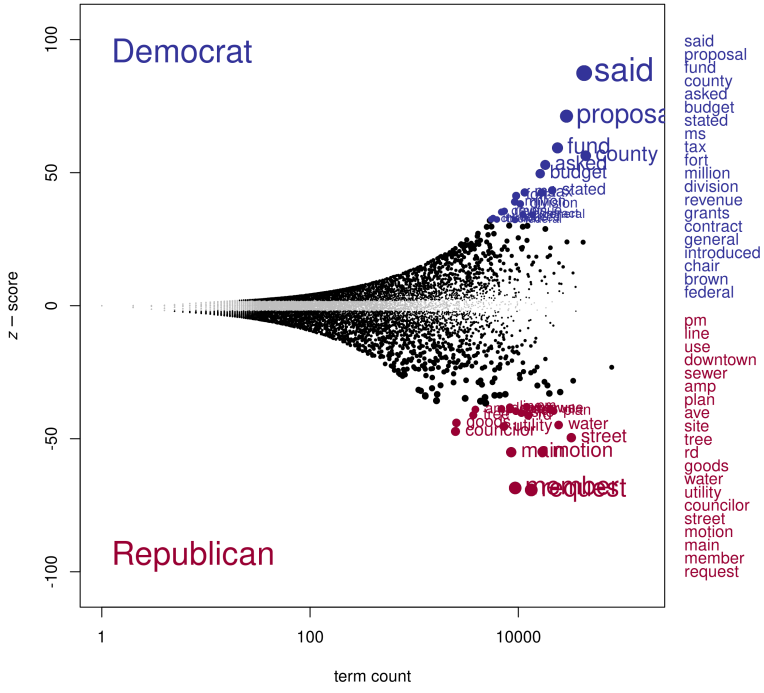
- Remove punctuation, dates, etc.
- To lowercase
- Duplicate content removal
- Spellchecking
- Lemmatization

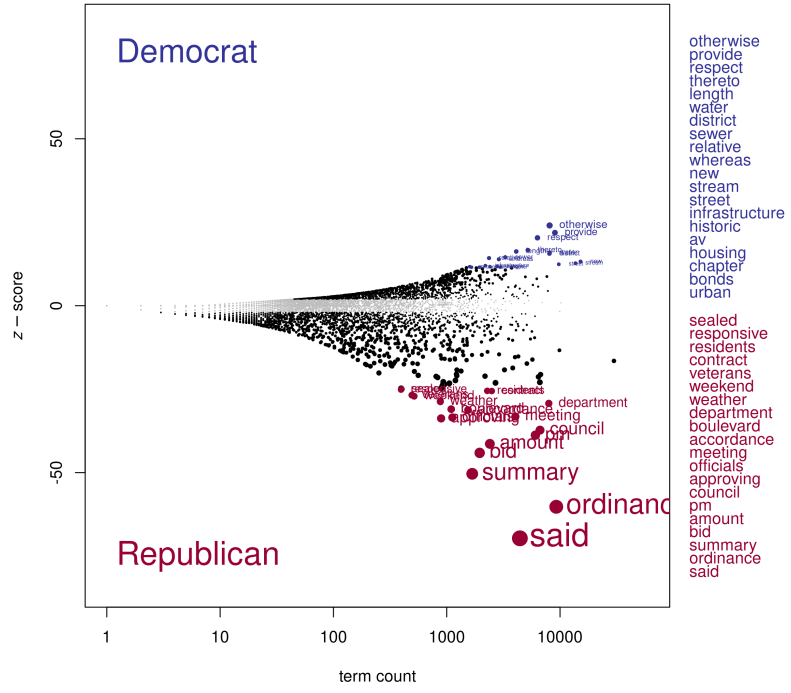
- Hierarchical clustering
- Fightin' Words
- Topic models
  - LDA
  - STM

Hierarchical Clustering



Fightin' Words





Latent Dirichlet Allocation

Word (D)	Instances (D)	Word (R)	Instances (R)
city	42493	will	53761
said	40480	city	36210
county	39209	street	21207
proposal	29019	board	19496
public	27070	water	18637
council	23492	plan	18241
shall	23162	public	14327
department	22926	use	13233
services	22703	information	13062
fund	21661	development	12916
will	20697	department	11554
new	19000	area	11270
stated	18794	shall	11247
project	18538	fire	10861
property	18378	can	10748
budget	16631	must	10633
community	16236	park	10493
asked	16231	building	10356
tax	14549	motion	10168
board	14363	ordinance	9625
state	13964	request	9512

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

- Hierarchical clustering only creates 'topics' if content is VERY similar
- Fightin words is extremely effective at forcing the text into two categories, but also somewhat limited
- LDA is sensitive to preprocessing, but the most flexible
- STM works too, but also very sensitive to duplicate content

- Democrats focus on raising and spending money, Republicans on infrastructure and basic utilities