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

Markus Neumann Fridolin Linder Bruce Desmarais

December 29, 2017

Abstract

A local government's website is a standard and general source of information for citizens and other community stakeholders. Accordingly, government websites have become prominent sources of data for a variety of research agendas in public administration, public policy, and political science. Existing research has relied on manual methods of website data collection of processing. Reliance on manual collection and processing limits the scale and scope of website content analysis. We develop a methodological pipeline that researchers can followed in order to gather, process, and analyze website content with established text analysis techniques. First, for the acquisition of website data, we cover approaches to automated scraping methods. Second, pre-processing is a particularly vital step in text analysis, but when websites are concerned, additional measures need to be taken in order to guard against potential sources of bias. We propose a new method for dealing with the kind of duplicated content that is commonly found in government websites. Finally, we illustrate methods of text analysis using automatically gathered and pre-processed website content. We illustrate our methodological pipeline through a new and innovative dataset—the websites of local governments in Indiana and Louisiana. We build upon recent research that analyzes how change and variation in the partisan control of government relates to content made available on the government's website. We explore the association between mayoral partisanship and the content of city websites.

1 Introduction

Local governments convey voluminous information about all aspects of their policymaking, policy implementation, and public deliberation, via their official websites. The vital role of official websites in connecting the government and the governed has motivated a wave of research on the contents of government websites (e.g., Grimmelikhuijsen 2010; Wang, Bretschneider and Gant 2005; Osman, Anouze, Irani, Al-Ayoubi, Lee, Balcı, Medeni and Weerakkody 2014). Despite the potential for automated scraping of website contents, the conventional approach to data collection in projects focused on government websites involves manual content extraction from each website in the dataset. Though highly accurate, the manual approach to data collection is costly, and cannot be scaled to capture even a fraction of the volume of content available on government websites. In this paper we present a methodological pipeline that can be used to automatically scrape government websites in order to build datasets that can be used for text analysis. We provide an illustrative application in which we explore the ways in which the textual contents on city government websites in Indiana and Louisiana correlate with the partisanship of the city mayor.

Though there exists a variety of software tools that are designed to automatically scrape all of the files available at a website (Glez-Peña, Lourenço, López-Fernández, Reboiro-Jato and Fdez-Riverola 2013), raw website downloads have to be processed significantly before the files are adequately prepared for text analysis. We describe and provide solutions to two central challenges in automatically gathering and analyzing website textual contents. First, plain text must be extracted from the files. This involves purging the files of syntax in HTML and other programming languages, and discarding any other character encoding errors that result from reading the files. This challenge would arise in any context in which researchers sought to study the textual contents of websites, and is not unique to comparative analysis of government websites. The second challenge we address in our methodological pipeline is, however, specific to the research objective of comparing websites on the basis of a common lexicon. For any two governments, the textual signatures that most dramatically differentiate the textual contents of their websites consist of what we can call "boilerplate" text—header, footer, or other titling text that is designed to identify the website as being associated with a specific government entity (e.g., "Welcome to the city of Santa Cruz", "The City of Los Angeles welcomes you"). This boilerplate text is replicated across many files that are associated with a government's website, but it provides little information regarding the form and/or function of the government. The second methodological innovation we offer in our pipeline is designed to minimize the impact of this boilerplate text on the comparative analysis of government website content.

Government websites provide information about how public policies shape the lives of local residents, and how local residents can engage with government to shape public policy. As such, government websites reflect both the results of, and inputs to, the political leadership in the city. In our illustrative application we explore the ways in which the contents of city government websites differ on the basis of the partisanship of the city's elected executive. A substantial body of research has found that the partisanship of the mayor affects city governance along multiple dimensions, including city budget priorities (de Benedictis-Kessner and Warshaw 2016), policies affecting inequality in cities (Einstein and Glick 2016), and framing of criminal justice policy (Marion and Oliver 2013). Furthermore, recent media coverage of changes to government websites that follow transitions in party control suggest that changes in web content are salient government actions, as perceived by the general public (Sharfstein 2017; Kirby 2017; Duarte 2017) . We study whether significant differences between city governments based on mayoral partisanship are reflected in the contents of city websites.

2 Background: The Study of Government Websites

Grimmelikhuijsen (2010) run an experiment in which citizens are exposed to randomly selected levels of information about local government council minutes. They find a negative relationship between the information level and perceptions of competence in the local government. This raises an interesting question regarding whether citizens or more likely to participate when they perceive competence or when they perceive incompetence.

Wang, Bretschneider and Gant (2005) present a widely cited 'model' for evaluating the accessibility of information on government websites. This is an important paper with which we should be familiar at a very detailed level as we use archived web content to assess the volume/accessibility of information provided by local governments.

Osman et al. (2014) is less relevant, but they develop a multi-item measure to predict the level of citizen satisfaction with e-government services.

Grimmelikhuijsen and Welch (2012) conduct an enormously relevant study. Insofar as we analyze what predicts openness of government websites, we will be replicating and building upon this study. They focus on Dutch municipal websites, and their approach is fairly limited in scope and highly manual (which we can compliment). For example, one of the dependent variables "Decision-making transparency," is measured "using a discrete (1/0) indicator for whether the underlying principles or reasons for local air pollution policies were given on the Web site."

Why does the content of city websites matter? According to Mayhew (1974), politicians engage in advertising, credit claiming and position taking in order to get re-elected. Official city websites allow mayors to do all three. Their offices frequently take a prominent position on the frontpage, and many websites also feature a picture of the candidate. In local politics, where campaign funds are low, this lends the incumbent a crucial advantage in becoming more well-known among her constituents. Furthermore, municipal politics gives incumbents clear and tangible achievements they can point to, such as completed infrastructure projects, the acquisition of federal or state funding, or the hosting of city-wide events. City websites present an opportunity for local officials to brandish these accomplishments. Finally, they also give mayors a platform from which they can advertise their political beliefs. On municipal websites, this may not manifest in the form of brazen partisanship, but more subtle avenues are available. As noted by Einstein and Glick (2016), there are stark differences in the spending preferences of Democratic and Republican mayors. City websites can then be used to communicate the stance of a mayor on social or economic programs. Another advantage of websites with regard to communication is that unlike direct social interactions, officials have full control over them.

In addition to the use of city websites for the politicians that control them, variance in content also matters with regard to the people who visit them. Local residents likely rely on city websites to get news about events, hot-button political issues specific to their city, contact city officials or find out addresses or opening hours of city institutions. Visitors use city websites to look up local attractions, which are often described in great detail. Similarly, prospective residents looking to move, might rely on city websites to inform their decision on whether to relocate there. An inviting website emphasizing the city's receptiveness to new residents might make a real difference here. Finally, city websites frequently feature sections on business, but there is a lot of variance in this area: Some emphasize economic development, properties, or transportation, whereas others focus on undeveloped land and other business opportunities. Differences in websites likely say something about a city's economic profile, with potential repercussions for the political realm.

The literature making use of scraped websites clusters into a number of categories. One, and most pertinent to our own endeavors, the e-governance literature which discusses the online presence of governments from a usability and public service point of view. For the most part, research in this category develops a classification scheme to rate websites in terms of accessibility, ease-of-use and function, and then hand-codes a set of websites according to these criteria (Urban 2002; Armstrong 2011; Feeney and Brown 2017). As an example, Grimmelikhuijsen and Welch (2012) study local government websites with the goal of uncovering how they aid the goal of transparency. To this end, they analyze a set of Dutch municipalities in which air quality had deteriorated. The authors test whether local governments provide citizens with information about potential complications and solutions associated with this issue. Like most e-government studies however, this publication does not make any use of automated text analysis.

Websites have also played a major role in the field of media studies, as scholars have scraped and analyzed the online presence of newspapers, as well as the more diffuse world of online political blogs (Adamic and Glance 2005; Gentzkow and Shapiro 2010). Lin, Bagrow and Lazer (2011) provide a good example for a study which makes extensive use of automated content analysis - a necessity arising from its dataset of 66830 blog posts and 57221 online news articles. The authors estimate the political slant of these entities by counting the frequencies with which politicians of either side are mentioned and determine that blogs are generally more biased. Unfortunately for us, the authors don't go into the details of their text analysis, and offer no information on the acquisition and pre-processing of the data.

Another well-known example fitting into this area of study is the set of studies conducted by King et al. (King, Pan and Roberts 2013, 2014, 2017), in which the authors study censorship by the country's government on its lively blogosphere. However, the authors also provide no information on how their data was collected "our extensive engineering effort, which we do not detail here for obvious reasons [...]".

The websites of politicians and their parties have also fallen under scholarly scrutiny. Researchers have found that in order to identify the constituencies, motives and modes of communication of these actors, their websites can be very illuminating sources of information (Druckman, Kifer and Parkin 2009; Druckman, Hennessy, Kifer and Parkin 2010; Cryer 2017; Esterling, Lazer and Neblo 2011; Esterling and Neblo 2011; Norris 2003; Therriault 2010). Druckman, Kifer and Parkin (2009); Druckman et al. (2010) rely on the Mational Journal to find the websites, then hand-coded them. Cryer (2017) provides fairly little information, but does mention the fact that she relied on Archive-it, the webservice of the Internet Archive we discussed recently. Esterling, Lazer and Neblo (2011); Esterling and Neblo (2011) rely on hand-coded data by the Congressional Management Foundation, a nonprofit organization which aims to assist Congress. Therriault (2010) (a working paper) actually portends to use automated text analysis, and also has the most extensive overview of the associated methodology. However, the division of the website into sections (home page, topics, issues, details) is done by hand, and the actual analysis is incomplete. The author acquired the websites from the Library of Congress (which only collected them from legislators who actually consented, and Therriault notes that this causes nonrandom missingness).

Importantly for us, research analyzing and improving the scraping, pre-processing and analysis methods of this literature is scarce. Eschenfelder, Beachboard, McClure and Wyman (1997) provide something of an overview of how how federal websites should be assessed from an e-governance point of view, but they largely focus on the substantive criteria that should be fulfilled, rather than the technical aspects of website acquisition and analysis.

3 Running Application: Party Differences in Municipal Websites

Names	Year	Journal	Findings	Important?
Benedictis-Kessner, Justin De Warshaw, Christopher	2016	JOP*	Regression discontinuity design. Democratic mayors spend more (but it is unclear on what, not the typical Democratic issue-areas), issue more debt, pay more interest	Yes
Caughey, Devin Warshaw, Christopher Xu, Yiqing	2015	Working Paper	Regression discontinuity design. Partisan composition of state governments affects state policy liberalism (composite index for the areas of social welfare, taxation, labor, civil rights, womens rights, moral legislation, family planning, environment).	Somewhat
Einstein, Katherine Levine Kogan, Vladimir	2015	Urban Affairs Review	Cities with more Democratic citizens spend more; more progressive (rather than regressive) forms of taxation; pursue intergov. aid more; spend more on police, fire, parks & recreation	Somewhat
Einstein, Katherine Levine Glick, David M.	2015	Working Paper	Survey of 72 mayors. Unlike Republican mayors, roughly half of Democrats seem to agree that cities should aim to reduce inequality. Democratic mayors also seem to favor redistribution to accomplish that goal.	Somewhat
Kiewiet, D Roderick Mccubbins, Mathew D	2014	Annual Review	City budgets have been severeley constrained since the Great Recession. Spending has thus decreased in general. Lack of funds means that there is not much discretion for partisanship.	Somewhat
Tausanovitch, Chris Warshaw, Christopher	2014	APSR*	Cities are responsive (taxes, expenditures, regressiveness of taxation) to citizens' conservatism/liberalism. Partisan elections do not make cities more or less responsive.	Yes
Guillamón, Ma Dolores Bastida, Francisco Benito, Bernardino	2013	European Journal of Law and Economics	Police spending in Spain. Conservative parties spend more on police. Spending is higher before elections. Also contains a useful overview of the literature.	Yes

Names	Year	Journal	Findings	Important?
Gerber, Elisabeth R.	2013	Cityscape	Partisanship of both citizens and elected city officials separately affect climate policy.	Yes
Solé-Ollé, Albert Viladecans-Marsal, Elisabet	2013	Journal of Urban Economics	Spanish cities. The authors "employ a regression discontinuity design to document that cities controlled by left-wing parties convert much less land from rural to urban uses than is the case in similar cities con- trolled by the right". Partisanship might also affect housing construction and price growth.	Yes
Gerber, Elisabeth R. Hopkins, Daniel J.	2011	AJPS	Regression discontinuity design. Democratic mayors spend less on public safety. All other policy areas (including taxation) are unaffected.	Yes
Trounstine, Jessica	2010	Annual Review	Race and ethnicity in local elections (not relevant to us). Partisan elections have higher turnout; non-partisan elections still tend to have some partisanship in them because voters learn about party of candidates from media. Non-partisan elections favor Republicans/upper class. Mixed evidence for whether partisanship of mayor is important for policy.	Somewhat
Palus, Christine Kelleher	2010	State and Local Government Review	Ideology (liberal/conservative) of citizen is well represented by gov. spending in five areas: (1) community development, housing, and conservation, (2) health and human services, (3) culture, the arts, and recreation, (4) environmental programs, and (5) transportation.	Somewhat
Ferreira, Fernando Gyourko, Joseph	2009	The Quarterly Journal of Economics	Regression discontinuity design. Null results for spending and city gov. size with regard to mayor partisanship.	Yes
Ansolabehere, Stephen Snyder, James M.	2006	Scandinavian Journal of Economics	Despite the journal, this is about the U.S. The important finding (for us) is the fact that counties whose government is controlled by the same party as the state government, receive more funding (county's share of state transfers, normalized by county pop.) from the state.	Somewhat
Murphy, Russell D.	2002	Annual Review	Not useful. Too philosophical; mostly cites papers written a hundred years ago. Also exclusively about larger cities.	No

Names	Year	Journal	Findings	Important?
Armstrong, Cory L.	2011	Government Information Quarterly	Comparison of county and school board websites in Florida (where the two align) with regard to transparency (presence or absence of public records). Manual content analysis (undergrads told to look around for 15 minutes). School board websites, more professional websites, and websites in Republican-dominated counties are found to be more transparent.	Yes
Cegarra-Navarro, Juan Pachón, José Cegarra, José	2012	International Journal of Information Management	Survey of Spanish municipal government officials (specifically, the city website managers). Respondents are asked about the features of their websites, the level of civic engagement and the size of their municipality. More sophisticated websites are correlated with greater civic engagement and greater use of e-government functions.	Yes
Dolson, Jordan Young, Robert	2012	Canadian Journal of Urban Research	Determinants of website content. Three categories: e-content (city information on website), e-participation, social media use. Tables on page 15 show frequencies of these categories across sites, and might be useful to inform our topics. Larger cities have better websites. Population growth and immigration are also tested, but the findings are somewhat inconclusive.	Yes
Feeney, Mary K. Brown, Adrian	2017	Government Information Quarterly	500 U.S. city websites at two points in time (2010-2014). Count model of website features regarding information, e-services, utilities, transparency and civic engagement. Having a larger population leads to more features. Relying on a website contractor leads to more information and transparency. The authors say that mayor-councils are negatively correlated with website sophistication, but their regression tables state the opposite.	Yes
Kaylor, Charles Deshazo, Randy Van Eck, David	2001	Government Information Quarterly	Model of best practices of e-government. Table 1 lists a number of possible ways this manifests, could be useful for our theory.	Somewhat
Ansolabehere, Stephen Urban, Florian	2002	Cities	Websites of 20 major cities across the world. Is website content correlated with city characteristics? Not particularly systematic, and the findings are inconclusive.	Somewhat
Jeffres, Leo W. Lin, Carolyn A.	2006	Journal of Computer-Mediated Communication	50 largest metropolitan areas in the U.S. Features include information about city, opportunities for citizen feedback, galleries of photos, links, etc. Purely descriptive analysis, doesn't contain anything that isn't covered in any of the other aricles.	No

There doesn't seem to be much literature on transitions of party control, if anything, that question is mostly phrased with regard to political representation. However, if we want to tie our paper to a larger theory, we could go with dynamic representation. Under dynamic representation, policy-makers are responsive to trends in citizens opinions, which mainly manifest/become apparent through election results, especially when incumbents are voted out of office. This fits our topic quite well. Also, of the few papers that do exist on the effects of partisan transitions, virtually all use regression discontinuity designs.

3.1 Data

The General Services Administration (GSA) maintains all .gov addresses, and provides a complete[1] list of all such domains to the public through GitHub[2]. This list is updated once per month - we rely on the version released on January 16, 2017. The data from the GSA contains the following variables: One, domain name, specifically, the all-uppercase version of domain and top-level domain (for example, 'ABERDEENMD.GOV'). Two, the type of government entity to which the domain is registered, such as city, county, federal agency, etc. Three, for federal agencies, the name is specfied. Finally, the city in which the domain is registered, is noted.

Here, we focus only on cities. As a first step, we use a webdriver-controlled browser (Firefox/Selenium/Geckodriver) to test whether all of the city websites actually work. Of the 2425 domains listed by the GSA as cities, 292 are not accessible. Furthermore, the .gov domain, as registered at the GSA, is frequently not the website a city actually uses. In many cases, these sites redirect to another address, sometimes not a .gov domain (in this case, we simply use this domain). We record these URLs, as they are required to retrieve the images websites stored in the Wayback Machine (WbM).

$T_1 + \beta_1$

In order to provide an overview of our coverage (as not all cities, towns and villages use .gov addresses), we merge this list with U.S. Census data[3]. Here, several limitations in the GSA data need to be accounted for: One, even though the GSA nominally separates websites of cities and counties, some of the domains categorized as cities actually belong to counties. The same is true for townships and boroughs. Ergo, we eliminate all websites belonging to these three types of entities by hand. Furthermore, the city name, as given by the GSA, refers to the city in which the domain is registered, which is not necessarily equivalent to the city the website serves. In many cases, a website of a larger city may be registered to one of its subdivisions (for example, the website of New York is registered to Brooklyn), or vice versa (for example, the website of Homecroftin, a small town within Indianapolis, is registered to the city as a whole). Consequently we fix mismatches between websites and cities manually. Finally, a number of cities are simply misspelled, which we also correct by hand.

[1]Domains used for testing and internal programs are excluded.

[2]https://github.com/GSA/data/tree/gh-pages/dotgov-domains

[3]http://www2.census.gov/programs-surveys/popest/datasets/2010-2015/cities/totals/sub-est2015_all.csv

After the counties, townships and cities that cannot be matched to the Census data[4] and duplicate websites (some cities have more than one website) are removed, 1813 domains/cities remain.

