Appendix: Automated Estimates of State Interest Group Lobbying Populations

Alex Garlick
Assistant Professor of Political Science, The College of New Jersey, email: garlicka@tcnj.edu
John Cluverius
Assistant Professor of Political Science, University of Massachusetts Lowell, email: John_Cluverius@uml.edu

This appendix accompanies the article published in *Interest Groups & Advocacy*. First, it describes the data collection process for this project. Next, it relates the code books from the different data sources. Then it has details on the different iterations of the automated method, and reports the results from each iteration. It also has an example of the bootstrap procedure described in the main text. Finally, it lists the stop words that are omitted from the bags of words.

Data collection

The first layer of data for this project is a census of the businesses, non-profits and governments that hire lobbyists in all 50 states from the National Institute for Money in State Politics (NIMSP). The NIMSP data is available to the public on its website,[1] although we were provided with the underlying data through 2017.[2] There are 2.07 million observations between 2000 (although national coverage picks up in 2006) through 2017 in this list.

Table 6 describes how the data sources used to build the estimates described in Section 2 were acquired. Table 6 shows the number of unique group registrations in each data source over time.

Table 6: Data sources for lobbying group registrations

Venue	Years	Source	No. of Codes
Colorado	2005-2014	CD-ROM from Sec. of State, also at: sos.state.co.us/lobby/Home.do	31
Massachusetts	2005-2016	Scraped from Sec. of State sec.state.ma.us/LobbyistPublicSearch/	26
Pennsylvania	2007-2018	Downloaded from Dept. of State palobbyingservices.state.pa.us/	42
All 50 States	2007	Gray and Lowery Team (GL 2007)	26
All 50 States	2005-2017	ZIP file from NIMSP, also at: followthemoney.org/lobbyist-link	350

Mapping subject codes

The GL 2007 codes assign each organization to one of 26 sectors (Lowery et al., 2012, p.596). The last column of Table 6 shows how many industry codes are used in the different states. Tables

[1] e.g. Illinois 2011 directory followthemoney.org/lobbyist-link?s=IL&y=2011

[2] We thank Peter Quist and Denise Roth Barber at NIMSP for supplying their source files that is surfaced at classic.followthemoney.org/database/search.phtml.

Table 7: Group registrations by year: 2005-2018

Year	Colorado	Massachusetts	Pennsylvania*	GL 2007	NIMSP
2005	1,093	1,036			17,640
2006	1,130	1,046			48,612
2007	1,150	1,067		51,061	50,093
2008	1,173	1,138	426		50,123
2009	1,200	1,221			46,857
2010	1,167	1,352	471		47,068
2011	1,280	1,355			49,255
2012	1,245	1,360	493		42,539
2013	1,336	1,406			54,941
2014	1,231	1,440	405		52,594
2015		1,437			57,920
2016		1,368	417		56,390
2017					54,161
2018			1,872		
Total	12,005	12,077	4,082	51,061	628,733
Average	1,201	1,006	681	51,061	48,364

*Penn. only reports groups in their latest lobbying period, therefore groups are observed once.

8 & 9 align those codes to the GL 2007 codes. In many cases the alignment is clear: 'Health care" in Pennsylvania corresponds to the GL 2007 'Health" code. However, these pathways can be muddled. Therefore, we use the Policy Agenda Project's codebook as a crosswalk, specifically the version written for the Pennsylvania Policy Agendas Project (McLaughlin et al., 2010)[3] that is sensitive to state-level issues, and include the appropriate major or minor codes in the second column of Table 8.

There are two important features of the subject codes to note in Tables 8 & 9. First, there are gaps. For example, Massachusetts does not have a category for Agriculture. This does not mean there are no dairy farmers in Massachusetts, just that the state apparently did not think there were enough to populate a category; as a result, they probably file as a business association and organization. Second, the bottom of Table 8 lists all of the unused subject categories for the legislative venues. There is a general 'business' category in all of these states, however, the GL 2007 coding scheme does not have such a broad category, so these extra-broad delineations are dropped. The 350 NIMSP codes were mapped to the PAP scheme using the GL 2007 coding scheme (Lowery et al., 2012).

Table 10 relates the GL 2007 codes to Holyoke (2019). Holyoke puts groups in 17 economic sectors, which do not neatly align with the GL codes. As a result, there is a higher degree of measurement error when comparing the aggregate totals from this scheme than when the automated estimates are compared to the GL estimates (in Figure 1).

[3]This version of the Comparative Agendas Project is hosted by Temple University: http://www.cla.temple.edu/papolicy/.

