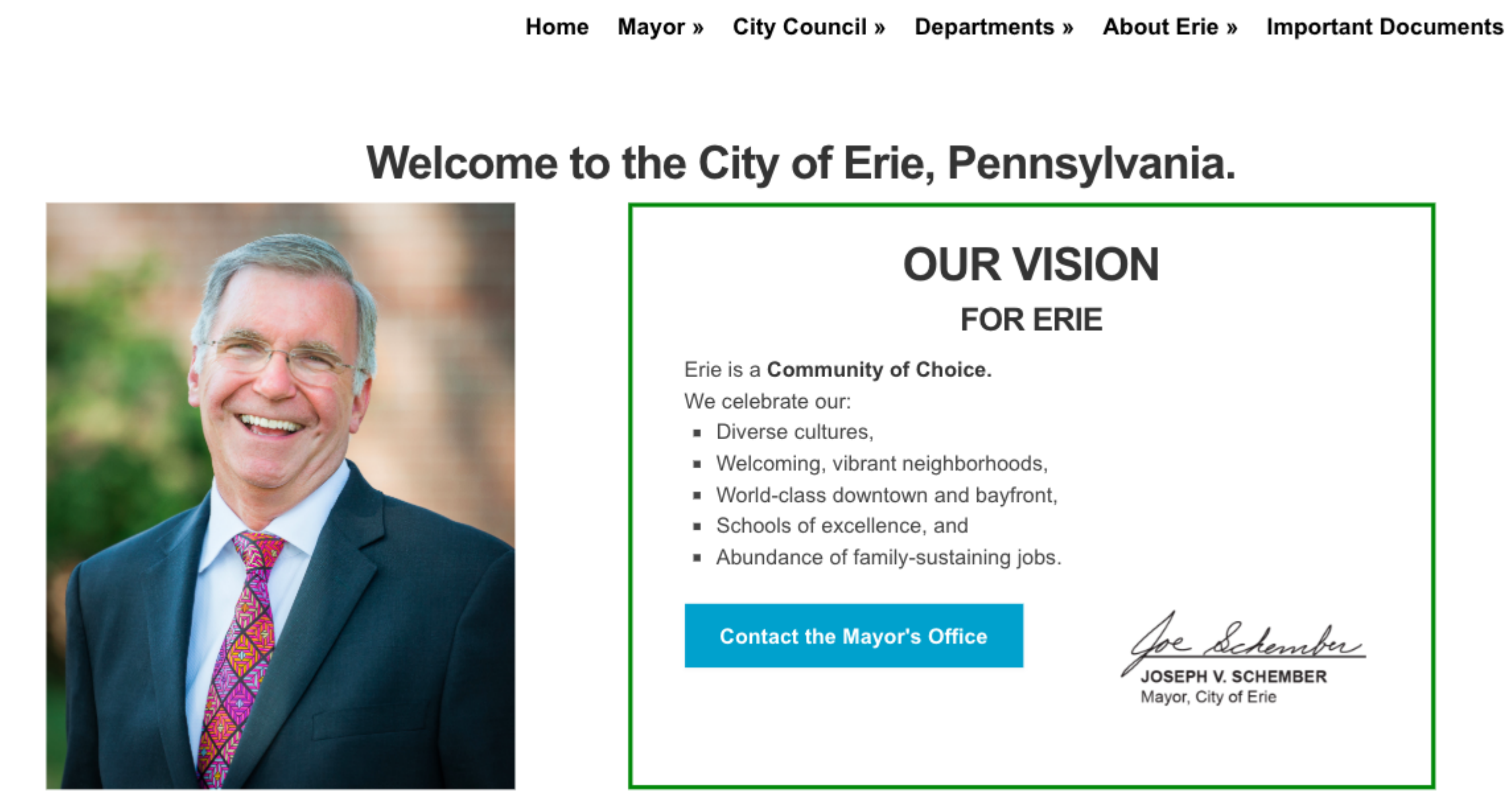
Government websites as data: A methodological pipeline with application to the websites of municipalities in the United States

Markus Neumann (mvn5218@psu.edu), Fridolin Linder (flinder@gmail.com), Bruce Desmarais (brucedesmarais@psu.edu)
Department of Political Science - Penn State University

Overview



Home Mayor » City Council » Departments » About Erie » Important Documents

Welcome to the City of Erie, Pennsylvania.

OUR VISION
FOR ERIE

Erie is a **Community of Choice.**
We celebrate our:

- Diverse cultures,
- Welcoming, vibrant neighborhoods,
- World-class downtown and bayfront,
- Schools of excellence, and
- Abundance of family-sustaining jobs.