These cities contain 90,616,865 people, and thus about 28% of the U.S. population (see figure 1).

We use the resulting list of websites to acccess their copies stored in the Internet Archive's Wayback Machine. To this end, we rely on the Ruby Gem 'Wayback Machine Downloader'[5] (WbMD). We supply the URL that each .gov website redirects to to the WbMD, which then downloads every file present in the WbM from a snapshot in October 2016, or, if not available, as soon as possible after this point.

<Note: We have not actually done this last step for all websites (however, the R script which runs the Ruby package is already set up to do so once we need to). Instead 10 websites were randomly sampled from an older version of the GSA list, which still contained counties and townships, which is why one of the 10 websites is from Dutchess County, NY.>

It would be fine to focus on Indiana as a case. First, we need to answer some preliminary questions about the data.

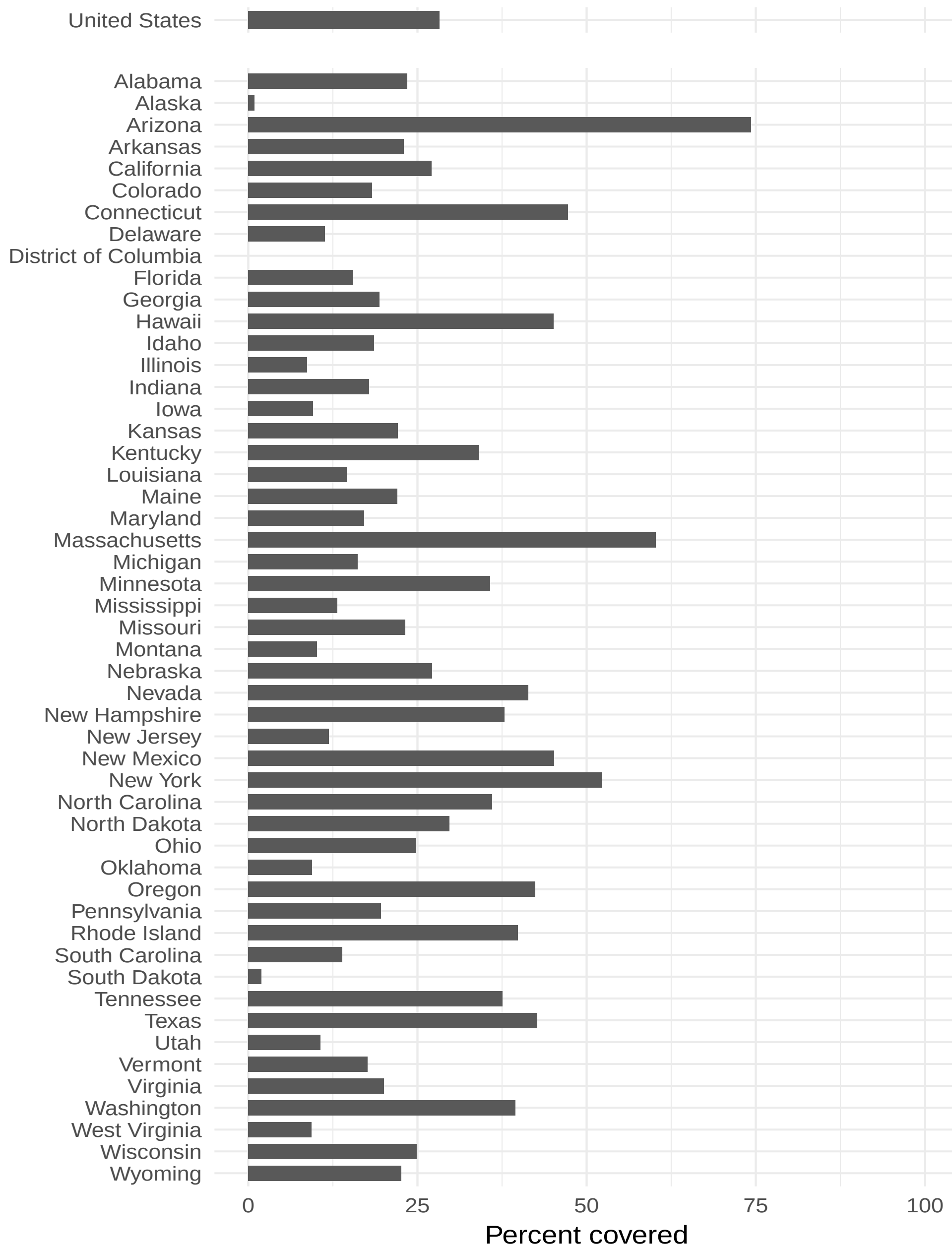
1. For what percentage and number of IN cities can we find data from the WBM?
2. For how many election cycles can we find political leadership data for these matched cities?
3. In what number and percentage of cities is the local leadership majority Republican?
4. Relatedly, in a typical election cycle, for how many cities do we see a transition in party leadership (i.e., a shift from majority D (R) to majority (R) D).

1. 30 cities, with a combined population of 1,180,435. However, since only cities (as opposed to towns and villages) hold mayoral elections, only 16 of these, with a combined population of 1,094,383 can be matched to the election data.
2. 2015, 2011, 2007, 2003.
3. Of the 16 cities, 7 have Republican mayors after the 2015 elections.
4. In 6 cases, a shift of party control occurs, with 4 of these being Republican –> Democratic.

[4] There are five cities that are not contained in the Census data

[5] https://github.com/hartator/wayback-machine-downloader

Figure 1: Percentage of state population covered.



Filetype	current	before	after
	51455	13866	19199
pdf	9646	5489	7544
jpg	5216	1988	3512
html	3767	17842	17596
aspx	2832	4356	3271
png	2714	2327	3684
gif	1068	664	1077
JPG	478	182	263
1	443	61	54
css	390	265	518
js	350	255	468
htm	264	295	256
docx	203	106	120
doc	167	70	130
asp	161	201	211
svg	87	55	69
php	83	157	241

Table 1: The most common file types in scraped websites

Website	current_size	current_files	before_size	before_files	after_size	after_files	size_change	files_change	control_change
attica-in.gov	61988	1417	7528	164	55956	1390	7.43	8.48	0.00
bedford.in.us	57628	560	27452	182	46388	525	1.69	2.88	0.00
cityofboonvilleindiana.com	9848	110	16996	172	20784	229	1.22	1.33	0.00
frankfort-in.gov	205368	2652	12208	242	138360	1077	11.33	4.45	0.00
warsaw.in.gov	298440	2117	26844	539	360400	2036	13.43	3.78	0.00
www.bloomington.in.gov	131128	2713	443360	14384	247096	9640	0.56	0.67	0.00
www.brazil.in.gov	43056	845	34472	625	55152	1214	1.60	1.94	0.00
www.carmel.in.gov	2270016	8727	1919344	5361	899900	2219	0.47	0.41	0.00
www.ci.auburn.in.us	183296	1025	21444	345	23564	211	1.10	0.61	0.00
www.cityoffortwayne.org	2136424	4378	266784	3582	233600	3018	0.88	0.84	0.00
www.cityofhobart.org	722000	2463	44192	650	62660	1037	1.42	1.60	0.00
www.evansvillegov.org	6345932	11844	290784	1281	1697224	6853	5.84	5.35	0.00
www.gary.in.us	373888	1227	121812	485	157140	719	1.29	1.48	0.00
www.huntingburg-in.gov	388680	2496	8644	213	375900	1953	43.49	9.17	0.00
www.jasperindiana.gov	561968	4013	55900	460	439072	2224	7.85	4.83	0.00
www.lakestation-in.gov	48	2	7724	84	257272	1097	33.31	13.06	0.00
www.linton-in.gov	32	1	24	2	24	2	1.00	1.00	0.00
www.madison-in.gov	531044	1848	36636	575	191624	1444	5.23	2.51	0.00
www.martinsville.in.gov	46792	1463	71628	1052	80944	800	1.13	0.76	0.00
www.monticelloin.gov	33656	753	18120	448	100680	2104	5.56	4.70	0.00
www.newhavenin.org	84364	626	2524	86	6792	334	2.69	3.88	0.00
www.richmondindiana.gov	250968	1042	217252	918	401672	2422	1.85	2.64	0.00
www.southbendin.gov	1264076	4749	454456	3286	1424136	2562	3.13	0.78	0.00
connersvillecommunity.com	170688	569	162316	815	187276	808	1.15	0.99	1.00
www.batesvilleindiana.us	166564	2348	39592	496	95696	1310	2.42	2.64	1.00
www.cityofrisingsun.com	994956	3311	321400	1268	80848	868	0.25	0.68	1.00
www.cityofrockport-in.gov	12068	98	5148	16	12068	98	2.34	6.12	1.00
www.elkhartindiana.org	1132828	2345	5588	123	6204	223	1.11	1.81	1.00
www.elwoodcity-in.org	224412	765	5000	123	139692	517	27.94	4.20	1.00
www.indy.gov	5726048	9675	6119260	10451	4984080	7981	0.81	0.76	1.00
www.northvernon-in.gov	272016	403	3132	112	289336	416	92.38	3.71	1.00
www.winchester-in.gov	364592	2480	6508	135	45488	567	6.99	4.20	1.00

Table 2: Number of files and size of websites

For some cities, whose websites make heavy use of JavaScript, this method does not lead to satisfying results. Consequently we restricted our corpus to cities with at least 3 documents.

4 The Web to Text Pipeline

In the methodological pipeline from native website files to text data that is appropriate for comparative analysis we address two methodological challenges. First, though they contain significant amounts of text, websites are not comprised of clean plain text files. Rather, the files available at websites are of multiple types, including HTML, PDF, word processor, plain text, and image files. The first step in the methodological pipeline is aimed simply at extracting clean plain text from this heterogeneous file base. The second step in our methodological pipeline is to process the text to remove boilerplate language—language that is effective at differentiating one website from another, but is uninformative regarding policy or process differences between governments. We describe these methodological steps in this section.

4.1 Site to Text Conversion

For the most part, the file type of a document can be correctly determined through its ending. However, there are exceptions to this, which, if ignored, can lead to large amounts of garbage text, stemming from incorrectly converted documents, as well as a general decrease in the amount of usable data. Two issues in particular need to be addressed: One, HTML files on city websites frequently do not have an ending, but are still perfectly readable if correctly identified as such. Second, some documents contain the incorrect file ending - for example, we found thousands of documents on the New Orleans city website that ended in .html, when they were actually PDFs. To accurately assess their type, we read in the first line of each document, which, if it is an HTML or PDF file, contains a string indicating as much. Consequently we rename all documents so that their file ending reflects their actual file type. This is strictly necessary, because we rely on the readText R package[6] - which determines a document's type solely through its ending - to convert the files to plain text.

The text documents are then read into R line by line, converted to UTF-8 and then stripped of dates, punctuation, numbers and words connected by underscores. At this point, the documents of one city still closely resemble one another in the form of boilerplate content, be it website elements (i.e. "You are here", "Home", "Directory" etc.) in html documents, or commonly used forms or phrases in pdfs, doc and docx files. This is an issue, because it clusters documents around the cities from which they originate in a way that has nothing to do with their actual content. In other words, the signal would be drowned out by the noise. Our solution to this problem is described in more detail in section 4.2. Preprocessing further includes setting every character to lowercase, as well as the removal of bullet points which frequently occur in html documents, extraneous whitespace, xml documents mislabeled as html files, and empty documents. Furthermore, some documents

[6]We have also experimented with several Unix-based alternatives, but found that they largely led to the same results.

contain gibberish, often as a result of faulty or impartial OCR. To combat this problem, we employ two solutions. One, we use spellchecking, implemented through the hunspell R package, to remove all non-English words. However, hunspell does not cover everything, either because some tokens are not actual words (for example artifacts from defective encoding), or because random sequences of characters just so happen to form words that exist in a dictionary (for example "eh" or "duh"). Since we rely on a bag-of-words model in which syntax does not matter, we can ameliorate these problems by removing all text except for whitespaces and the characters that appear in the English alphabet. Since a lot of the nonsensical text tends to be quite repetitive, we also delete all documents in which the proportion of unique to total number of tokens is less than 0.15. Furthermore, hunspell does not spellcheck individual characters or two-character words, so we remove these token types entirely (none of these words are of any substantive relevance to our research question). Since these pre-processing steps reduce documents which are largely unsuitable to only a few words of texts that don't make much sense, we also remove all remaining documents containing less than 50 tokens. Finally, to remove words that are extremely rare (which also has the advantage of eliminating any remaining oddities) and thus add nothing substantive to our models while increasing their computational cost, we also discard any token types that occur in only one document.

4.2 Boilerplate Removal

As noted above, city websites contain a large amount of text that is uninformative for its actual content and therefore a hindrance to correct analysis by automatic text processing methods. Consequently we remove this content as following: Each line of every document is compared to every line in every other document belonging to the same city. We count how many times each line is duplicated for that city. We remove any line occuring more than our chosen threshold of 10.[7] This means that each document only retains the information that is particular about it. We implement this algorithm through hash tables, which reduces the computational complexity from O(N^2) to O(N). Before this step is taken, we remove numbers and dates from the documents because they frequently make lines unique, despite the fact that they are virtually the same (for example different days on a city calendar).

5 Bag-of-Words Text Analysis

5.1 Topic modeling

Note: this chapter is mostly a wordy and less coherent version of the above.

We hypothesize that a change in leadership from one party to the other will lead to a change in website content because the two parties have different agendas. Democrats have a predilection towards policies that promote social and economic equality, whereas Republicans like to emphasize

[7]Empirically, lines tend to be duplicated either hundreds of times, or only once or twice, if at all.

small government as well as law and order. Documents uploaded to city websites are expected to be a reflection of these preferences.

The Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan 2003) is the most commonly used topic model. However, it is unable to account for the existence of two parties with very different policy agendas, translating to different preferred topics. There are two types of extensions to the LDA that fit our subject much better - the structural topic model, and the author-topic model.

The structural topic model, developed by Roberts, Stewart and Airoldi (2016) allows researchers to model a corpus as a function of metadata associated with its documents. Specifically, topic prevalence (the proportion of a document made up by a topic) and topical content (the rate at which words figure into a topic) are contingent on a set of covariates. In our case, the two most important covariates are (1) city and (2) authoring party (operationalized by whether a document was present before a change of power, or introduced afterwards). Furthermore, the population size of the city should be a predictor for both the number and kind of problems it faces, which thus need to be addressed on its website. Furthermore, city size also serves as an estimate for the budget and technical capacities of its staff in charge of maintaining the website[8]. Further demographic as well as economic data might also be useful to differentiate cities from one another. If we model the differences between cities properly, we might not have to/should not include city as a (categorical) variable, because it would probably interfere with these more meaningful covariates.

The author-topic model (Rosen-Zvi, Griffiths, Steyvers and Smyth 2004) would allow us to capture the fact that different authors have different topical preferences. Unfortunately, we have two types of 'authors' - cities, and parties. Given the largely divergent administrative needs of different types of cities, we would likely have to treat cities as the author. This would require us to capture the partisan authorship of different documents entirely on the basis of sub-sets of the website data - changed, added, or deleted documents. (Note: In the papers on author-topic models, the intention is often to analyze scientific articles. These articles are often co-authored. Would it be possible to have BOTH cities and parties as authors, so that a specific version of a website would then be 'co-authored' by its city and party?)

The critical element in this analysis is to accurately attribute authorship of documents to either party. Despite possible changes to websites due to a leadership transition, large parts of the content carry over. This means that unless the successor government decides to delete everything, some of the existing documents will be preserved, and in the model, also attributed to the new 'author'. But the reverse is not possible, because the predecessor government can't choose to retain documents from the future. *This is a very important point for municipal websites. We should investigate the possibility of modeling only the changes—documents that change, documents that are deleted, and documents that are added.*

Labeling newly added documents after a change of power is quite simple. As far as older documents are concerned, we would have to operate under the assumption that the incumbent

[8] Although this relationship is not exactly deterministic - when looking through .gov websites manually, I've noticed that a lot of websites of (presumably wealthy) towns of only a few thousand citizens often have extremely well-kept websites

didn't keep his or her successor's document's on the website for four years.[9] One problem here is the fact that the incumbent would have all the administrative topics assigned to them, simply because they have to have those on their website.

If we really do end up getting swamped with administrative terms in our topic models (and it does kind of look like that at the moment), we might be able to separate the signal from the noise by running a preparatory LDA once and using its results to create a new, corpus-specific list of stop words. After that, we run the actual model. This way, politically charged terms and topics, which likely are not as common, but present nevertheless, should be able to rise to the surface. It might be possible to refine this process by running an exploratory model on website data from cities in which party control never changes, and the incumbent always wins by large margins. 'Safe' cities like this should have fairly homogeneous populations, with little need for the incumbent to play politics on the municipal website. Hence, these websites should be filled with purely administrative content.

The use of asymmetric priors (Wallach, Mimno and Mccallum 2009) over the document topic distribution - i.e. the assumption that some topics, such as administrative content, are inherently more common - may be a more elegant way of dealing with this issue.

Another intervening factor is that for cities in Indiana, mayoral terms begin in January. Since a lot of clerical and administrative tasks tend to be year-specific, work tends to pile up around the new year. Thus it is possible that a spike in newly added documents is not due to a change in party control, but owed to a seasonal increase in activity. We can test for this by comparing election years to non-election years. Furthermore, since in Louisiana, mayors take office in May, we have another point of comparison.

Furthermore, if we only investigate cities in which control of government changes from one party to another, we may overestimate its effect. Not only does a transition in party control occur, but the person in charge also changes. Parties are fairly homogeneous, so that two mayors from the same party may have very different policy preferences and managerial styles. To remedy this problem, we [could] utilize matching, pairing our cases with similar cities in which the incumbent does not run for re-election, but party control stays the same nevertheless.