Table 8: Codebook for Policy Agendas Project, GL 2007 codes, Penn., Colorado

No.	(1) PAP	(2) GL 2007	(3) Pennsylvania	(4) Colorado
1	201	Civil Rights		
2	202	Women	Women's/Reproductive Issues	
3	205	Environment	Environment	
4	207	Religion	Religious	Clergy/Faith-based
5	208	Tax Policy		
6	209	Good Government	Public Interest Firearms	Political Organizations
7	300	Health	Health Care Mental Health	Healthcare/Medical
8	400	Agriculture	Agriculture Food Processing/Sales Forest Products	Agriculture
9	600	Education	Education	Education
10	701	Utilities	Utilities	Utilities
11	702	Natural Resource	Natural Resources Energy	Energy Mining
12	1000	Transportation	Motor Vehicle Transportation	Automotive Transportation
13	1200	Law	Legal	Attorney/Legal
14	1300	Welfare	Human Services	
15	1400	Construction	Construction	Construction/Engineering
16	1500	Bank	Banking/Finance Real Estate	Real Estate Financial/Investment
17	1501	Hotel	Food Service Tourism Alcoholic Beverages	Food Services Entertainment/Recreation
18	1502	Small Business	Retail Sales Commerce	Merchandise/Retail
19	1503	Sports	Wagering/Gaming Recreation/Entertainment	Gaming
20	1504	Business Services	Waste Management Accounting Information Technology	Consultant Environmental Services
21	1510	Insurance	Insurance	Insurance
22	1520	Manufacturing	Industry/Manufacturing Biotechnology	Manufacturing
23	1600	Military		Military
24	1700	Communication	Media Telecommunications	Media/Public Relations Science/Technology
25	2400	Local Government	Government	Government/Civil
26	2401	Police and Fire		Firefighters/Paramedics Law Enforcement
	9999	Uncategorized	Business Labor Union Tobacco Workers' Compensation Other	General Business Labor Unions

Table 9: Codebook for Massachusetts to GL 2007 codes

No.	Massachusetts Term	GL 2007
1	Automotive Industry	Transportation
2	Banking: Lending, Investment	Bank
3	Business Associations and Organizations	Small Business
4	Contractors and Subcontractors	Construction
5	Education: Institutions, Services, Programs	Education
6	Energy: Petroleum, Hydro, Nuclear, Oil	Natural Resource
7	Environment: Recycling, Packaging, Pollution	Business Services or Environment
8	Food & Beverage: Industry, Services	Hotel
9	Gaming: Casinos, Gambling	Sports
10	Government: Agencies, Associations, Organizations	Local Government
11	High Technology	Business Services
12	Hospitals: Health Care Systems, Medical Organizations	Health
13	Human Services	Welfare
14	Insurance: Auto, Home, Life, Other	Insurance
15	Insurance: Medical, Dental, Mental Health	Health
16	Legal Organizations and Services	Law
17	Pharmaceutical Industry	Manufacturing
18	Police, Fire, Law Enforcement Organizations	Police and Fire
19	Real Estate: Development and Property	Bank
20	Tobacco and Alcohol	Agriculture
21	Transportation: Air, Sea, Land, Rail	Transportation
22	Travel, Tourism, Entertainment	Sports
23	Utilities and Telecommunications: Gas, Electric, Cable, Telephone	Utilities or Communications
XX	*Unused categories*	
	Labor Unions, PACs	Civil Rights
	Lobbying Organizations	Good Government
	Other	Religion
		Tax Policy
		Uncategorized
		Women

Table 10: Codebook for GL 2007 to Holyoke 2019 codes

No.	PAP	GL (2007)	Holyoke (2019)
1	200	Civil Rights Women Environment Religion Tax Policy Good Government	Ideological
2	300	Health	Health
3	400	Agriculture	Agriculture
4	600	Education	Education
5	700	Utilities Natural Resource	Energy
6	1000	Transportation	Transportation
7	1200	Law	Lawyers
8	1300	Welfare	Social services
9	1400	Construction	Construction Real Estate
10	1500	Bank Hotel Small Business Sports Business Services	Finance Entertainment Leisure Business
11	1520	Manufacturing	Electronics
12	1700	Communication	Comm
13	2400	Local Government Police and Fire	Government
	Uncoded	Insurance Military Unknown	

Testing the Method

This section tests the estimates produced the method, specifically evaluating 36 iterations of the estimation procedure on two different sets of testing data. First, I will note the alternative specifications used to create the myriad iterations.

