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

Presented at SPSA

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

Presented at SPSA

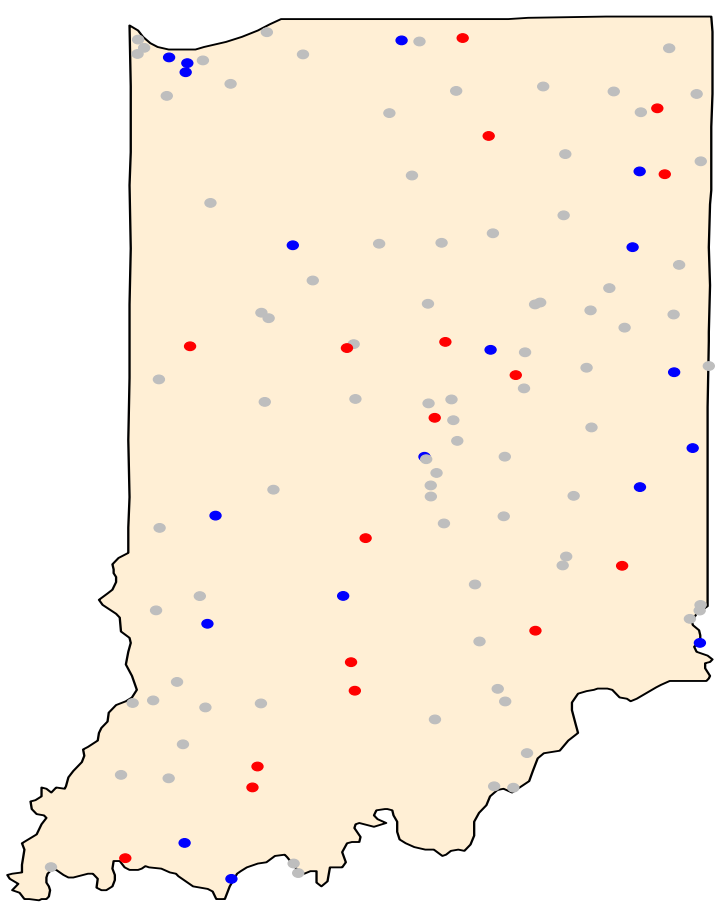
	Democratic	Republican	Total
Cities	16	17	33
Documents	10868	6438	17306
Token types	20774	17947	21697
Token instances	6532383	2651876	9184259