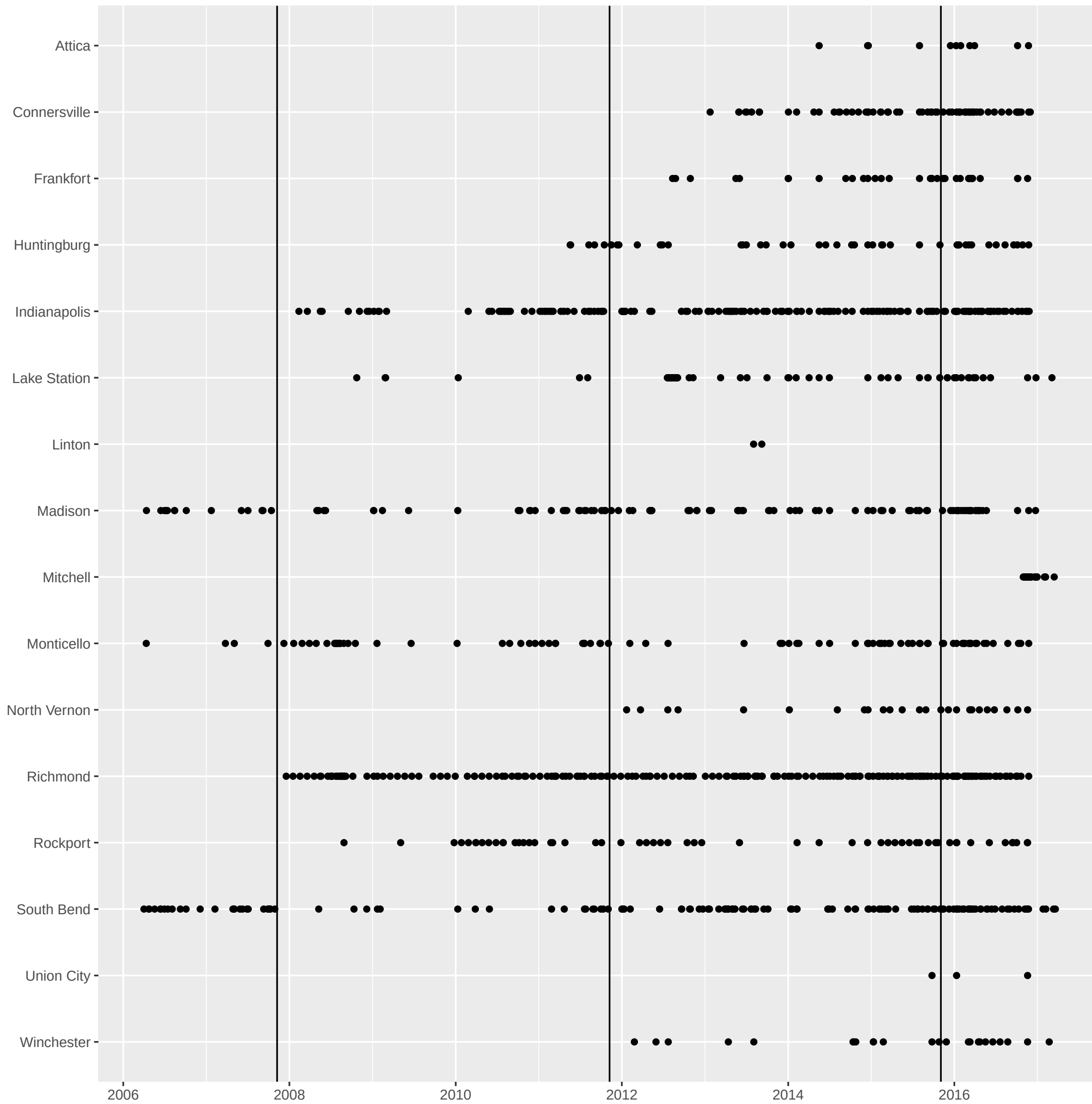
In order to determine the optimal number of topics for our corpus, we utilize the measure developed by **?**. In contrast to other approaches such as **?** or **?**, this measure includes as much information about the model as possible by taking into account both the document-term matrix, as well as the word-topic matrix. We find 200 topics to be the best value (see figure 13).

5.1.1 LDA diagnostics

Figure 3 plots the densities of topic weights across documents, where each line represents a topic. Distributions with lower peaks near zero and flatter bodies indicate a topic that is present

[9] Probably a safe assumption. However, we could, and probably should test how long documents tend to stay on a city website. Simple descriptive statistics (for example density plots) on the length of existance should likely be sufficient. If we want to be really fancy about it, we could create a duration model, with document topics as features. This would allow us to measure whether some documents tend to remain longer based on their topic (i.e. fire regulations are probably going to stay up longer than notes on a specific council meeting).

Figure 2: Dates of Wayback Machine snapshots. The vertical lines are municipal elections.



to varying degrees in multiple documents. This shape appears to be more common in documents from Republican cities. By contrast, a distribution with a high peak near zero and a very long tail is indicative of a topic that only appears in one specific document. This appears to be more common for documents in Democratic cities. The substantive conclusion then is that Democrats appear to be more prone to dedicating individual documents to one specific purpose, whereas Republicans produce general-purpose documents more often.

An alternative way to consider topic weights is to aggregate them across documents, in this case through their median or mean. Figure 4 shows this distribution of topic weights across topics. Most topics have a very low weight - meaning that they do not appear in a lot of documents. When aggregating across documents via the mean, this effect is more pronounced for Republicans, suggesting that most topics do not feature frequently (or even at all) in their documents. Democrats on the other hand appear to have a wider spectrum of topics from which they chose - supporting the hypothesis of Democrats as a 'big tent' party. However, this effect is reversed when using the median instead of the mean to aggregate across documents. Now, the the distribution has a higher peak and lower tail for Democrats. The cause for this stark divergence appears to lie in the fact the mean and median differ enormously - for the range displayed here, the mean is about a hundred times greater. Evidently, extremely large weights (i.e. some documents fitting specific topics perfectly) distort the picture. The median document is a better representation of the corpus as a whole, but the results do not fit our hypothesis. Their are however more consistent with figure 4.

Topics for which there are stark differences in their distribution across documents between Democrats and Republicans are of particular interest to us. To detect these topics, we take the absolute difference of the median document topic weights, and arrange them as a histogram, see figure 5. For the most part, differences are small. For a few topics however, a contrasts emerges.

Figure 6 displays the word-topic probabilities for the 10 topics with the largest partisan differences. Some of these, such as topic 82, with its focus on policing, safety and crime could be construed as politically charged. For the most part however, these topics focus on administrative matters.

To further investigate the topics with large partisan differences, figure 6 shows densities of topic weights just like figure 3, but reduced to only the topics with the largest differences. It appears that mere absolute differences between the medians (or some other measure of central tendency) obscures the fact that the distributions themselves are quite different.

To further investigate this issue, we look at the raw data itself - the document-topic matrices for Democrats and Republicans, displayed as heatmaps. Figure 9 shows that especially for Republicans (but also for Democrats, to a lesser extent), there appears to be extensive clustering for consecutive documents. Since the order of the documents in the corpus is dependent on the cities in which they appear in, it seems that topics are mere representation of city websites - each website 'owns' a number of topics, that appear across all of its documents, and hardly anywhere else. One possible cause for this type of clustering is the fact that documents frequently share common words, for example pertaining to navigation on the site, or standard forms that are shared throughout all documents.

5.2 Informative Dirichlet model

For the analysis of the data, we present two approaches, the first being the informative dirichlet model developed by (Monroe, Colaresi and Quinn 2008). This approach aims to account for the fact that some words naturally occur more than others by applying a Dirichlet prior based on the distribution of words in random text. Table 3 shows the top words for both Democrats and Republicans - and accomplishes, to some extent, the goal of (Monroe, Colaresi and Quinn 2008) of banishing frequent words from this list and supplanting them with text with greater semantic, and in our case, partisan meaning.

In Indiana, Democrats exhibit a preference for words related to public finance, such as 'fund', 'budget', or 'tax', indicative of a greater willingness to emphasize the city's efforts to raise and spend money. This finding is consistent with (Einstein and Kogan 2015), who show that Democratic mayors tend to favor greater spending. Beyond the focus on public finance, the words preferably used by Democrats do not fall into any particularly congruent categories, and largely sort into various areas related to city administration - i.e. 'council', 'services', 'budget', 'committee', 'contract', etc. If there is theme around the words preferred by Republicans, it seems to center around city planning - street, fire, water, building, construction, park. These words suggest that the hands-off approach favored by Republicans results in a focus on supporting infrastructure and logistics.

For Lousiana, the results (see table 4) are less coherent. Only one of the finance-related terms appears again for Democrats - specifically 'fund', although 'rate' might also be used in a financial context. Beyond that, some focus on a 'historic' ''district of a city seems evident, as is the use of some words - 'infrastructure', 'water', 'building' that were used for Republicans in Indiana. Conversely, Republicans are now missing these words, and their preferred terms generally do not seem to follow any particular theme.

The weakness of the fightin' words method is evident here, as a list of words does not necessarily provide sufficient information to glean preferred topics from. This is especially the case when the texts are spread across a broad number of issue-areas, with little semantic similarity. In (Monroe, Colaresi and Quinn 2008), the authors focus on the fairly constrained corpus of U.S. Senate speeches with respect to abortion - our context, by comparison, is far more eclectic.

5.2.1 Structural topic model

A more powerful approach with the capacity of addressing this problem is the use of topic models. This class of clustering methods relies on the co-occurrence of words within documents to form a set of semantically coherent topics. In order to compare the degree to which Republicans and Democrats prefer specific topics, we rely on the structural topic model, developed by (Roberts, Stewart, Tingley, Lucas, Leder-Luis, Gadarian, Albertson and Rand 2014). Theoretically, the most widely-used form of topic model, latent dirichlet allocation, can also be used to test for the impact of a single covariate through a post-hoc comparison, but the structural topic model allows for multiple covariates, and also produced more meaningful topics in our experiments.

We use 60 topics - the number recommended by the authors for medium- to large-sized corpora, and party as well as city population (the literature frequently emphasizes city size as a determinant of the issues it faces - see, for example, Guillamón, Bastida and Benito (2013)) as covariates. The results are shown in tables 5 to 8. The coefficients in the table headers describe the size of the party covariate on a given topic. In order to test statistical significance, we calculated credible intervals -

the topics shown here are all significant at the 0.1% level.

In Indiana, some of the topics associated with Democrats - one related to education, one to recycling - clearly seem to match the party brand. Interestingly enough, Democrats also 'own' the topic related to law enforcement, which might be somewhat unexpected given Republicans' usual focus on law and order (Gerber and Hopkins 2011). However, this kind of finding is not entirely without precedent in the literature (see (Einstein and Kogan 2015)). Similar to the informed dirichlet model, the structural topic model also finds the emphasis on construction and infrastructure by Republicans - in table 5, topics 2, 7 and 8 clearly focus on these issues.[10]

When comparing Indiana to Louisiana, it appears that the Democratic emphasis on law enforcement is robust. Furthermore, as with the fightin' words approach, some smaller degree of focus on money (see topic 1) is still evident. For Republicans, topics 2 to 4 seem to be, once again about infrastructure and utilities, pointing to a certain degree of robustness in these results, as well as the emergence of a trend. The results produced by the structural topic model are not flawless, but the two parties do seem to have somewhat consistent themes on which they focus on in both states. Furthermore, in comparison to the fightin' words approach, the ability of the structural topic model to form coherent topics is quite evident and helpful in the interpretation of the results.

6 Conclusion

We have developed a methodological pipeline for automatically gathering and preparing government websites for comparative analysis. This methodology holds the potential to vastly scale up the data collection efforts underpinning the rapidly growing body of research that is focused on government website analysis. Through an application to the analysis of municipal websites in Indiana and Louisiana, we show how our pipeline is capable of gathering corpora that shed light on the forms and functions of local government.

[10]The first Repuiblican topic in Indiana (library, stream, obj, etc.) is likely an artifact from incorrectly converted html, and since it presumably only happens only in one Republican city, the topic is classified as very Republican.

References

Adamic, Lada A. and Natalie Glance. 2005. "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog." *Proceedings of the 3rd international workshop on Link discovery - LinkKDD '05* pp. 36–43.

Armstrong, Cory L. 2011. "Providing a clearer view: An examination of transparency on local government websites." *Government Information Quarterly* 28(1):11–16.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Cryer, J. E. 2017. "Candidate Identity and Strategic Communication." pp. 1–42.

de Benedictis-Kessner, Justin and Christopher Warshaw. 2016. "Mayoral partisanship and municipal fiscal policy." *The Journal of Politics* 78(4):1124–1138.

Druckman, James N., Cari Lynn Hennessy, Martin J. Kifer and Michael Parkin. 2010. "Issue Engagement on Congressional Candidate Web Sites, 20022006." *Social Science Computer Review* 28(1):3–23.
URL: *http://journals.sagepub.com/doi/10.1177/0894439309335485*

Druckman, James N., Martin Kifer and Michael Parkin. 2009. "Campaign Communications in U.S. Congressional Elections." *American Political Science Review* 103(03):343–366.
URL: *http://www.journals.cambridge.org/abstract_S0003055409990037*

Duarte, Eugenio. 2017. "The Un/Deniable Threat to LGBTQ People." *Contemporary Psychoanalysis* pp. 1–6.

Einstein, Katherine Levine and David M Glick. 2016. "Mayors, partisanship, and redistribution: Evidence directly from US mayors." *Urban Affairs Review* p. 1078087416674829.

Einstein, Katherine Levine and Vladimir Kogan. 2015. "Pushing the City Limits: Policy Responsiveness in Municipal Government." *Urban Affairs Review* pp. 1–30.

Eschenfelder, Kristin R, John C Beachboard, Charles R McClure and Steven K Wyman. 1997. "Assessing U.S. federal government websites." *Government Information Quarterly* 14(2):173–189.
URL: *http://www.sciencedirect.com/science/article/pii/S0740624X97900186%5Cnpapers2://publication/doi/10.101*
624X(97)90018-6

Esterling, Kevin M, David Lazer and Michael A Neblo. 2011. "Representative Communication: Website Interactivity & Distributional Path Dependence in the U.S . Congress.".

Esterling, Kevin M. and Michael A. Neblo. 2011. "Explaining the Diffusion of Representation Practices among Congressional Websites." *Working Paper* pp. 1–42.

Feeney, Mary K. and Adrian Brown. 2017. "Are small cities online? Content, ranking, and variation of U.S. municipal websites." *Government Information Quarterly* 34(1):62–74.
URL: *http://dx.doi.org/10.1016/j.giq.2016.10.005*

Gentzkow, Matthew and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers." *Econometrica* 78(1):35–71.

Gerber, Elisabeth R. and Daniel J. Hopkins. 2011. "When Mayors Matter: Estimating the Impact of Mayoral Partisanship on City Policy." *American Journal of Political Science* 55(2):326–339.

Glez-Peña, Daniel, Anália Lourenço, Hugo López-Fernández, Miguel Reboiro-Jato and Florentino Fdez-Riverola. 2013. "Web scraping technologies in an API world." *Briefings in bioinformatics* 15(5):788–797.

Grimmelikhuijsen, Stephan G. 2010. "Transparency of Public Decision-Making: Towards Trust in Local Government?" *Policy & Internet* 2(1):5–35.

Grimmelikhuijsen, Stephan G and Eric W Welch. 2012. "Developing and testing a theoretical framework for computer-mediated transparency of local governments." *Public administration review* 72(4):562–571.

Guillamón, Ma Dolores, Francisco Bastida and Bernardino Benito. 2013. "The electoral budget cycle on municipal police expenditure." *European Journal of Law and Economics* 36(3):447–469.

King, G., J. Pan and M. E. Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science* 345(6199):1251722–1251722.
URL: *http://www.sciencemag.org/cgi/doi/10.1126/science.1251722*

King, Gary, Jennifer Pan and Margaret E. Roberts. 2017. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument." *American Political Science Review* 111(03):484–501.
URL: *https://www.cambridge.org/core/product/identifier/S0003055417000144/type/journal_article*

King, Gary, Jennifer Pan and Margaret Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107(02):326–343.
URL: *http://www.journals.cambridge.org/abstract_S0003055413000014*

Kirby, Reid. 2017. "The Trump?s administration?s misaligned approach to national biodefense." *Bulletin of the Atomic Scientists* 73(6):382–387.

Lin, Y-R, J P Bagrow and D Lazer. 2011. "More Voices than Ever? Quantifying Bias in Social and Mainstream Media." *arXiv preprint arXiv* 1111(1227).

Marion, Nancy E and Willard M Oliver. 2013. "When the Mayor Speaks... Mayoral Crime Control Rhetoric in the Top US Cities: Symbolic or Tangible?" *Criminal justice policy review* 24(4):473–491.

Mayhew, David. 1974. *Congress: The Electoral Connection*. Yale University Press.

Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2008. "Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16(4 SPEC. ISS.):372–403.

Norris, P. 2003. "Preaching to the Converted?: Pluralism, Participation and Party Websites." *Party Politics* 9(1):21–45.

Osman, Ibrahim H, Abdel Latef Anouze, Zahir Irani, Baydaa Al-Ayoubi, Habin Lee, Asım Balcı, Tunç D Medeni and Vishanth Weerakkody. 2014. "COBRA framework to evaluate e-government services: A citizen-centric perspective." *Government Information Quarterly* 31(2):243–256.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand. 2014. "Structural topic models for open-ended survey responses." *American Journal of Political Science* 58(4):1064–1082.