1. Create 6 sets of 'bags of words' for each policy sector S, by observing the full distribution of words with three different source materials.
   (a) Use the GL 2007 hand coded list as source material.
   (b) Use the GL 2007 hand coded list as source material, with the stop words in Table 15-18 removed.
   (c) Use the 2006-2018 Colorado and Pennsylvania lobbying registrations as source material.
   (d) Use the 2006-2018 Colorado and Pennsylvania lobbying registrations, with the stop words in Table 15-18 removed.
   (e) Use the GL 2007 hand coded list and 2006-2018 Colorado and Pennsylvania lobbying registrations as source material.
   (f) Use the GL 2007 hand coded list and 2006-2018 Colorado and Pennsylvania lobbying registrations as source material., with the stop words in Table 15-18 removed. The posterior probabilities built off these 'bags of words' are listed in Table 11.
2. There are no alternative specifications for estimating the posterior probabilities, or the likelihood of a word falling in a policy sector. Of note, the posterior probabilities are defined for each set of 'bags of words' listed in step 1.
3. There are two different equations to estimate the conditional probability Pc that each word k in a document fits each policy sector.
   (a) The standard method takes the raw count of words in each bag.

$$P_{c,k} = \frac{N_{c,k}}{N_s} \tag{1}$$

   (b) The *Laplace* method conducts a Laplace smoothing procedure by adding one for each quantity, this eliminates zeroes (so that logarithms may be taken).

$$L_{c,k} = \frac{(N_{c,k} + 1)}{N_s} \tag{2}$$

4. There are three different equations to combine the conditional probabilities of the words in each document.
   (a) Products:

$$P_c = \prod_{i=1}^{k} P_{c,k} \tag{3}$$

(b) Log: Note that this equation uses the Laplace smoothing procedure for each word.

$$P_c = \sum_{i=1}^{k} \log(L_{c,k}) \tag{4}$$

(c) Means

$$P_c = \frac{\sum_{i=1}^{k} P_{c,k}}{k} \tag{5}$$

5. There is only one method to estimate the Bayes probability of each class by multiplying the conditional probability for each policy sector by the sector's posterior probability.

$$B_s = P_c \times P_s \tag{6}$$

6. There are two different types of bootstraps used to randomly draw a single class for each document. After sorting the Bayes estimates in a descending fashion for each document that has $1 >= k \leq 26$ non-zero Bayes estimates, the methods vary in V, or the number of potential classes they keep. They also vary in how much the Bayes estimates are weighted in the random draw procedure. See Table 12 for examples of both.

   (a) Coinflip Bootstrap: For $0 > V \leq 2$, inclusive of ties, each bootstrap draws an unequal probability sample, proportional to B_s.

   (b) Weighted Bootstrap: For $0 > V \leq 5$, inclusive of ties, each bootstrap draws an unequal probability sample, proportional to ${B_s}^2$.

Example Bootstrap

Table 12 demonstrates the bootstrap procedure for the document estimated in Table **??**. Column (1) shows an example of a 'coinflip' bootstrap, while column (2) shows a weighted bootstrap. Table 12 shows the Bayes estimates produced by the 'Log" estimation procedure to demonstrate a document with a large number of potential policy areas is reduced down by the bootstrap procedures.

Optimizing the algorithm

A researcher could argue for any one of 36 iterations above. This section evaluates each iteration twice, first on a test set of the codes provided in the NIMSP data. While there is no pattern to the missingness in these data, there is no reason to believe the codes that are present are biased, therefore they provide a reasonable set to test estimates produced by the automated procedure. It also uses data from Massachusetts. As a reminder, Schütze, Manning and Raghavan (2008, p. 154-155) suggest that these methods should be judged by its precision and retrieval. Precision is the share of estimations that are correct. Recall is the share of possibly relevant documents that are retrieved. They suggest balancing these two criteria by calculating an F1-score, which is the weighted harmonic mean of precision and recall show in equation **??**. where P is precision, and R is recall.

Table 11: Posterior probabilities for the 26 policy areas using stop words

	Policy Sector	Words in bag	Posterior Probability
1	Health	15,389	0.133
2	Education	11,251	0.097
3	Banking	9,274	0.080
4	Manufacturing	8,132	0.070
5	Government	8,102	0.070
6	Business Services	7,533	0.065
7	Small Business	5,573	0.048
8	Legal	4,786	0.041
9	Construction	4,672	0.040
10	Insurance	4,520	0.039
11	Utilities	4,366	0.038
12	Sports/Recreation	4,041	0.035
13	Police/Fire	3,636	0.031
14	Social Welfare	3,209	0.028
15	Good Government	3,023	0.026
16	Transportation	2,950	0.026
17	Agriculture	2,902	0.025
18	Natural Resources	2,892	0.025
19	Communications	2,715	0.023
20	Hotel/Entertainment	2,099	0.018
21	Environment	1,908	0.017
22	Civil Rights	945	0.008
23	Religious	710	0.006
24	Women's Groups	379	0.003
25	Tax	344	0.003
26	Military	182	0.002
	Total	**115,533**	