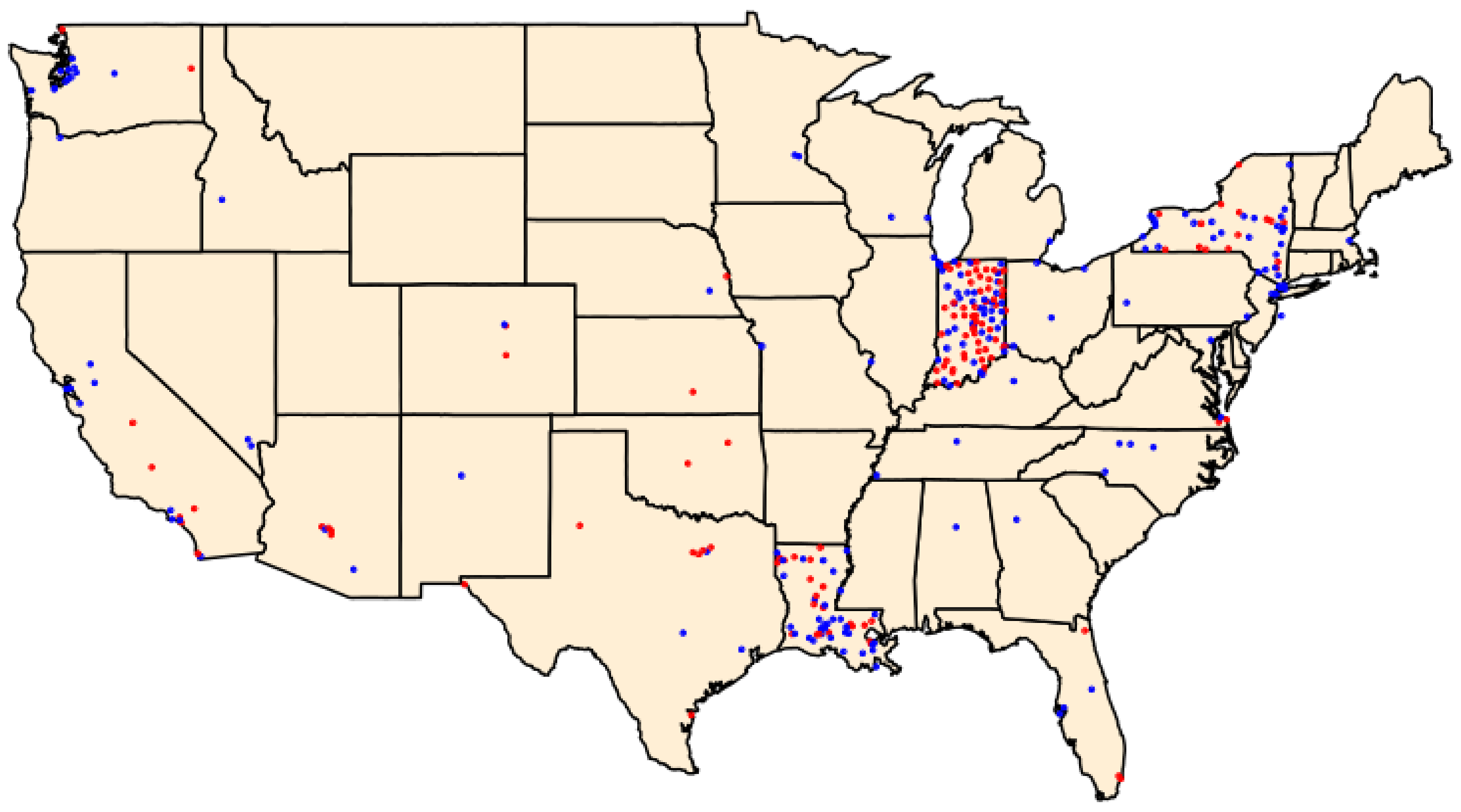
Contact the Mayor's Office

Joe Schember
JOSEPH V. SCHEMBER
Mayor, City of Erie

- Government websites contain important information
- Research has largely relied on manual coding
- Our contribution: pipeline for automated analysis
- Application: websites of partisan municipalities