Roberts, Margaret E, Brandon M Stewart and Edoardo M. Airoldi. 2016. "A model of text for experimentation in the social sciences." *Journal of the American Statistical Association* pp. 1–49.
URL: *http://www.tandfonline.com/doi/full/10.1080/01621459.2016.1141684*

Rosen-Zvi, M., T. Griffiths, M. Steyvers and P. Smyth. 2004. "The author-topic model for authors and documents." *Proceedings of the 20th conference on Uncertainty in artificial intelligence* pp. 487–494.
URL: *http://portal.acm.org/citation.cfm?id=1036902*

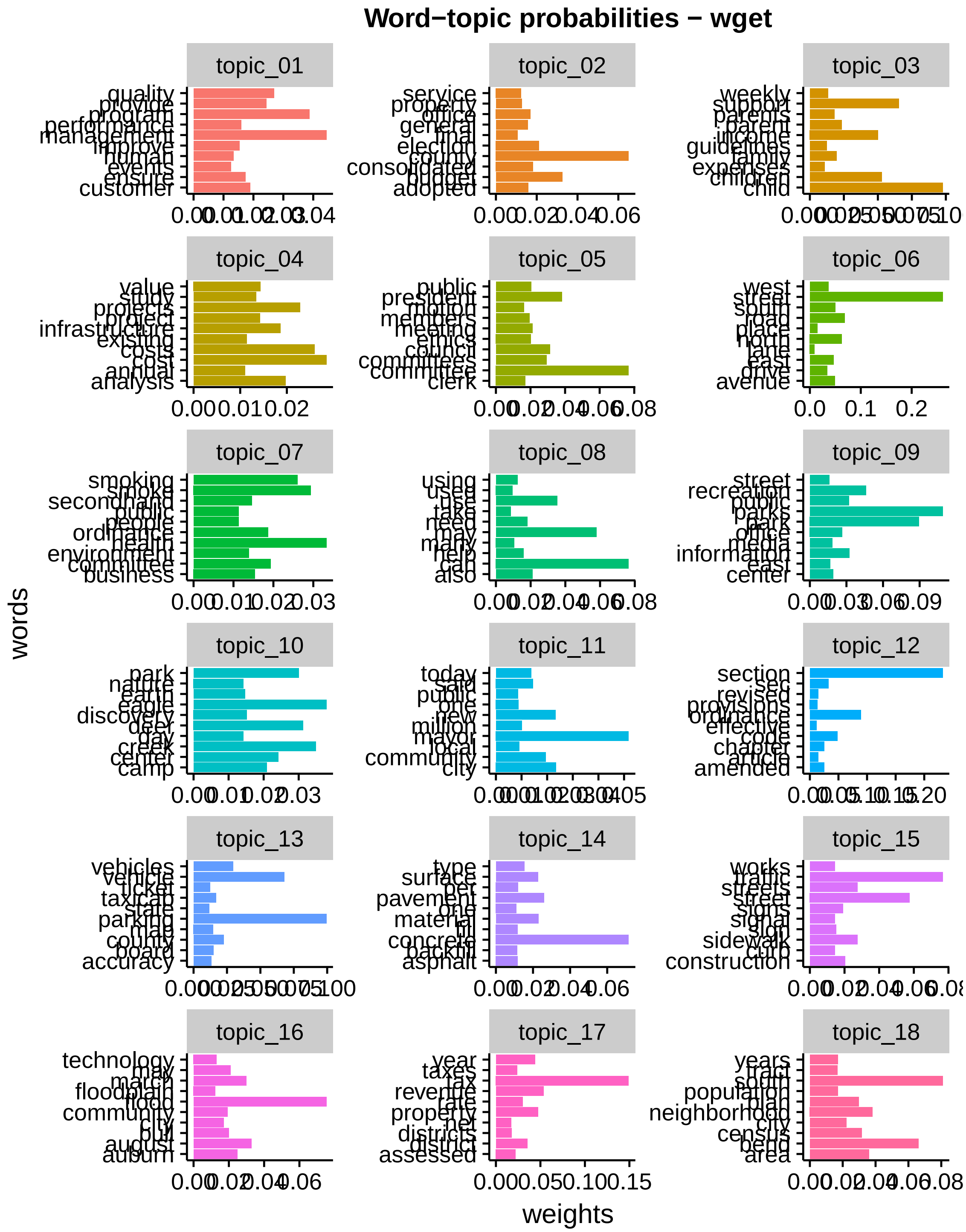
Sharfstein, Joshua M. 2017. "Science and the Trump Administration." *Jama* 318(14):1312–1313.

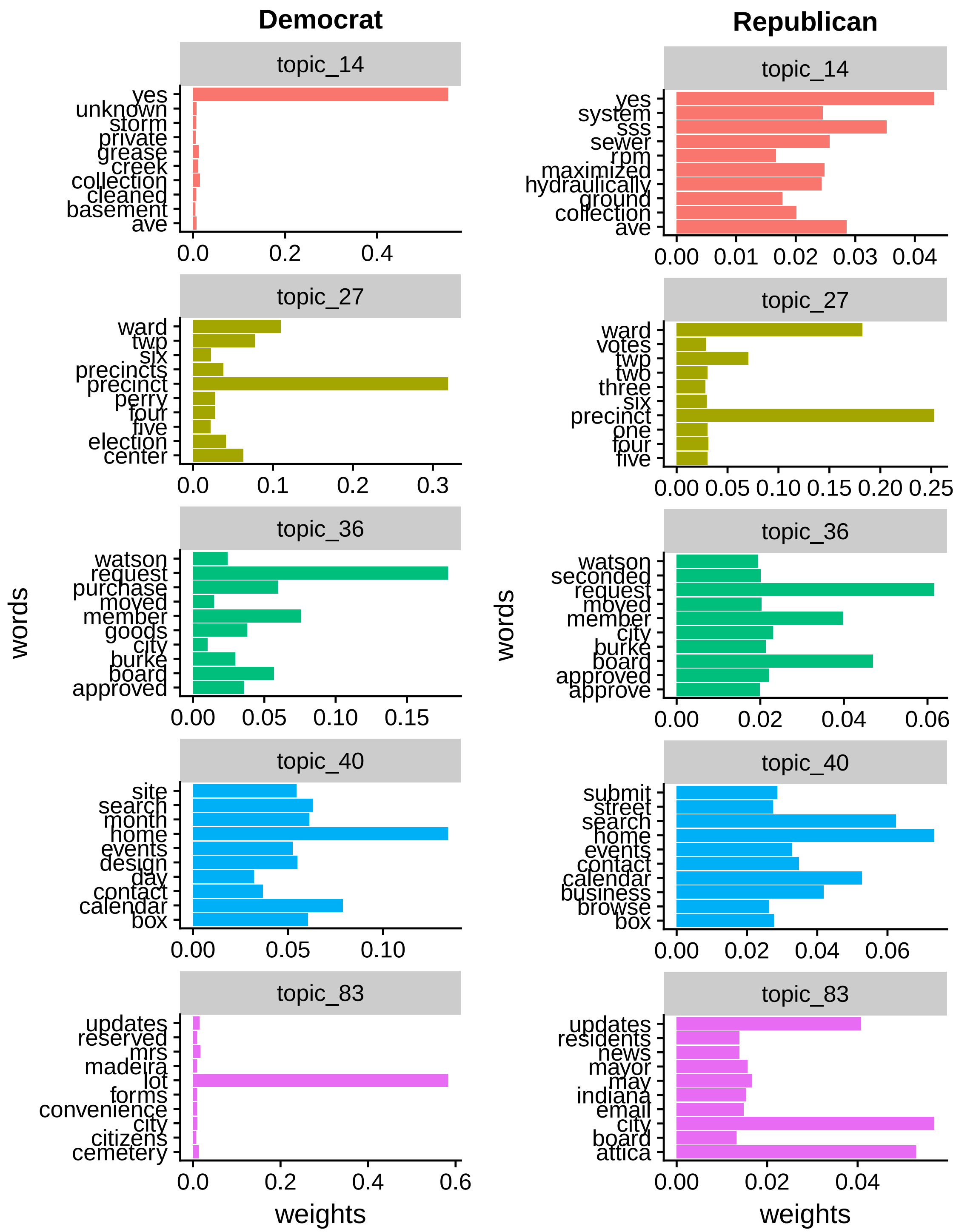
Therriault, Andrew. 2010. "Taking Campaign Strategy Online: Using Candidate Websites to Advance the Study of Issue Emphases." pp. 1–23.
URL: *http://poseidon01.ssrn.com/delivery.php?ID=588125096113080101107007109104101121035031077054017*

Urban, Florian. 2002. "Small town, big website? Cities and their representation on the internet." *Cities* 19(1):49–59.

Wallach, Hanna M, David Mimno and Andrew Mccallum. 2009. "Rethinking LDA: Why Priors Matter." *Advances in neural information processing systems* .

Wang, Lili, Stuart Bretschneider and Jon Gant. 2005. Evaluating web-based e-government services with a citizen-centric approach. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*. Ieee pp. 129b–129b.


Word−topic probabilities − wget
topic_01
topic_02
topic_03
topic_04
topic_05
topic_06
topic_07
topic_08
topic_09
topic_10
topic_11
topic_12
topic_13
topic_14
topic_15
topic_16
topic_17
topic_18
words
weights



Democrat
Republican
topic_14
yes
unknown
storm
private
grease
creek
collection
cleaned
basement
ave
system
sss
sewer
rpm
maximized
hydraulically
ground
topic_27
ward
twp
six
precincts
precinct
perry
four
five
election
center
votes
two
three
one
topic_36
watson
request
purchase
moved
member
goods
city
burke
board
approved
seconded
approve
topic_40
site
search
month
home
events
design
day
contact
calendar
box
submit
street
business
browse
topic_83
updates
reserved
mrs
madeira
lot
forms
convenience
citizens
cemetery
residents
news
mayor
may
indiana
email
attica
words
weights

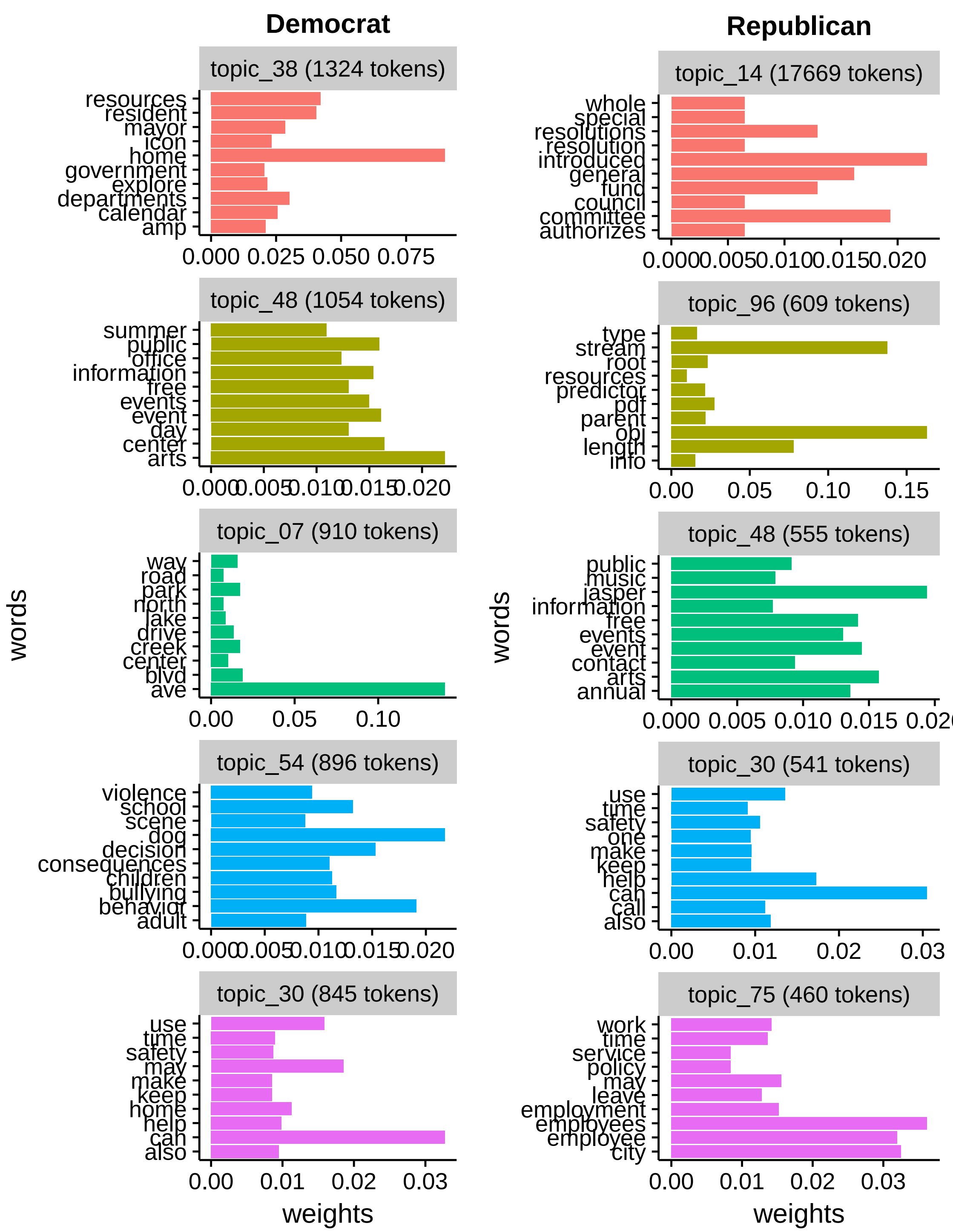

Democrat
Republican
topic_38 (1324 tokens)
resources
resident
mayor
icon
home
government
explore
departments
calendar
amp
0.000 0.025 0.050 0.075
topic_14 (17669 tokens)
whole
special
resolutions
resolution
introduced
general
fund
council
committee
authorizes
0.000 0.005 0.010 0.015 0.020
topic_48 (1054 tokens)
summer
public
office
information
free
events
event
day
center
arts
0.000 0.005 0.010 0.015 0.020
topic_96 (609 tokens)
type
stream
root
resources
predictor
pdf
parent
obj
length
info
0.00 0.05 0.10 0.15
topic_07 (910 tokens)
way
road
park
north
lake
drive
creek
center
blvd
ave
0.00 0.05 0.10
topic_48 (555 tokens)
public
music
jasper
information
free
events
event
contact
arts
annual
0.000 0.005 0.010 0.015 0.020
words
topic_54 (896 tokens)
violence
school
scene
dog
decision
consequences
children
bullying
behavior
adult
0.000 0.005 0.010 0.015 0.020
topic_30 (541 tokens)
use
time
safety
one
make
keep
help
can
call
also
0.00 0.01 0.02 0.03
topic_30 (845 tokens)
use
time
safety
may
make
keep
home
help
can
also
0.00 0.01 0.02 0.03
topic_75 (460 tokens)
work
time
service
policy
may
leave
employment
employees
employee
city
0.00 0.01 0.02 0.03
weights

Figure 3: Densities of topic weights for documents in Republican and Democratic cities (Indiana).

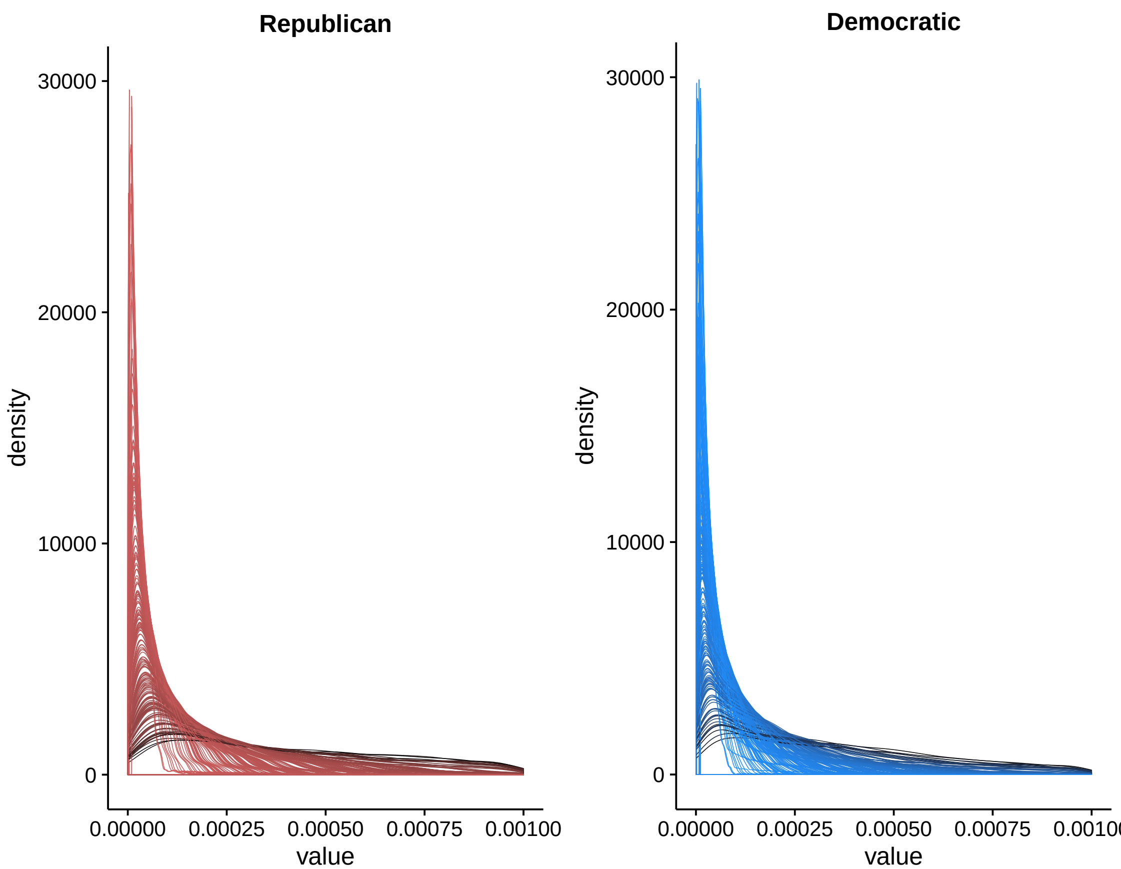


Figure 4: Densities of topic weights for documents in Republican and Democratic cities (Indiana).

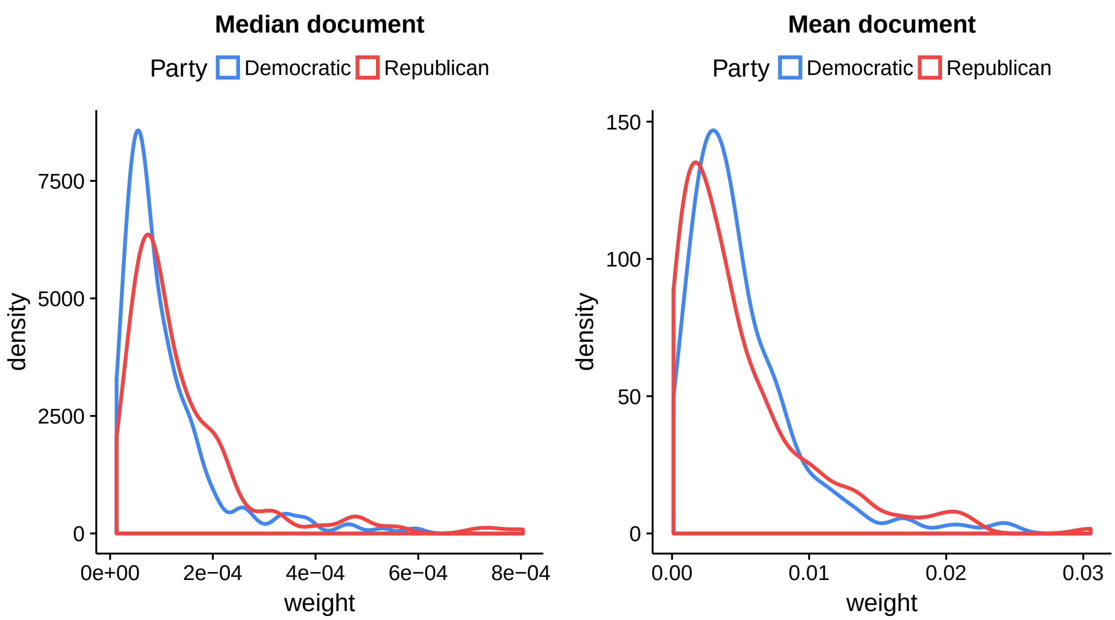


Figure 5: Word-topic probabilities for topics with big partisan differences (Indiana).