Source: Hand codes, CO, PA

Table 12: Example of weighted and coinflip bootstraps for '1st Farm Credit Services" calculated using Log estimation

No. (n)	Rank	Sector	(1) Coinflip Bootstrap Bayes Est.($\times 10^{-4}$)	In	Pred.	(2) Weighted Bootstrap In	Bayes^2($\times 10^{-8}$)	Pred.
1	1	Banking	0.0848	✓	0.59	✓	0.0719	0.49
2	2	Agriculture	0.0591	✓	0.41	✓	0.0349	0.24
3	3	Insurance	0.0496			✓	0.0246	0.17
4	4	Health	0.0259			✓	0.0067	0.046
5	5	Government	0.0250			✓	0.0062	0.043
6	6	Education	0.0215					
7	7	Good Gov't	0.0155					
8	8	Manufacturing	0.0139					
9	9	Social Welfare	0.0120					
10	9	Small Business	0.0120					
11	11	Business Services	0.0120					
12	12	Sports/Rec.	0.0095					
13	13	Legal	0.0060					
14	13	Tax	0.0060					
15	13	Environment	0.0060					
16	13	Communications	0.0060					
17	13	Utilities	0.0060					
18	13	Hotel/Ent.	0.0060					

Notes: (1) (2) (3) (4) (5)

(1) The top two Bayes estimates (with ties) are kept.

(2) STATA randomly selects a sector using the Bayes estimate as an analytic weight.

(3) Up to 5 estimates are kept, inclusive of ties.

(4) The Bayes estimate is squared ($B_s{}^2$).

(5) STATA randomly selects a sector using the squared Bayes estimate as an analytic weight.

Tables 13 and 14 show the performance of each iteration of the model on the two out-of-sample tests. The procedure used in the main text had the highest average F1-score across the two models. It was the second highest performing model under both tests. It used both the GL 2007 codes, as well as the Colorado and Pennsylvania lobbying registrations as the source for the bags of words and posterior probabilities. It also used the 1557 'stop words' and used the 'weighted' bootstrap method. Each of these methods was tested over 50 bootstraps.

The F1 levels shows in Table 14 are lower than in Table 13; however, this discrepancy is due to the lack of alignment between the number of categories Massachusetts offers groups (23) and the 26 hand-codes. There were a number of uncoded groups in Table 9 that do not map onto the Massachusetts coding scheme, and some of the codes that do match are not exhaustive.

List of Stop Words

Tables 15, 16, 17, and 18, list the 1557 stop words that were excluded from the bag of words calculation. This list was created by highlighting words that were prominent in multiple bags but did not differentiate the type of groups effectively.

References

Holyoke, Thomas T. 2019. "Dynamic state interest group systems: a new look with new data." *Interest Groups & Advocacy* pp. 1–20.

Lowery, David, Virginia Gray, John Cluverius and Jeffrey J Harden. 2012. "Explaining the anomalous growth of public sector lobbying in the American States, 1997–2007." *Publius: The Journal of Federalism* 43(4):580–599.

McLaughlin, Joseph P, Paul Wolfgang, J Wesley Leckrone, Justin Gollob, Jason Bossie, Jay Jennings and Michelle J Atherton. 2010. "The Pennsylvania policy database project: A model for comparative analysis." *State Politics and Policy Quarterly* 10(3):320–336.

Schütze, Hinrich, Christopher D. Manning and Prabhakar Raghavan. 2008. *Introduction to Information Retrieval*. Cambridge Univ Press.