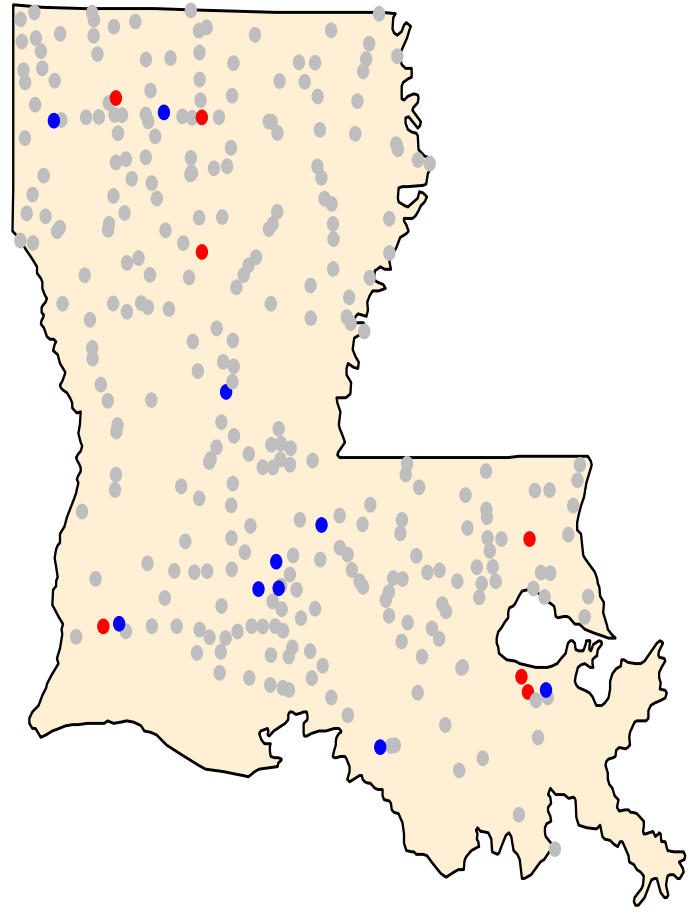
Table: Indiana

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

Table: Louisiana

Presented at SPSA





Presented at SPSA

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

Presented at SPSA

-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

Presented at SPSA

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

SPSA feedback

- overall quite positive
- there seems to be some demand in publican administration for this kind of research
- threshold of ten for duplicates
- the usual concerns with bag-of-words
- describe methods more clearly
- **“Does your method improve the external validity so greatly that the internal validity becomes less of a concern?”**
- comparison with non-partisan cities/websites
- city covariates

Planned covariates

- population
- GDP per capita
- percent non-white
- City area
- democratic vote share/magnitude of victory
- log median house price
- (most of these are from Einstein & Glick 2015)

Since SPSA - ground truth test

- party manifestos (didn't work - not enough data)
- mayors' campaign websites (LA/IN - didn't work - not much data, and strange results)
- mayors' campaign websites (top 100 cities - worked somewhat)

	dem.groundtruth	rep.groundtruth	dem.cities
dem.groundtruth	1, 1	0.807, 0.896	0.714, 0.729
rep.groundtruth	0.807, 0.896	1, 1	0.647, 0.697
dem.cities	0.714, 0.729	0.647, 0.697	1, 1
rep.cities	0.68, 0.698	0.641, 0.693	0.937, 0.944

Table: Ground truth test, comparing campaign websites of mayors of the 100 largest cities in the US and cities in Indiana and Louisiana. The values are bootstrapped confidence bounds for cosine similarities between concatenated document collections.

Since SPSA - extending the sample

- New York
- big cities
- Washington
- Oregon (unsuccessful)
- extended LA/IN (55 -> 165)
- 314 cities total, 230 downloaded so far