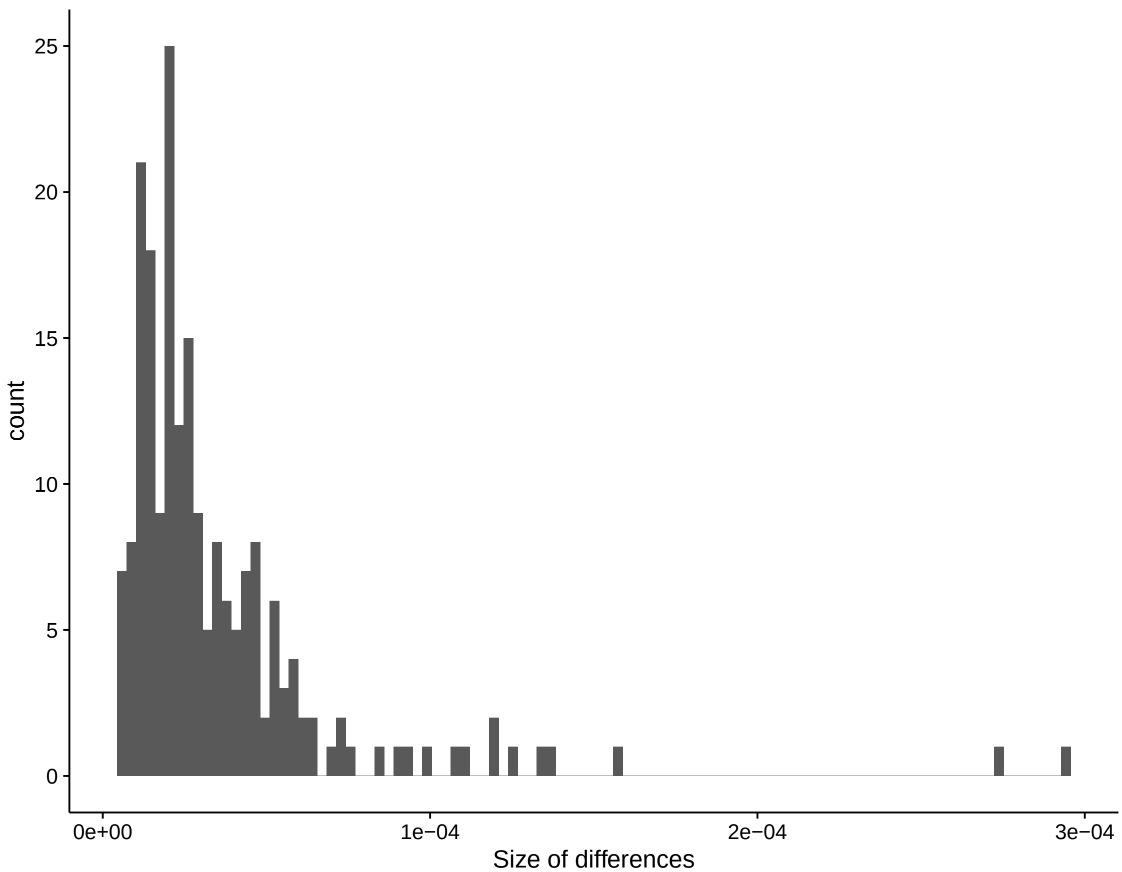


Figure 6: Word-topic probabilities for topics with big partisan differences

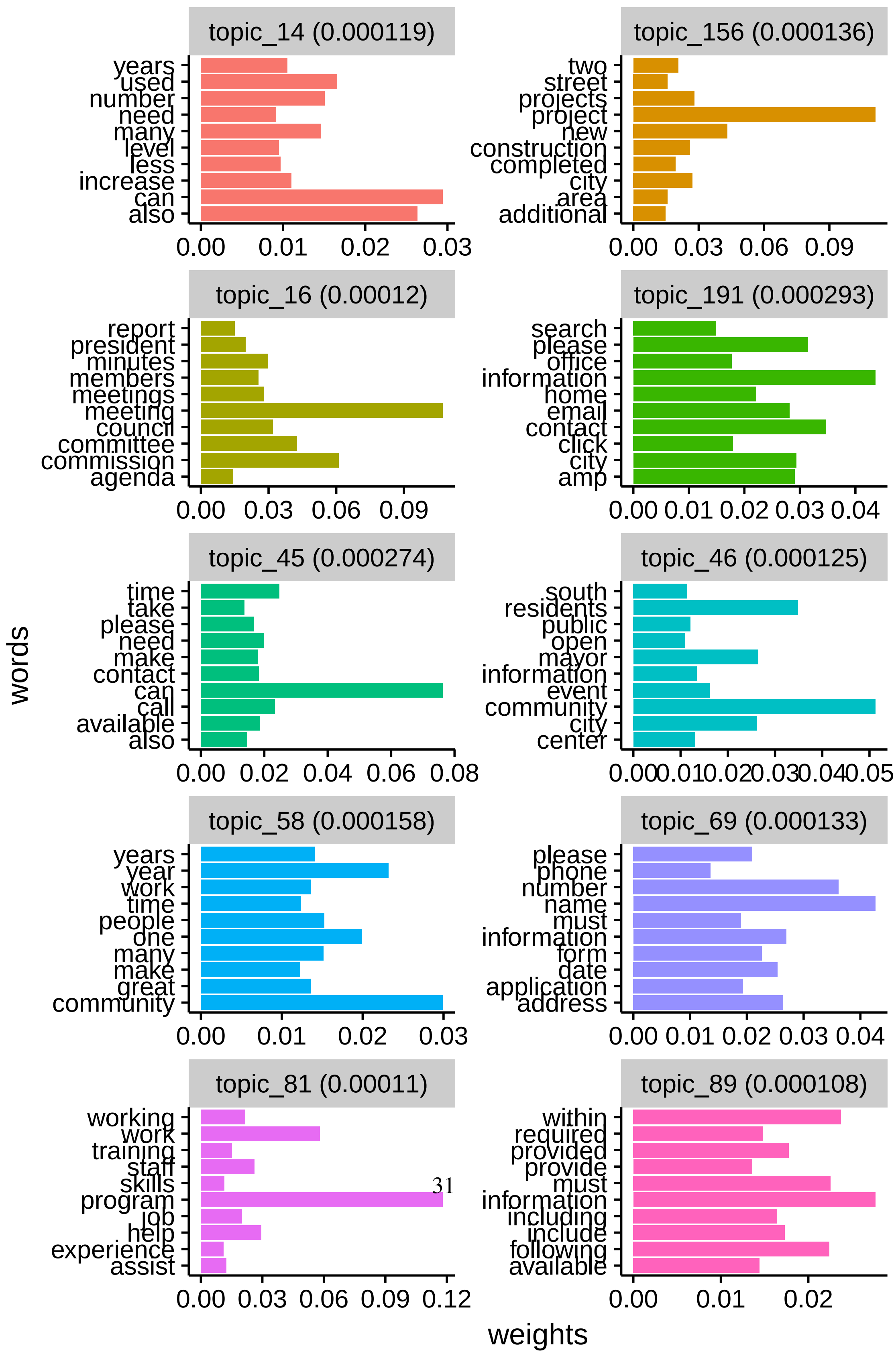
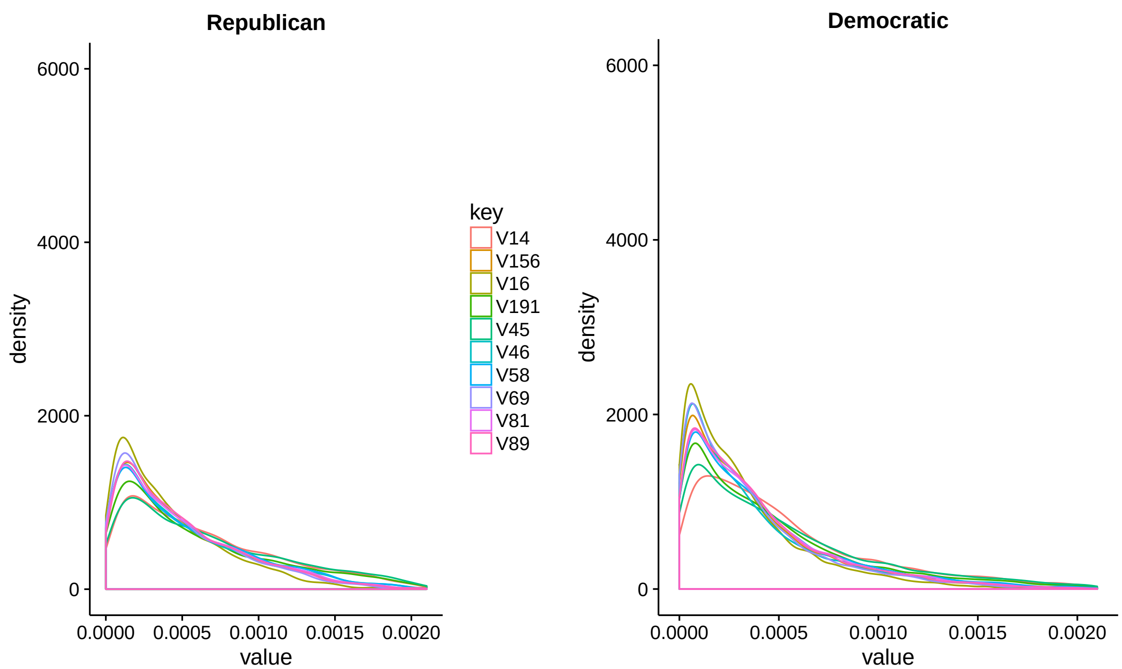


Figure 7: Word-topic probabilities for topics with big partisan differences



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
brown	32.78	downtown	35.86
management	29.69	old	35.22
clerk	29.66	root	34.96
increase	29.30	area	34.82
dollar	29.16	docket	34.81
appoint	29.10	rider	34.79
technology	29.07	station	34.45
service	28.32	variance	34.12
digest	28.30	use	34.00
recognize	27.90	carter	33.66
year	27.73	residential	33.56
justice	27.46	request	32.98
court	26.72	foot	32.76
criminal	25.99	clean	32.27
appropriation	25.60	obstruction	31.72
enterprise	25.54	rep	31.69
financial	25.45	overflow	31.42
sander	25.27	lateral	31.08
public	25.09	tablet	30.91
fiscal	24.77	river	30.70
corporation	24.58	road	30.32
whereas	24.46	ordinance	30.10
vendor	24.43	drive	29.96
sec	24.31	pump	29.95
prosecutor	24.30	clay	29.63
pursuant	24.02	secondary	29.61
crime	23.93	fence	29.54

Table 3: Top 50 Democratic and Republican words (Indiana), according to the informed Dirichlet model of Monroe et al. (2008).

Word (D)	z-Score (D)	Word (R)	z-Score (R)
otherwise	20.73	say	86.18
health	18.65	ordinance	77.67
respect	17.98	summary	59.81
use	16.62	bid	58.98
officer	16.22	council	46.92
staff	15.87	amount	41.21
district	15.82	official	39.79
historic	15.51	mayor	39.07
datum	15.19	accordance	37.91
fund	15.02	boulevard	37.78
thereto	14.86	weekend	35.41
building	14.70	weather	34.34
street	14.69	seal	33.27
total	14.60	responsive	33.15
window	14.50	veteran	31.96
applicant	14.41	resolution	29.52
exist	14.19	hold	28.71
housing	14.13	gathering	28.32
provide	13.84	furnish	27.36
review	13.58	councilman	27.19
source	13.54	meeting	26.74
neighborhood	13.09	exceed	26.54
revenue	12.99	show	26.44
target	12.88	emergency	26.01
policy	12.75	resident	25.23
training	12.52	city	24.89
process	12.51	accept	24.73
actual	12.45	visit	24.67
population	12.04	wheeler	24.21
green	11.95	night	24.11
rate	11.70	purchase	24.00
infrastructure	11.68	theater	23.76
urban	11.46	parish	23.63
average	11.45	sweep	23.39
retention	11.22	inc	23.27
master	11.03	tonight	22.09
bureau	10.93	recreation	21.92
roof	10.90	mike	21.82
strategy	10.89	park	21.78
water	10.82	department	21.71
construct	10.79	movie	21.65
residence	10.57	tropical	21.50
reduce	10.47	hall	21.49
relative	10.46	contract	21.31
construction	10.46	pet	21.24
monthly	10.46	morning	21.08
chapter	10.43	begin	20.84
individual	10.35	information	20.78
design	10.29	beach	20.60
standard	10.24	approve	20.56

Table 4: Top 50 Democratic and Republican words (Louisiana), according to the informed Dirichlet model of Monroe et al. (2008).

0.023	0.021	0.019	0.017	0.017	0.014	0.013	0.012
library	foot	team	ave	request	board	amp	building
stream	sign	game	inc	board	meeting	traffic	historic
obj	use	play	cross	member	member	stop	build
length	lot	league	creek	service	committee	vehicle	material
branch	building	camp	construction	street	council	block	preservation
type	zone	class	blvd	approve	commission	sign	wall
flag	area	age	park	city	meet	airport	roof
filter	district	must	lake	purchase	public	ave	window
rim	parking	child	hill	move	director	theft	floor
page	residential	participant	ridge	good	president	signal	new

Table 5: Top Republican topics and words (Indiana), according to STM. The words are the top words for the most Democratic/Republican topic, determined by the size (and significance) of the coefficient (see table header) of the party covariate.

-0.027	-0.022	-0.016	-0.015	-0.012	-0.011	-0.011	-0.01
city	school	downtown	service	contract	city	trash	housing
ordinance	community	business	division	bid	department	city	property
approve	program	project	provide	contractor	mayor	waste	program
resolution	student	city	city	city	police	day	fund
property	education	development	management	agreement	officer	recycle	home
purchase	university	new	public	work	public	street	city
area	national	center	department	service	citizen	collection	project
department	award	economic	program	department	work	resident	neighborhood
contract	high	company	include	bidder	safety	recycling	grant
service	year	community	office	move	resident	snow	unit

Table 6: Top Democratic topics and words (Indiana), according to STM. The words are the top words for the most Democratic/Republican topic, determined by the size (and significance) of the coefficient (see table header) of the party covariate.

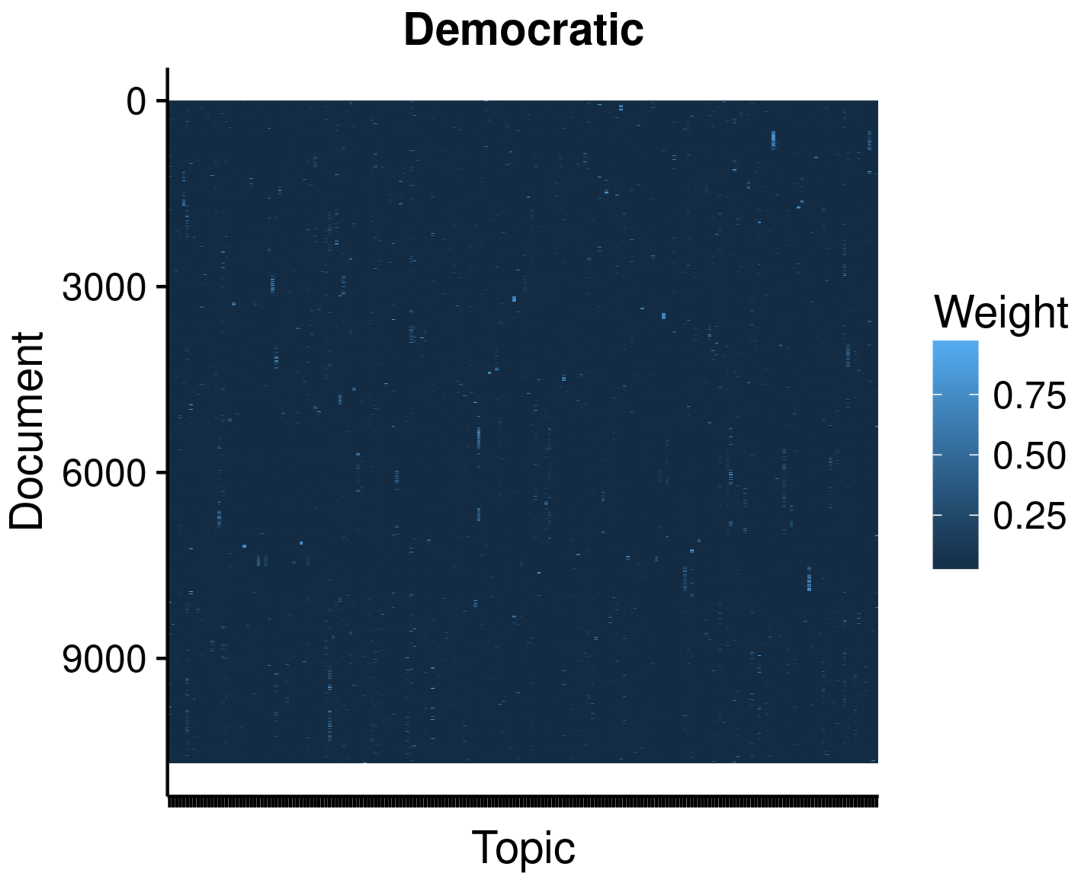
0.071	0.054	0.054	0.034	0.033	0.024	0.023	0.02
event	ordinance	water	street	say	city	city	mayor
information	department	emergency	traffic	can	business	meeting	city
show	summary	city	parking	make	new	council	parish
park	amount	resident	lane	get	mayor	commission	town
music	bid	storm	project	take	development	plan	office
food	city	weather	work	people	economic	member	hall
visit	public	waste	bike	work	million	public	contact
weekend	police	system	downtown	need	continue	board	day
festival	approve	power	public	city	work	committee	official
begin	inc	service	bicycle	help	local	planning	state

Table 7: Top Republican topics and words (Louisiana), according to STM. The words are the top words for the most Democratic/Republican topic, determined by the size (and significance) of the coefficient (see table header) of the party covariate.