Table 13: Group estimation validations: NIMSP

Rank	F1	Precision	Recall	Matches	Guesses	Stops	Source	Method	Coin
1	0.70	0.74	0.67	12,690.3	17,174	✓	H	Prod.	
2	0.70	0.73	0.67	12,739.8	17,334	✓	H, CO, PA	Prod.	
3	0.70	0.71	0.68	12,891.9	18,050		H, CO, PA	Prod.	
4	0.69	0.72	0.67	12,799.6	17,872		H	Prod.	
5	0.68	0.72	0.65	12,347.5	17,174	✓	H	Prod.	✓
6	0.68	0.70	0.66	12,627.5	18,050		H, CO, PA	Prod.	✓
7	0.68	0.71	0.65	12,375.6	17,334	✓	H, CO, PA	Prod.	✓
8	0.68	0.70	0.66	12,556.2	17,872		H	Prod.	✓
9	0.64	0.66	0.63	11,969.9	18,104	✓	H	Mean	
10	0.64	0.66	0.63	11,988.4	18,182	✓	H, CO, PA	Mean	
11	0.61	0.63	0.60	11,398.8	18,104	✓	H	Mean	✓
12	0.61	0.63	0.60	11,384.9	18,182	✓	H, CO, PA	Mean	✓
13	0.54	0.55	0.53	10,046.3	18,104	✓	H	Log	✓
14	0.54	0.55	0.53	10,025.1	18,182	✓	H, CO, PA	Log	✓
15	0.50	0.51	0.48	9,210.2	18,104	✓	H	Log	
16	0.49	0.50	0.48	9,137.9	18,182	✓	H, CO, PA	Log	
17	0.47	0.52	0.42	8,020.5	15,292	✓	CO, PA	Mean	
18	0.45	0.45	0.44	8,424.5	18,818		H	Log	✓
19	0.44	0.50	0.40	7,633.6	15,292	✓	CO, PA	Mean	✓
20	0.44	0.45	0.44	8,398.4	18,833		H, CO, PA	Log	✓
21	0.44	0.56	0.36	6,820.6	12,217	✓	CO, PA	Prod.	
22	0.42	0.54	0.35	6,572.4	12,217	✓	CO, PA	Prod.	✓
23	0.41	0.46	0.37	6,978.1	15,292	✓	CO, PA	Log	
24	0.39	0.44	0.36	6,767.9	15,292	✓	CO, PA	Log	✓
25	0.38	0.46	0.32	6,072.3	13,279		CO, PA	Prod.	
26	0.37	0.45	0.31	5,919.2	13,279		CO, PA	Prod.	✓
27	0.34	0.35	0.34	6,502.3	18,818		H	Mean	✓
28	0.34	0.34	0.34	6,485.4	18,833		H, CO, PA	Mean	✓
29	0.34	0.34	0.34	6,459.9	18,818		H	Mean	
30	0.34	0.34	0.34	6,455.4	18,833		H, CO, PA	Mean	
31	0.32	0.32	0.32	6,109.2	18,818		H	Log	
32	0.32	0.32	0.32	6,067.9	18,833		H, CO, PA	Log	
33	0.29	0.30	0.28	5,405.6	18,083		CO, PA	Log	✓
34	0.27	0.27	0.26	4,950.3	18,083		CO, PA	Mean	
35	0.25	0.26	0.25	4,665.2	18,083		CO, PA	Mean	✓
36	0.25	0.26	0.25	4,661.4	18,083		CO, PA	Log	
Notes				(1)	(2)	(3)	(4)	(5)	(6)

(1) Average matches from 50 bootstrap draws.

(2) There are 19,019 unique group names in the NIMSP database.

(3) Uses the 1557 stop words in Table 15-18.

(4) Source for bags of words. Hand-coded (H), Colorado, Pennsylvania.

(5) Methods: Log, Products, Means.

(6) Bootstrap type: Coin flip or weighted.