Data Collection

- City URLs are scraped from Wikipedia and the GSA



Site to Text Conversion

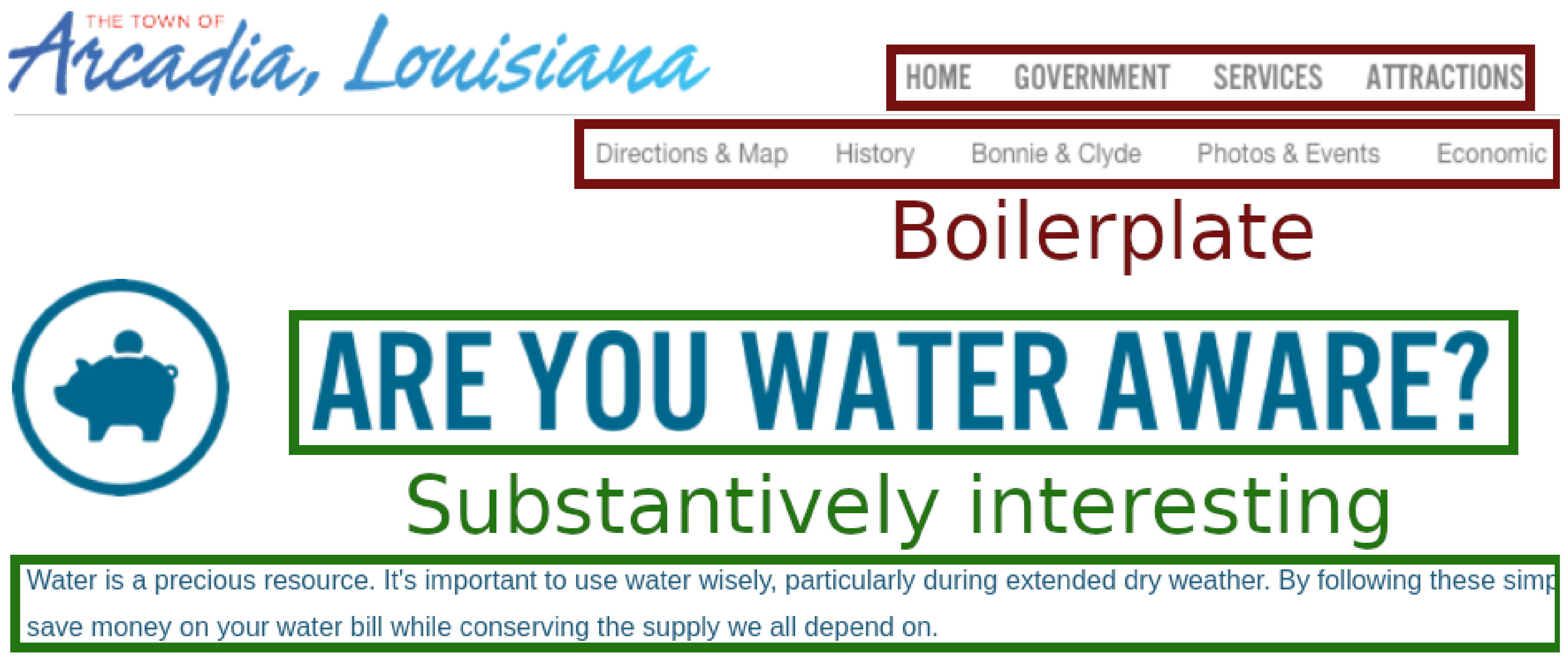
Filetype	Occurances Before	Occurances After
html	211682	887362
pdf	464842	638802
jpg	0	36958
xml	0	29638
Other	162681	9475
ics	435	8950
png	0	8863
doc	6972	8430
txt	317	6025
	793990	5234
docx	3137	4319
TOTAL	1644056	1644056

Table: Number of files per type, before and after detecing them via their magic number. The table shows that a lot of files originally have the wrong type, and that converting them correctly has a large impact on how many of them end up being usable.

- The file endings from city websites are sometimes wrong
- We use file signatures to identify the correct type
- Preprocessing: lowercase; removal of punctuation, numbers, dates, etc.; removal of non-English words; lemmatization