-0.136	-0.102	-0.043	-0.02	-0.02	-0.012	-0.012	-0.012
art	otherwise	whereas	water	street	shall	police	event
call	provide	city	main	inc	city	crime	city
cost	respect	ordinance	sewer	drive	agreement	officer	park
home	city	bond	project	construction	party	suspect	rental
sponsor	thereto	provide	infrastructure	permit	provide	arrest	use
church	authorize	resolution	street	service	property	report	hour
amp	ordinance	code	system	avenue	owner	victim	hotel
free	district	chapter	improvement	oak	provision	information	public
museum	amend	shall	remark	park	section	murder	provide
artist	locate	otherwise	phase	lane	agree	block	term

Table 8: Top Democratic topics and words (Louisiana), according to STM. The words are the top words for the most Democratic/Republican topic, determined by the size (and significance) of the coefficient (see table header) of the party covariate.

Figure 8: Word-topic probabilities for topics with big partisan differences, across documents (Indiana).



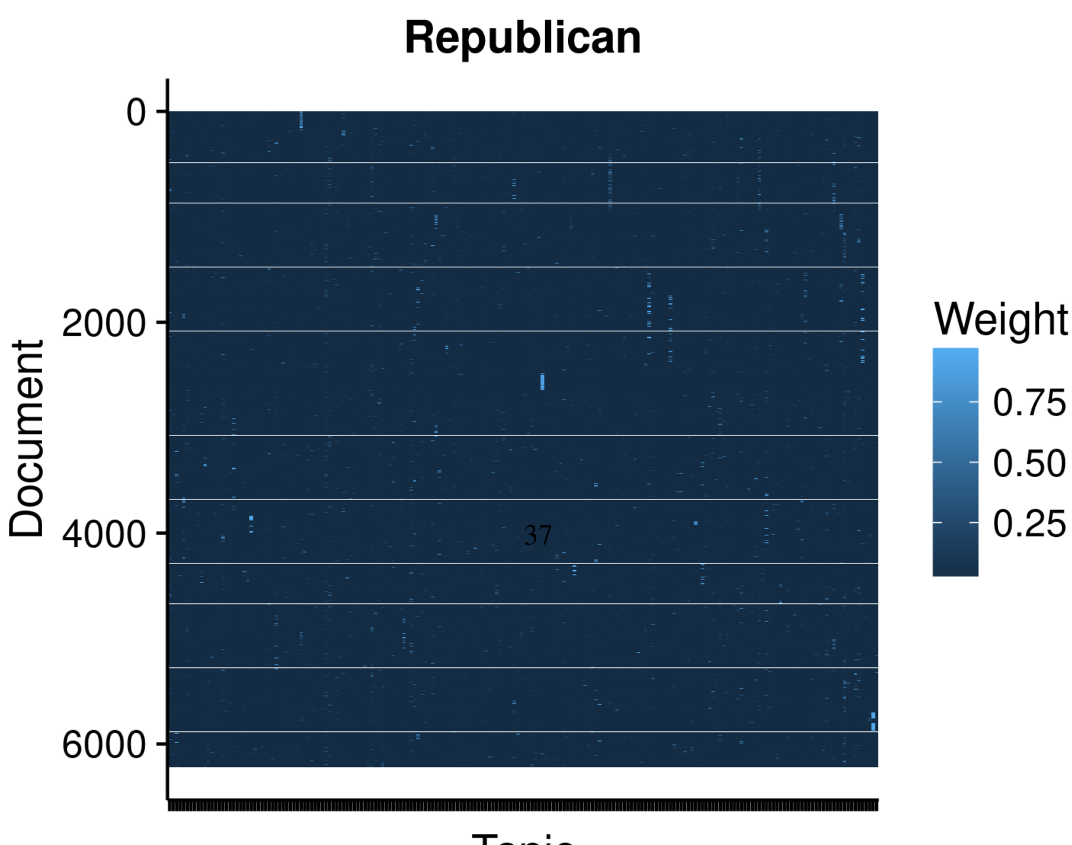


Figure 9: Word-topic probabilities for topics with big partisan differences, across documents (Indiana). No duplicate line removal.

Figure 10: Topic coherence, varying the number of topics. The red line represents the mean topic coherence for each number of topics.

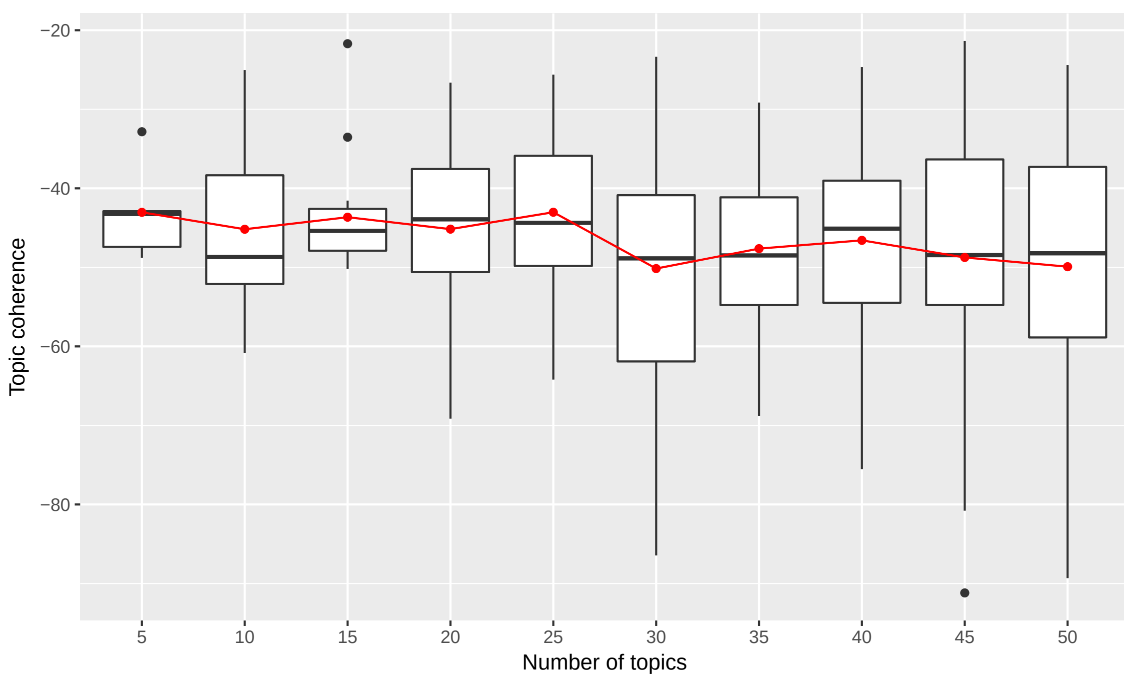


Figure 11: Cities in the corpus, by partisanship of mayor.

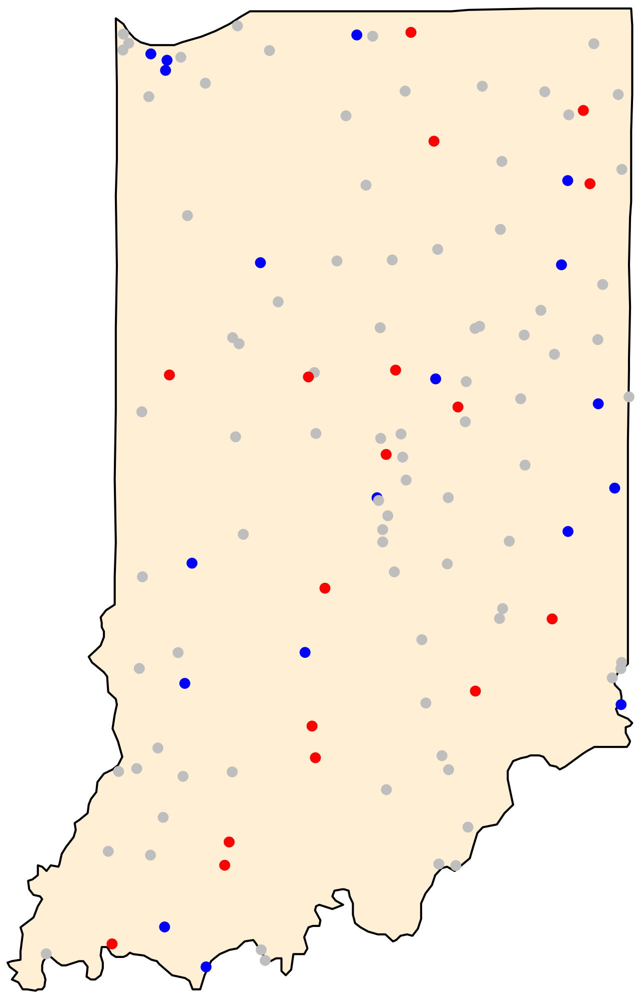


Figure 12: The most partisan topics. The number in parenthesis indicates how many times more the topic appears on average (measured through the number of words of the topic throughout the documents) in the respective party's corpus (indicated by the color).

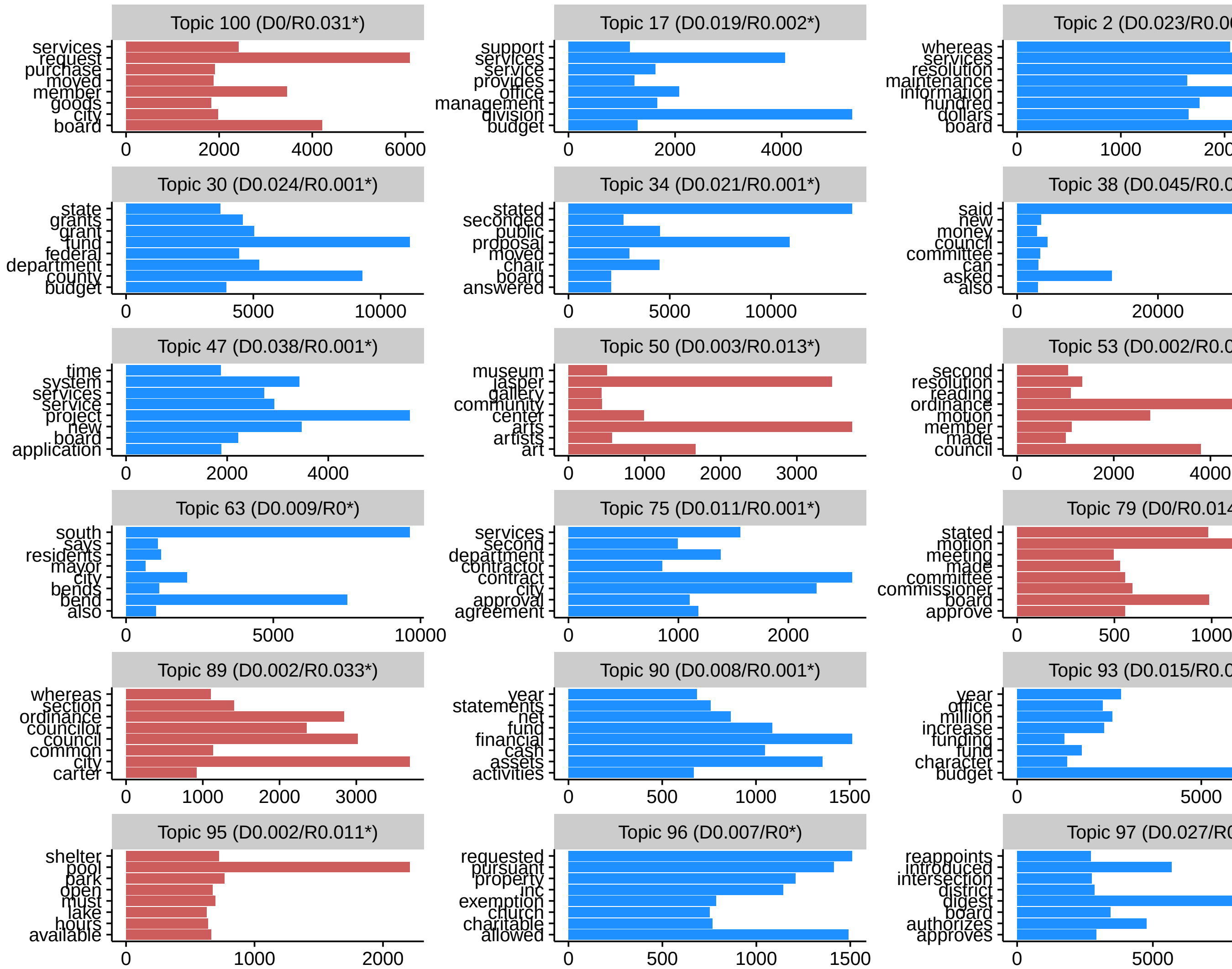


Figure 13: The optimal number of topics for our corpus. The measures of Griffths 2004 and Cao 2009 indicate the best number of topics at their minimum, whereas the measure of Arun 2010 points to the best value at its maximum.

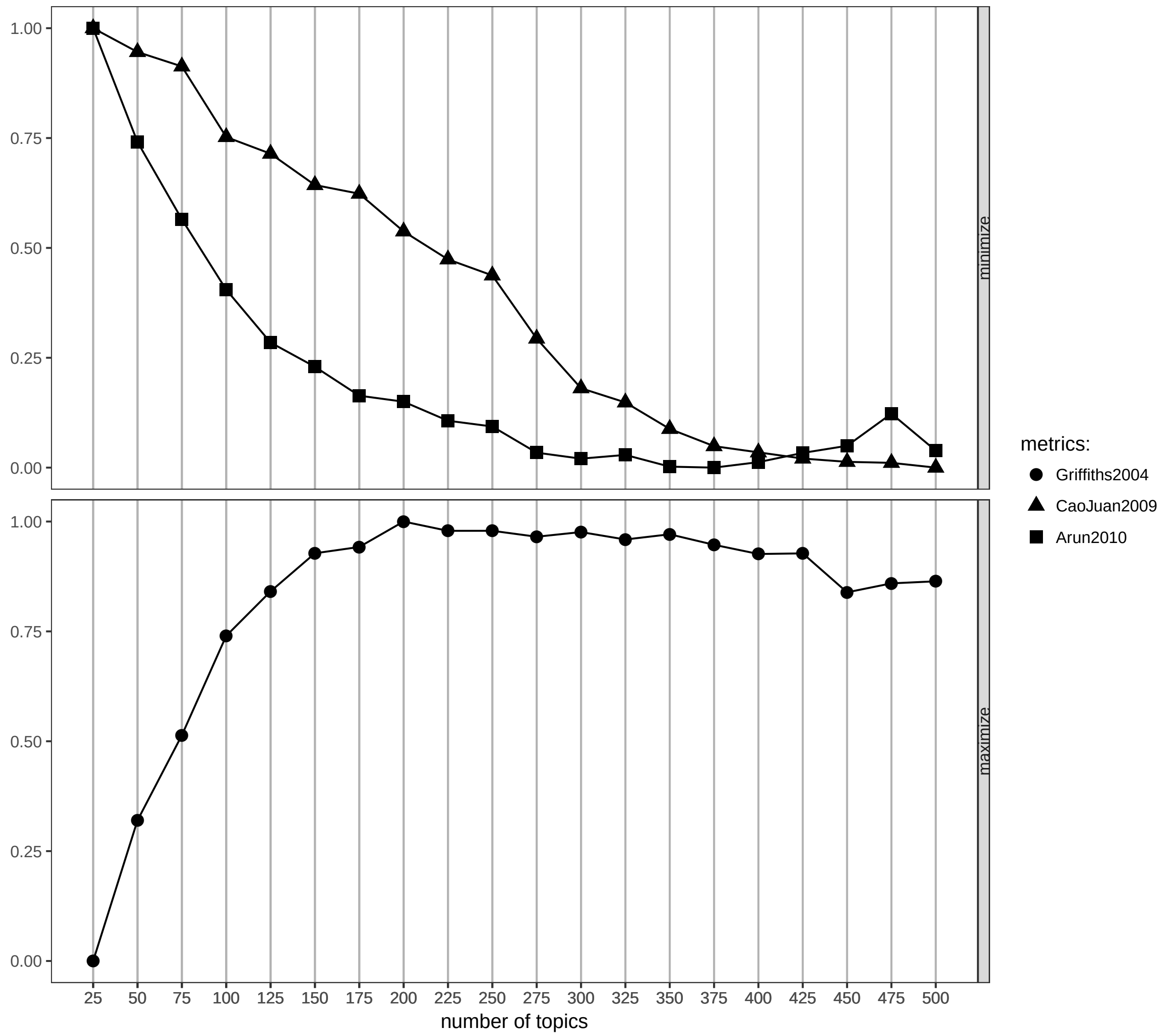


Figure 14: Results from a structural topic model, displayed as the p-values for each variable for each topic. This would normally be somewhat nonsensical, but here it illustrates why the model does not work.

Figure 15: Top democratic and Republican words (Indiana), according to the informed Dirichlet model of Monroe et al. (2008). Ordering is top to bottom for Democrats, and vice versa for Republicans.