Table 14: Group estimation validations: Mass

Rank	F1	Precision	Recall	Matches	Guesses	Stops	Source	Method	Coin
1	0.35	0.38	0.32	1,671.3	4,375	✓	H, CO, PA	Mean	
2	0.35	0.40	0.31	1,593.1	4,000	✓	H, CO, PA	Products	
3	0.34	0.38	0.31	1,634.8	4,346	✓	H	Mean	
4	0.34	0.39	0.30	1,578.1	4,020	✓	H, CO, PA	Products	✓
5	0.34	0.38	0.31	1,621.5	4,283		H, CO, PA	Products	✓
6	0.34	0.37	0.31	1,632.9	4,375	✓	H, CO, PA	Mean	✓
7	0.34	0.39	0.30	1,560.3	4,030		H, CO, PA	Products	
8	0.33	0.37	0.31	1,596.4	4,346	✓	H	Mean	✓
9	0.33	0.39	0.29	1,516.7	3,886	✓	H	Products	
10	0.33	0.40	0.29	1,498.4	3,793	✓	CO, PA	Mean	
11	0.33	0.38	0.29	1,497.5	3,902	✓	H	Products	✓
12	0.32	0.38	0.27	1,434.5	3,793	✓	CO, PA	Mean	✓
13	0.32	0.36	0.28	1,464.9	4,059		H	Products	✓
14	0.31	0.37	0.27	1,424.2	3,863		H	Products	
15	0.31	0.42	0.24	1,248.1	2,947	✓	CO, PA	Products	
16	0.30	0.33	0.28	1,459.1	4,382	✓	H, CO, PA	Log	✓
17	0.30	0.42	0.23	1,223.7	2,947	✓	CO, PA	Products	✓
18	0.30	0.33	0.27	1,420.3	4,351	✓	H	Log	✓
19	0.29	0.35	0.25	1,326.4	3,796	✓	CO, PA	Log	
20	0.29	0.35	0.25	1,321.3	3,796	✓	CO, PA	Log	✓
21	0.29	0.37	0.24	1,238.7	3,338		CO, PA	Products	
22	0.29	0.36	0.23	1,225.7	3,360		CO, PA	Products	✓
23	0.27	0.30	0.25	1,309.9	4,382	✓	H, CO, PA	Log	
24	0.27	0.30	0.25	1,286.3	4,351	✓	H	Log	
25	0.26	0.28	0.25	1,289.1	4,672		H, CO, PA	Log	✓
26	0.26	0.27	0.24	1,270.9	4,652		H	Log	✓
27	0.24	0.25	0.22	1,151.2	4,537		CO, PA	Log	✓
28	0.22	0.23	0.21	1,089.8	4,661		H	Mean	✓
29	0.22	0.23	0.21	1,083.0	4,670		H, CO, PA	Mean	✓
30	0.22	0.23	0.21	1,071.7	4,661		H	Mean	
31	0.21	0.23	0.20	1,056.6	4,670		H, CO, PA	Mean	
32	0.21	0.23	0.20	1,021.8	4,538		CO, PA	Mean	✓
33	0.20	0.22	0.19	999.5	4,538		CO, PA	Mean	
34	0.20	0.21	0.19	997.3	4,672		H, CO, PA	Log	
35	0.20	0.21	0.19	989.4	4,652		H	Log	
36	0.20	0.21	0.19	967.4	4,537		CO, PA	Log	
Notes				(1)	(2)	(3)	(4)	(5)	(6)

(1) Average matches from 50 bootstrap draws.

(2) There are 5,219 unique group names in the Mass. database.

(3) There are 1557 stop words in Table 15-18.

(4) Source for bags of words. Hand-coded (H), Colorado, Pennsylvania.

(5) Methods: Log, Products, Means.

(6) Bootstrap type: Coin flip or weighted.