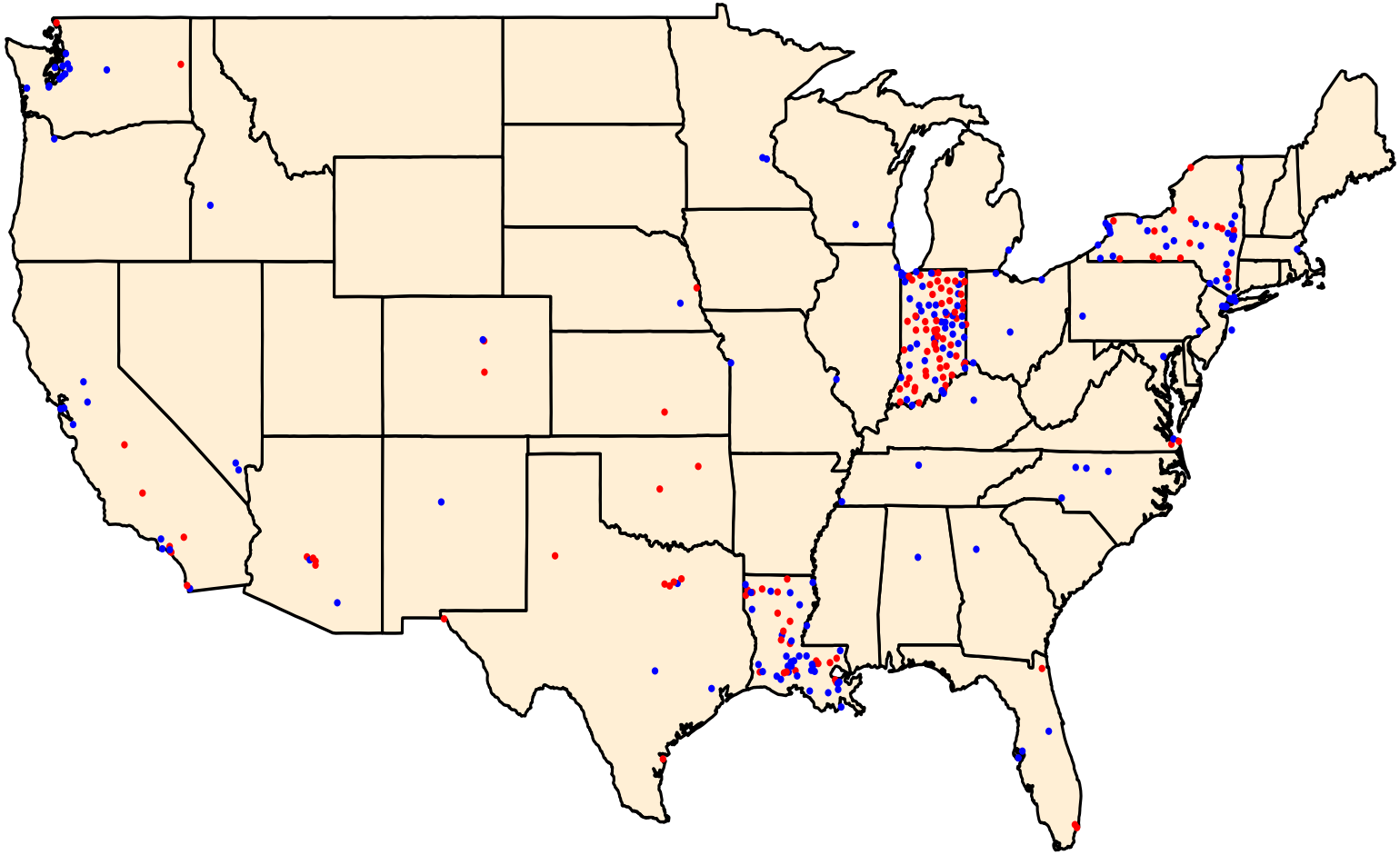
wget and www

- www.townoflockport.com/ – doesn't work
- http://townoflockport.com/ – works
- inconsistent across websites
- solution: check every website with Selenium, record the url it redirects to
- currently re-scraping the 84 websites still missing

Next steps: compression

- Since, so far, something has always gone wrong when adjusting file endings and converting files to text, I've always made zipped backups so far
- compression of millions of files of more than 1TB
  - time
  - some paths are too long
  - some filenames have non-ascii characters

Next steps: map



Next steps: covariates

- I already downloaded the above some time ago, but only for LA and IN
- this was already a little convoluted, I'll have to automate it for it to work with all states
- there is a problem with the census data:

Next steps: covariates

- I already downloaded the above some time ago, but only for LA and IN
- this was already a little convoluted, I'll have to automate it for it to work with all states
- there is a problem with the census data:

61	36	101	0	18256	0	0	F	Corning city	New York	11183
71	36	101	18256	18256	0	1	A	Corning city	New York	11183
61	36	101	0	18267	0	0	A	Corning town	New York	6270
71	36	101	62061	18267	0	1	A	Riverside village	New York	497
71	36	101	68847	18267	0	1	A	South Corning village	New York	1145
71	36	101	99990	18267	0	1	F	Balance of Corning town	New York	4628

- Not a priority right now, need to finish all the document-related stuff first

Next steps: optimize code

- preprocessing already took 4-5 hours for 25000 documents
- we currently have 1.3 million; this will get reduced a bit, but it will still be a huge increase
- the functions that aren't already vectorized are currently "sapply'd"
- with this much data, parallelization is probably worth it (currently only hashtables, spellchecking and some other more computationally intensive stuff is parallelized)
- MAYBE it would be worth it to translate everything to quanteda - but unfortunately that won't work with the hashtables stuff
- fightin' words is pretty fast, but stm also took hours already

Next steps: structural topic model

- currently we use separate models for IN, LA
- state fixed effects
- package has problems with overly complex models
- maybe use stm only for IN, LA, NY, WA, CA