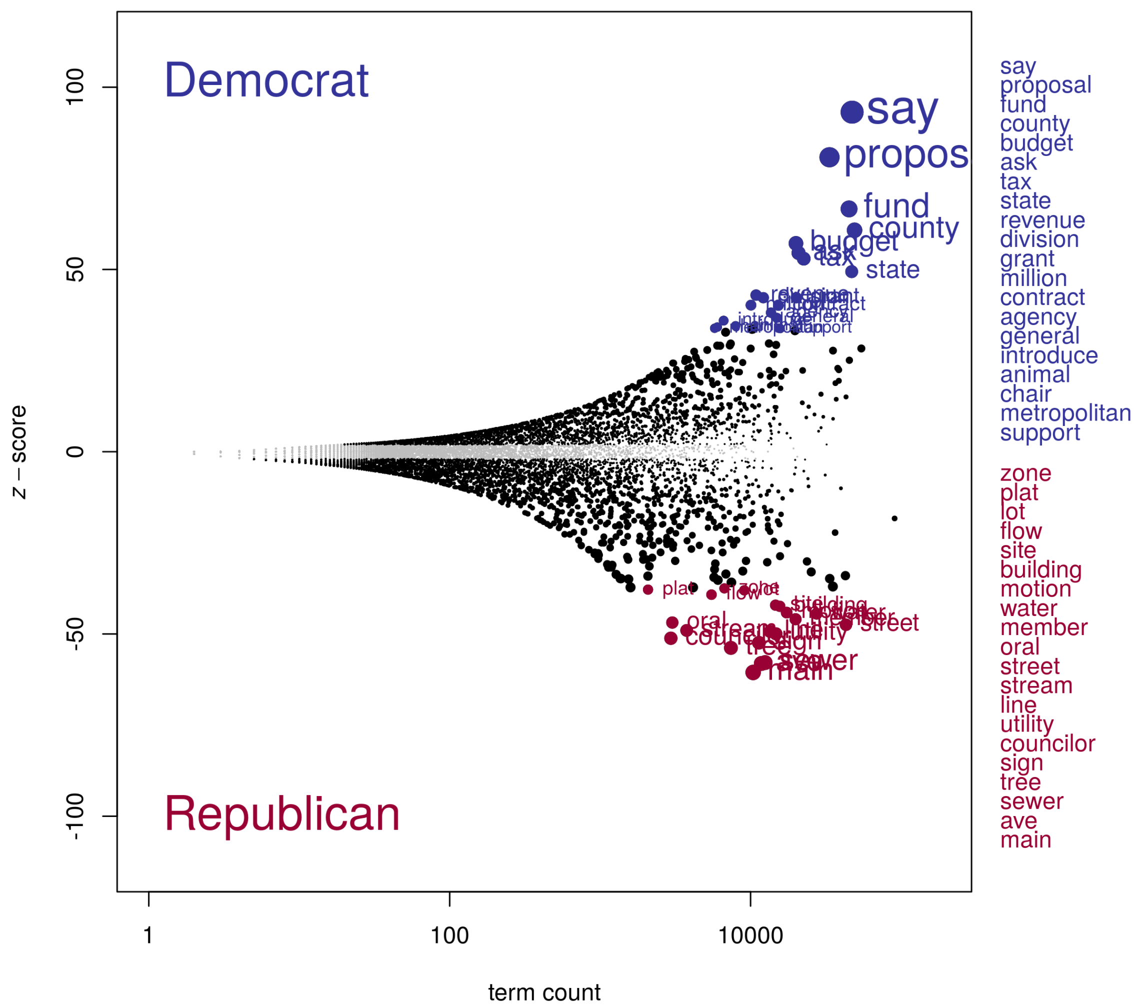
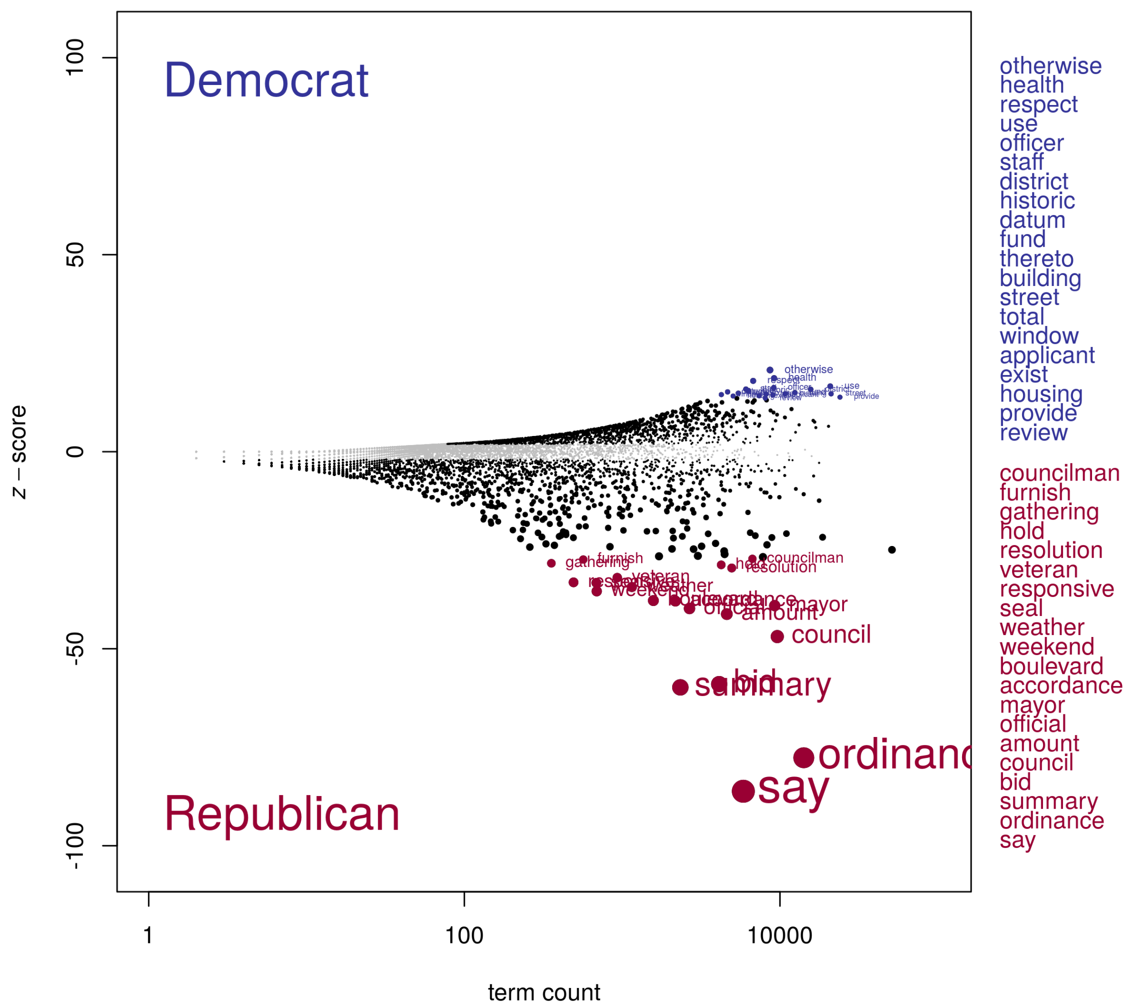


Figure 16: Top democratic and Republican words (Louisiana), according to the informed Dirichlet model of Monroe et al. (2008). Ordering is top to bottom for Democrats, and vice versa for Republicans.



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
office	13818	council	9098
program	13536	community	9072
year	13376	meeting	8990
service	13312	ave	8555
provide	13138	service	8040
one	13066	construction	7999
section	12669	one	7885
work	11986	property	7741
information	11886	also	7492
development	11854	per	7442
committee	11802	required	7407
district	11584	home	7334
time	11466	center	7316
total	10965	made	7301
general	10731	site	7279
parks	10704	business	7222
system	10668	time	7157
digest	10481	services	7140
police	10474	housing	7111
management	10433	new	7006
park	10356	within	6910
also	10112	date	6818
division	9964	year	6768
street	9853	following	6754
resolution	9768	road	6629
contract	9763	member	6450
ordinance	9456	inc	6367
safety	9362	number	6360
code	9342	day	6254

Table 9: Top 50 Democratic and Republican words (Indiana), according to LDA. Topic ownership is determined by the ratio of Democratic to Republican tokens in it (both weighted by the total number of tokens per party). The instances of each token type are then summed across all topics owned by the party.

Word (D)	Instances (D)	Word (R)	Instances (R)
city	19306	city	9930
stream	13397	ordinance	4413
new	13001	information	3756
obj	10440	council	3422
otherwise	8271	said	3301
street	7990	plan	3194
provide	7647	department	2991
district	7449	state	2598
property	7031	public	2594
public	6864	meeting	2392
shall	6750	mayor	2258
respect	6698	one	2166
water	6085	application	2105
thereto	5686	development	2017
development	5124	parish	1809
use	5086	can	1807
ordinance	4963	new	1807
business	4763	water	1780
department	4757	program	1691
community	4705	project	1674
authorizing	4440	time	1648
located	4315	code	1641
mayor	4266	year	1560
length	4215	date	1556
project	3918	number	1548
section	3863	name	1516
service	3831	street	1504
councilman	3824	motion	1500
services	3782	day	1483
zoning	3771	park	1471
parish	3731	home	1469
providing	3641	address	1415
one	3636	office	1408
system	3617	amount	1392
building	3607	ave	1384
can	3557	budget	1382
code	3532	please	1375
office	3305	community	1334
drive	3223	area	1326
work	3171	contact	1319
permit	3165	emergency	1308
following	3153	summary	1282
within	3123	also	1271
must	3088	make	1265
plan	3064	two	1224
neighborhood	3048	work	1213
construction	3016	fire	1184
chapter	2973	bid	1134
ordinances	2885	planning	1124
fire	2878	people	1108

Table 10: Top 50 Democratic and Republican words (Louisiana), according to LDA. Topic ownership is determined by the ratio of Democratic to Republican tokens in it (both weighted by the total number of tokens per party). The instances of each token type are then summed across all topics owned by the party.

	Democratic	Republican
Cities	15	17
Documents	10257	5859
Tokens	6101752	2310072
Token assignments	6006202	2259362
Topics	103	97

Table 11: Descriptive statistics for Indiana. "Tokens" describes the number of words in each party's documents, "token assignments" the tokens assigned to each party in the topic model depending on the ratio of Democratic to Republican tokens in it (both weighted by the total number of tokens per party).

	Democratic	Republican
Cities	11	7
Documents	6287	1327
Tokens	1955198	322915
Token assignments	1789373	314628
Topics	143	57

Table 12: Descriptive statistics for Louisiana. "Tokens" describes the number of words in each party's documents, "token assignments" the tokens assigned to each party in the topic model depending on the ratio of Democratic to Republican tokens in it (both weighted by the total number of tokens per party).

Figure 17: Total number of lines retained at a given threshold for removing duplicated lines. For example, at x = 10, all lines occurring more than 10 times within a city's documents are removed.

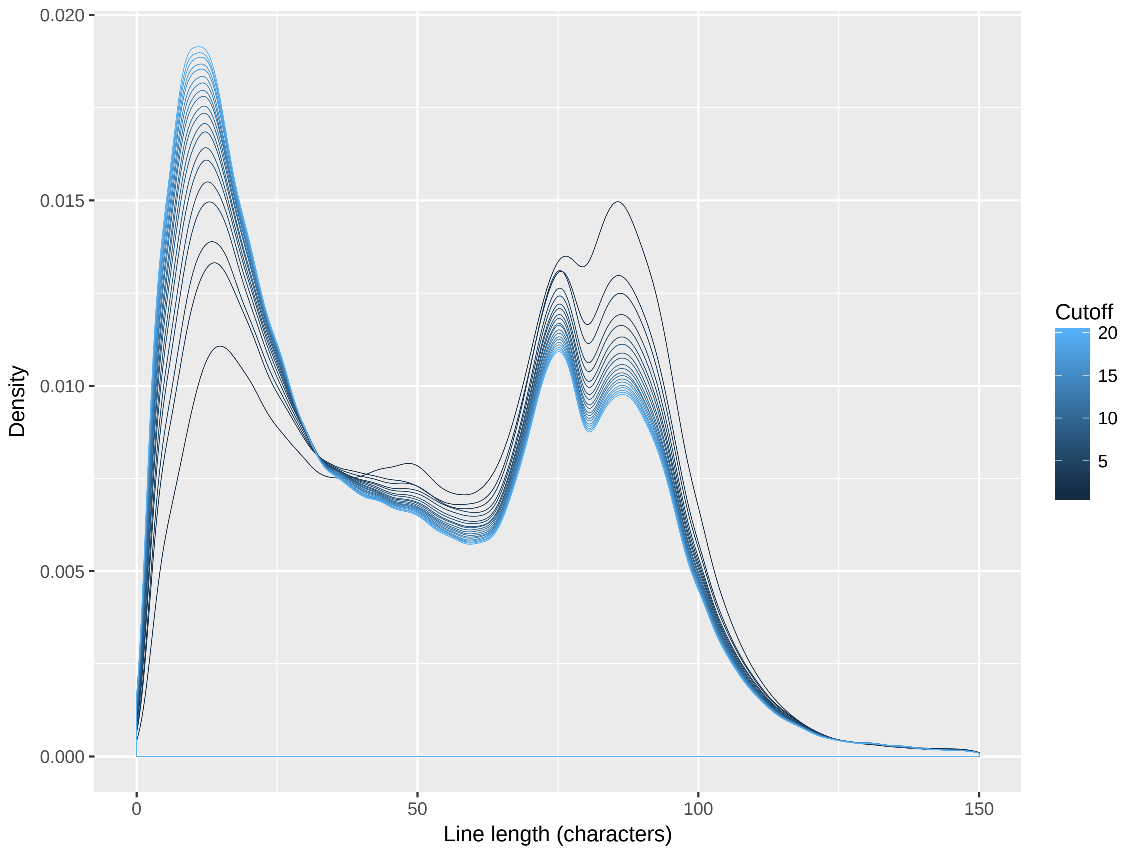


Figure 18: Hierarchical clustering (Indiana).

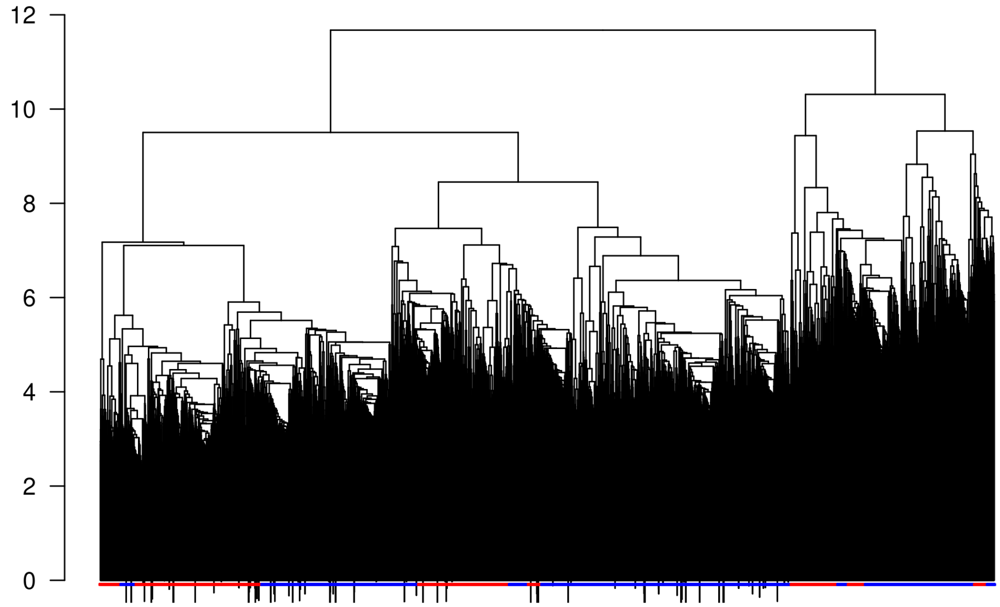


Figure 19: Hierarchical clustering (Louisiana).

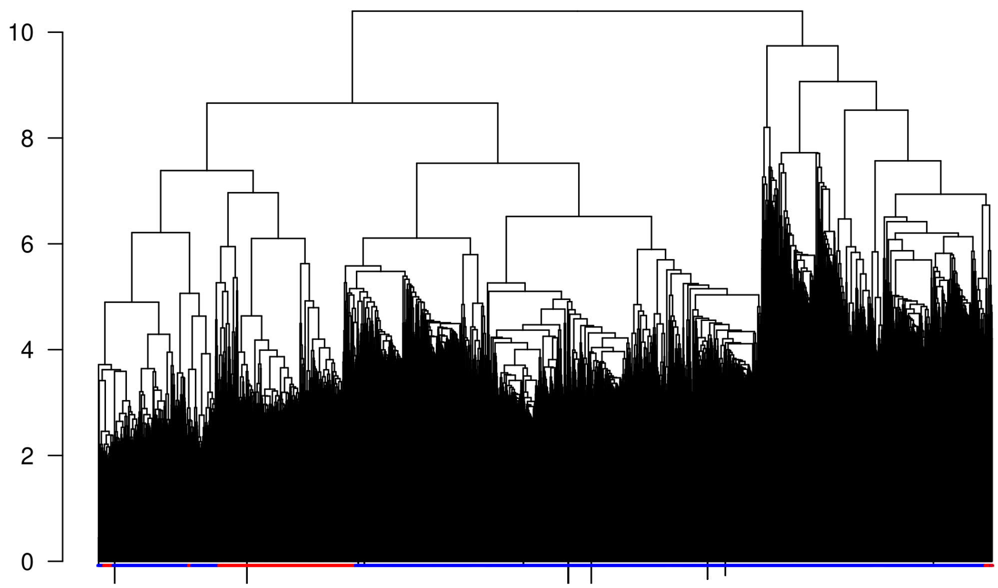


Figure 20: Entropy (between cities, for each topic).

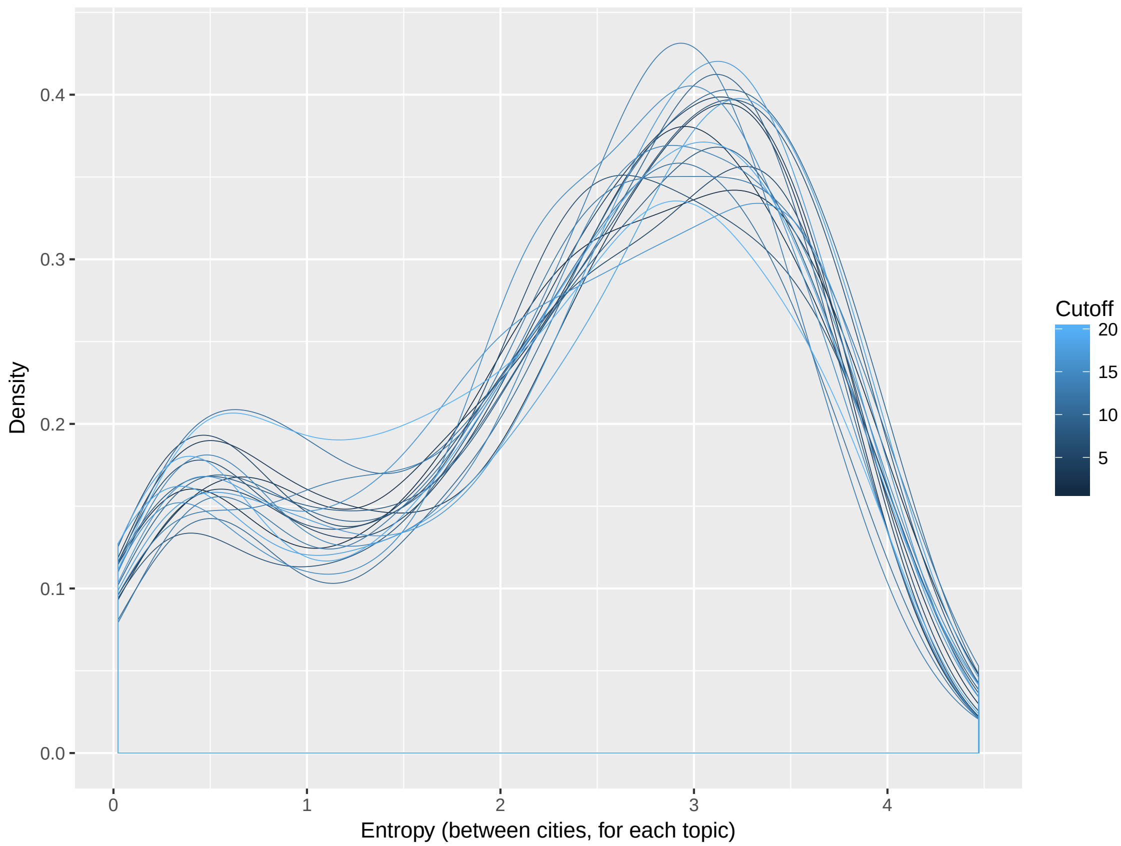


Figure 21: Weighted (multiplied by tokens assigned to the topic) entropy (between cities, for each topic).