Table 15: 1557 Stop Words: Part 1

0	52	109	169	246	344	465	565	700	876	1091	1518	2002	3000	6000	57048
1	54	110	170	247	345	469	566	701	880	1099	1521	2004	3001	6062	75202
2	55	111	172	249	346	470	567	706	881	1100	1522	2005	3005	6083	95014
3	56	112	174	250	350	472	568	707	883	1101	1526	2006	3030	6200	101286
4	57	113	175	251	351	473	569	709	884	1104	1547	2007	3080	6206	105566
5	58	114	176	252	357	475	570	710	888	1107	1600	2008	3119	6264	106450
6	59	115	177	257	359	476	572	711	891	1109	1664	2009	3183	6365	112211
7	60	116	180	259	360	478	576	712	893	1112	1707	2010	3203	6488	122011
8	61	117	183	261	361	480	580	714	900	1113	1739	2011	3210	6700	132011
9	62	118	184	264	365	481	583	716	903	1117	1742	2012	3215	7076	136140
10	63	119	185	265	367	484	589	718	908	1121	1752	2013	3299	7102	185225
11	64	120	186	268	368	486	595	719	909	1132	1765	2014	3393	7200	192011
12	65	121	187	269	369	489	597	720	911	1140	1771	2015	3400	7214	206210
13	66	122	188	272	371	490	600	728	912	1142	1776	2016	3419	7250	234238
14	67	123	190	273	372	491	601	731	916	1177	1792	2017	3470	7374	317483
15	68	124	191	277	374	495	605	732	918	1180	1800	2020	3484	7704	323433
16	69	125	192	279	375	499	609	743	920	1182	1805	2021	3500	7777	341363
17	70	126	193	280	376	500	611	751	921	1184	1808	2024	3530	7810	361363
18	71	127	194	281	380	501	612	757	922	1186	1812	2030	3560	8000	536540
19	72	128	195	282	381	502	613	760	924	1199	1818	2055	3570	8031	612011
20	73	129	198	284	384	503	614	766	925	1200	1819	2102	3621	8330	900097
21	74	130	199	285	385	508	615	768	932	1209	1827	2115	3631	8622	1800411
22	75	131	200	286	390	509	616	773	945	1211	1863	2116	3720	8858	2010548
23	76	132	201	287	391	511	620	775	949	1222	1868	2120	3721	8868	2037839
24	77	133	202	288	392	512	623	777	951	1234	1877	2130	3846	8884	8450797
25	79	134	203	290	396	513	624	779	959	1245	1889	2150	3895	9294	11811061
26	80	135	205	291	399	514	625	780	962	1250	1890	2151	3930	9400	21712200
27	81	136	206	292	400	515	630	786	965	1260	1899	2222	4000	9413	1234567890
28	82	137	208	293	401	518	631	788	967	1271	1901	2250	4034	9501	8550...*
29	83	139	209	296	402	521	632	789	970	1290	1908	2260	4041	9801	1420...**
30	84	140	210	298	403	522	633	792	974	1298	1920	2320	4100	10000	a
31	85	141	211	300	404	525	638	799	975	1303	1922	2323	4110	10402	aberdeen
32	86	142	212	301	405	526	640	802	980	1320	1926	2325	4121	10670	abilene
33	88	143	215	302	407	527	646	807	983	1325	1931	2327	4145	10717	ac
34	89	145	216	303	409	528	648	808	987	1348	1948	2350	4201	12003	advocacy
35	90	146	217	307	411	529	650	810	996	1350	1950	2401	4321	13214	affairs
36	91	148	223	308	412	530	655	811	998	1360	1957	2404	4345	14614	affiliates
37	92	149	224	311	415	535	659	813	1000	1400	1965	2428	4502	15013	agencies
38	93	150	225	317	416	537	668	815	1001	1404	1971	2498	4566	15145	ak
39	94	151	226	320	417	540	669	817	1005	1405	1975	2500	4600	16807	aka
40	95	153	227	321	419	542	675	825	1010	1408	1980	2507	4731	16813	akron
41	97	155	229	324	420	544	676	827	1013	1414	1984	2600	4900	16941	al
42	98	157	230	325	424	547	680	831	1017	1416	1985	2623	4951	18032	ala
44	99	158	233	326	426	548	682	846	1022	1422	1989	2630	4995	20091	alabama
45	100	160	234	328	429	549	683	850	1025	1436	1991	2650	5000	21925	alabamas
46	101	161	235	333	431	550	685	853	1029	1445	1992	2665	5280	24115	alaska
47	103	162	237	334	433	551	687	854	1031	1446	1994	2715	5285	30025	alaskaalaska
48	104	163	238	338	435	554	689	860	1032	1454	1995	2750	5657	32836	alaskan
49	105	164	241	340	444	555	690	863	1039	1457	1996	2857	5696	37083	alaskans
50	106	165	242	341	449	562	691	866	1055	1466	2000	2881	5732	39521	albany
51	108	167	244	343	453	563	692	872	1084	1500	2001	2907	5890	41424	albuquerque

Unabridged long codes: *85500000000000, **1420000000000000

Table 16: 1557 Stop Words: Part 2

alexandria	assocs	burlington	congress	district	gerogia
all	assoction	c	connecticut	district of columbia	gilbert
allentown	assof	ca	consulting	division	glendale
alliance	assoiation	calif	coral springs	downey	grand prairie
amarillo	assoicates	californa	corona	duluth	grand rapids
ameerican	assoication	california	corp	durham	grayslake
america	assoications	californias	corpo	e	green bay
american	assoiciation	calilfornia	corperation	east	greenbay
americas	assoociation	cambridge	corpopration	el monte	greensboro
amrican	assosciation	canton	corporacion	el paso	greenville
anaheim	assotiation	cape coral	corporate	elizabeth	group
anchorage	asspciates	carrollton	corporatio	elk grove	gulfport-biloxi
and	asssn	cary	corporatioin	elkhart	h
ann arbor	asssociates	cathedral city	corporation	erie	hagerstown
antioch	asssociation	cedar rapids	corporations	escondido	hampton
apple valley	at	center	corporationthe	eugene	harlingen
appleton	athens	central	corporatoin	evansville	harrisburg
ar	atlanta	champaign	corporaton	f	hartford
arizona	atlantic city	chandler	corportaion	fairfield	havre de grace
arizonas	augusta	chapter	corportate	family	hawaii
arkanas	aurora	charleston	corportation	fargo	hawaiian
arkansa	austin	charlotte	corportations	fayetteville	hayward
arkansans	authority	chattanooga	corportion	federation	hemet
arkansas	az	chesapeake	corpration	first	henderson
arlington	b	chicago	corproation	fitchburg	hesperia
arvada	bakersfield	chula vista	corps	fl	hi
asheville	baltimore	cincinnati	corpus christi	flint	hialeah
assn	barnstable	clarke county	costa mesa	florida	hickory
associated	baton rouge	clarksville	council	floridal	high point
associates	beaumont	clearwater	creek	fontana	holdings
associatess	bel air	cleveland	cross	for	hollywood
associati	bellevue	clubs	ct	fort collins	home
associaties	berkeley	co	d	fort lauderdale	honolulu
associatiion	bethlehem	coalition	dallas	fort smith	houma
associatin	billings	college station	daly city	fort walton beach	houston
associatio	birmingham	colo	danbury	fort wayne	howell
association	bloomington	coloardo	davenport	fort worth	huntington
associations	board	coloradans	davidson county	foundation	huntington beach
associatioon	boise	colorade	dayton	frederick	huntsville
associatition	boise city	colorado	daytona beach	fremont	i
associatoin	bonita springs	colorado springs	dc	fresno	ia
associaton	boston	columbia	de	fullerton	id
associatted	boulder	columbus	delaware	g	idaho
assocication	bradenton	community	delawareans	ga	idahoan
associtation	bremerton	compamy	delawaremaryland	gainesville	il
associtition	bridgeport	compan	delawaremd	garden grove	illiana
associtiona	brighton	companies	deltona	garland	illinios
assoclp	brownsville	companu	denton	gastonia	illinois
assocmmta	bryan	company	denver	general	in
assocof	buffalo	companys	des moines	georgia	inc
assocoiates	burbank	concord	detroit	georgian	incorporated