Boilerplate Removal

- Websites contain a lot repetitive and uninformative content



THE TOWN OF Arcadia, Louisiana

HOME GOVERNMENT SERVICES ATTRACTIONS

Directions & Map History Bonnie & Clyde Photos & Events Economic

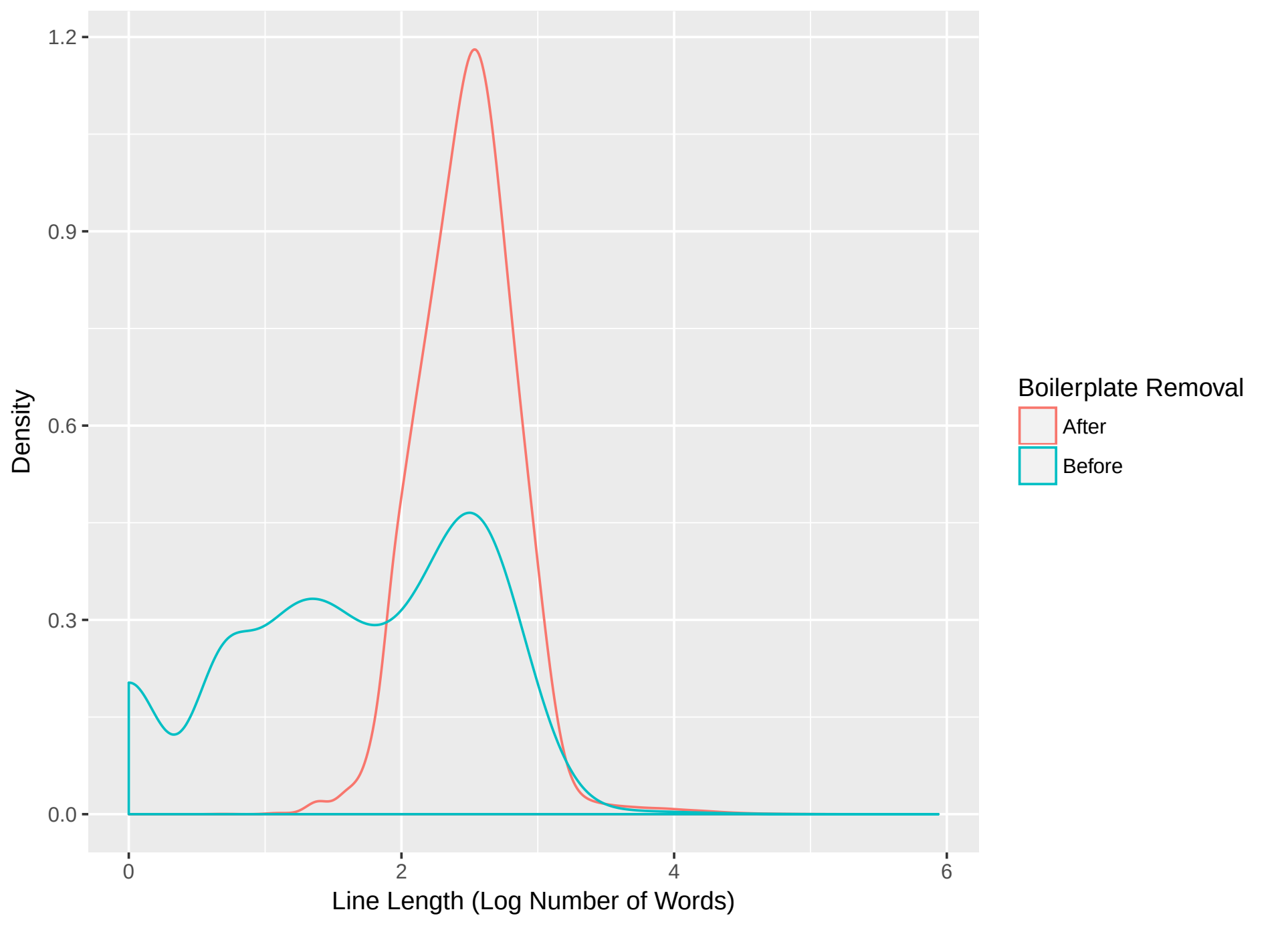
Boilerplate

ARE YOU WATER AWARE?

Substantively interesting

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

- A classifier (random forest) is trained to identify and remove boilerplate lines



	Value
Percent Correctly Predicted	0.87
Precision	0.87
Recall	0.91
F1-Score	0.89

Table: Performance metrics for random forest boilerplate classifier.

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
42	breastfeed	vaccine	infection	symptom	asthma	mosquito	2497
17	alarm	disaster	fire	rescue	preparedness	evacuation	989
53	drinking	wastewater	water	pipeline	pump	disinfection	461
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
1	storm	runoff	infiltration	discharge	drainage	drain	516
38	youth	student	parent	teacher	immigrant	literacy	714
59	sampling	sample	analytical	concentration	hydrocarbon	toxicity	1241
45	premise	licensee	violation	license	permit	inspection	509
60	exhaust	fugitive	aircraft	airport	aviation	diesel	731
4	facade	awning	porch	roof	balcony	exterior	1108
15	complainant	defendant	allegation	complaint	allege	discrimination	1384
14	yes	agency	federal	recipient	compliance	entity	205
25	cannabis	marijuana	senate	dispensary	ballot	cultivation	1188
47	employee	overtime	sick	wage	grievance	bargaining	511
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

- Republican cities feature information on basic utilities and protection from natural disasters
- Democratic cities feature information on policy deliberation, crime control and city budgeting
- Politics at the municipal level is not entirely non-partisan
- We plan to implement the pipeline in an R package