Figure 22: Log number of city documents per cluster. Brighter tile colors indicate more documents. The color of the tile border indicates the city's party.

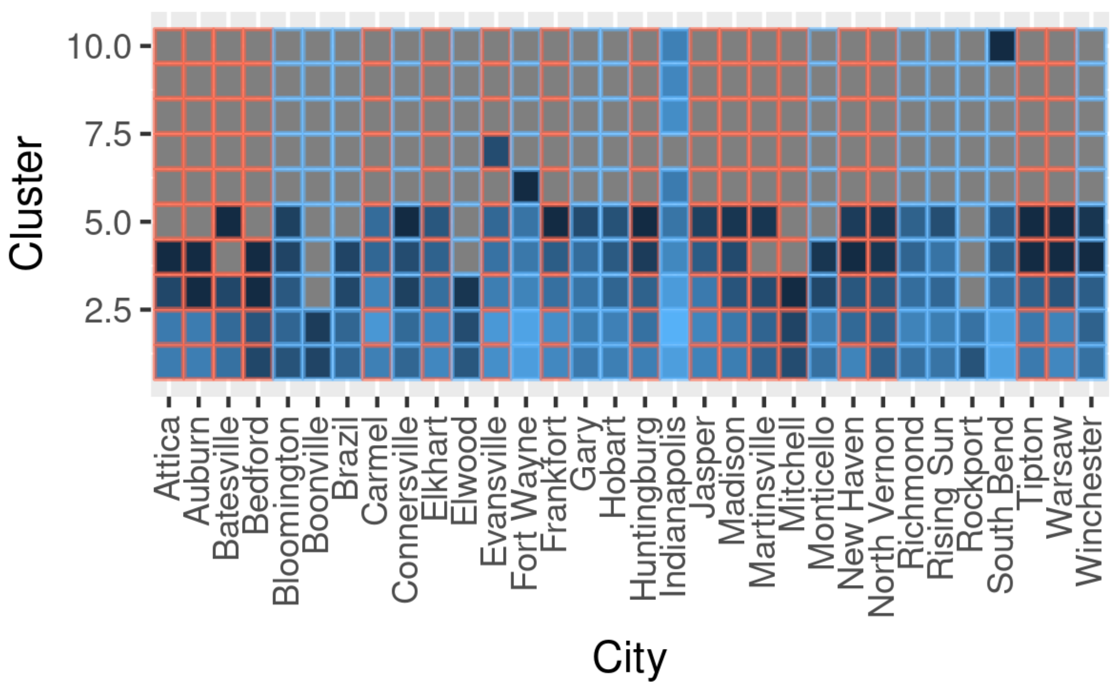


Figure 23: Log number of city documents per cluster. Brighter tile colors indicate more documents. The color of the tile border indicates the city's party.

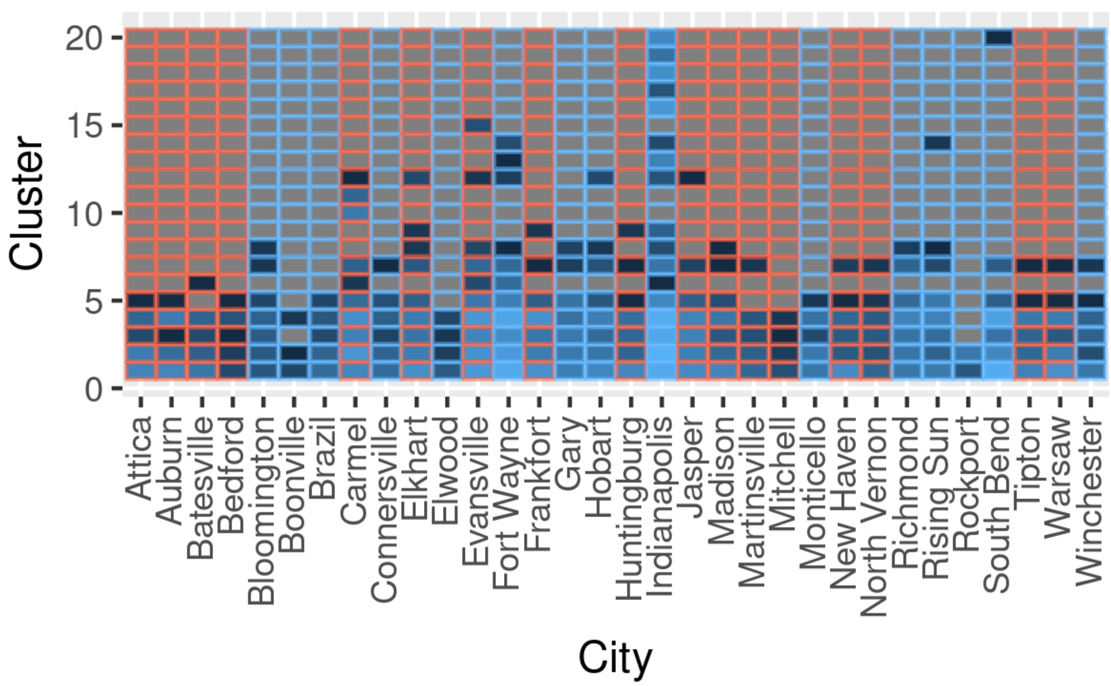


Figure 24: Log number of city documents per cluster. Brighter tile colors indicate more documents. The color of the tile border indicates the city's party.

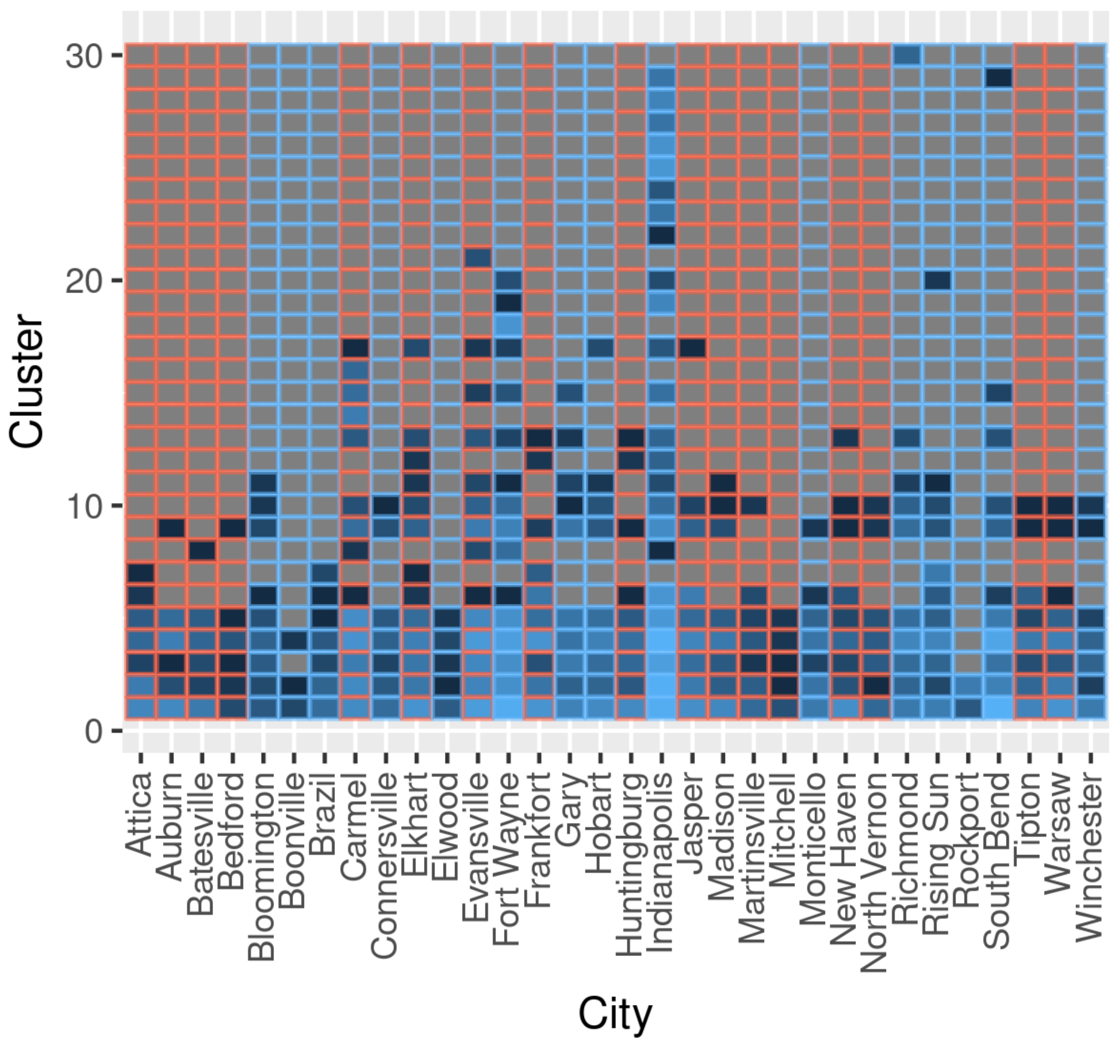


Figure 25: Log number of city documents per cluster. Brighter tile colors indicate more documents. The color of the tile border indicates the city's party.

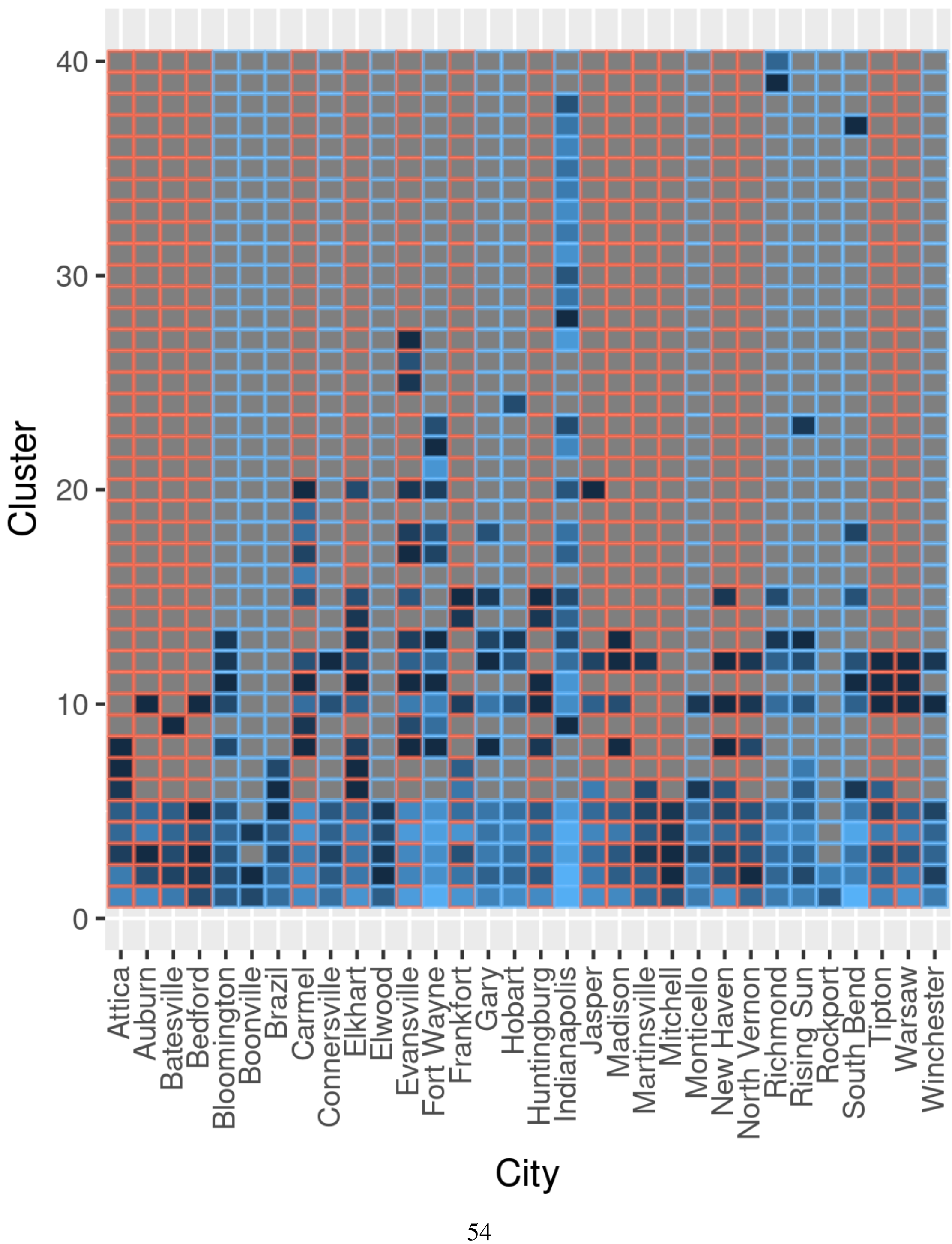


Figure 26: Log number of city documents per cluster. Brighter tile colors indicate more documents. The color of the tile border indicates the city's party.

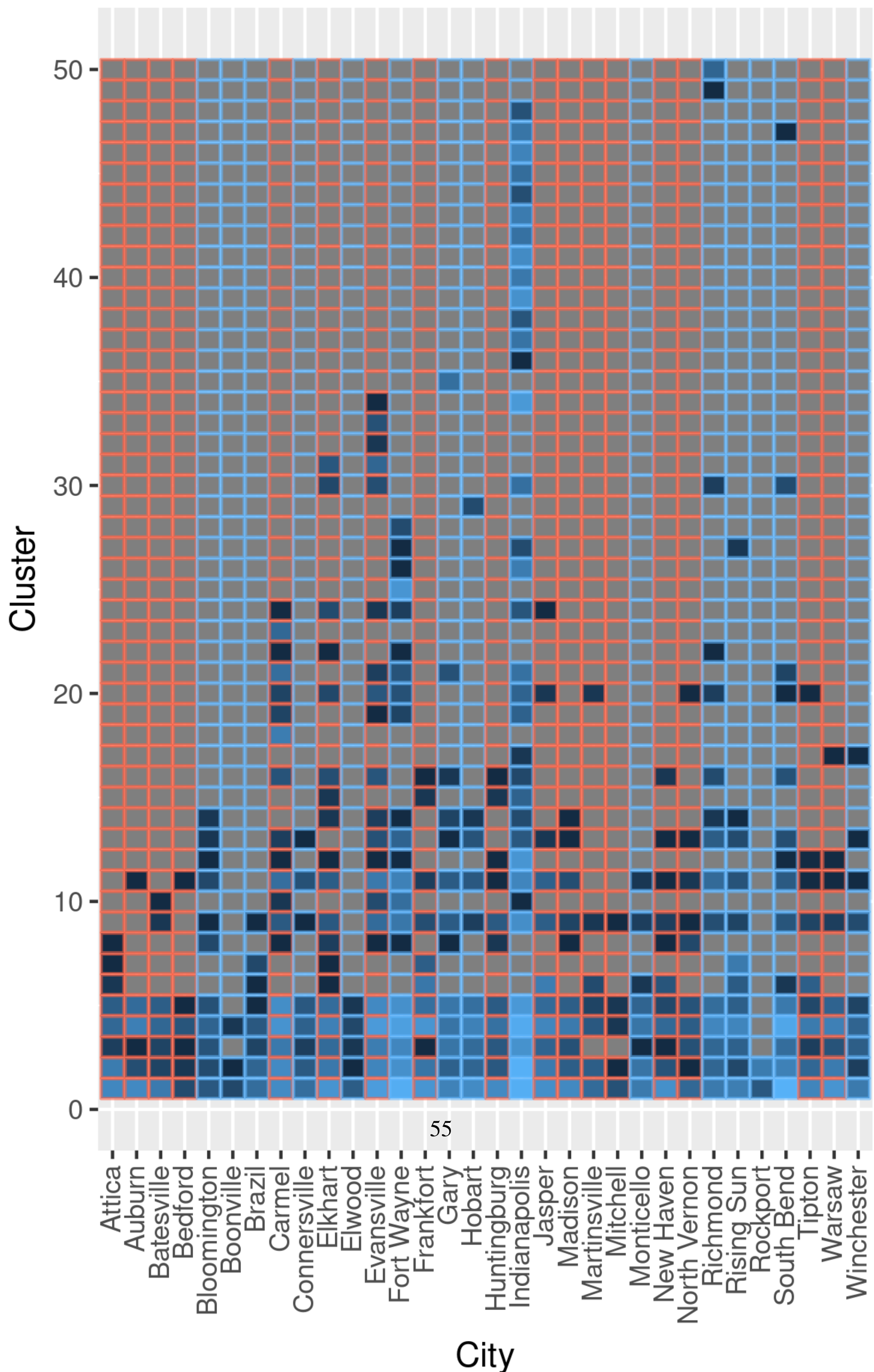


Figure 27: Log number of city documents per cluster. Brighter tile colors indicate more documents. The color of the tile border indicates the city's party.