Table 17: 1557 Stop Words: Part 3

independence	las cruces	miami	new haven	one
independent	las vegas	michigan	new jersey	ontario
indiana	lawrence	milwaukee	new london	or
indianapolis	layton	minneapolis	new mexico	orange
inglewood	league	minnesota	new orleans	oregon
institute	leominster	minnesotagt	new york	oregonians
international	lewisville	miramar	new york	oregonlng
intl	lexington	mission viejo	new york city	oregons
into	lincoln	mississisippi	newark	orem
iowa	little rock	mississipi	newburgh	orlando
irvine	llc	mississippi	newport news	overland park
irving	long beach	mississippians	nh	owners
its	lorain	missouri	nj	oxnard
j	los angeles	missourians	nm	p
jackson	louisia	missouris	norfolk	pa
jacksonville	louisian	mn	normal	palm bay
jefferson	louisiana	mo	norman	palm springs
jersey city	louisville	mobile	north	palmdale
johnson city	lousiana	modesto	north carolina	panama city
joliet	lousisiana	monroe	north charleston	partners
jr	lowell	montana	north dakota	pasadena
junior	lp	montano	north las vegas	paterson
k	lubbock	monterey	north port	pc
kailua	m	montgomery	northeast	pembroke pines
kalamazoo	m	moreno valley	northeastern	penn
kaneohe	ma	mountain	northwest	pennsylvania
kansas	macon	mountains	northwestern	pennsylvanians
kansas city	madison	ms	norwalk	pennsylvanias
kennewick	maine	mt	norwich	pennvylvania
kenosha	maines	multistate	nv	pensacola
kentuckey	management	murfreesboro	ny	peoria
kentuckians	manchester	murrieta	nyc	philadelphia
kentucky	marina	muskegon	o	philly
keystone	maryland	myrtle beach	oakland	phoenix
killeen	marysville	n	ocala	pittsburgh
kissimmee	massachsetts	naperville	oceanside	plano
knoxville	massachusets	naples	odessa	pllc
ks	massachusetts	nashua	of	pomona
ky	mcallen	nashville	ofnew	pompano beach
l	mchenry	national	ogden	port arthur
la	md	nc	oh	port orange
la	me	nd	ohio	port saint lucie
lacey	medford	ne	ohioans	port st. lucie
lafayette	melbourne	nebraska	ohios	portland
lake charles	memphis	nebraskans	ok	portsmouth
lakeland	merced	nevada	oklahoma	poughkeepsie
lakewood	mesa	nevadalas	oklahoma city	products
lancaster	mesa	nevadastate	olathe	professional
land	mesquite	new	olympia	providence
lansing	metro	new bedford	omaha	provo
laredo	mi	new hampshire	on	public

Table 18: 1557 Stop Words: Part 4

pueblo	seaside	thru	west
punta gorda	seattle	tn	west covina
q	sebastian	to	west valley city
r	service	toledo	west virginia
racine	services	topeka	western
raleigh	shreveport	torrance	westminster
rancho cucamonga	simi valley	trenton	wi
reading	sioux city	tucson	wichita
redding	sioux falls	tulsa	wilmington
reno	society	tuscaloosa	winston
rhode island	solutions	tx	winter haven
ri	south	tyler	wis
richland	south bend	u	wisc
richmond	south carolina	united	wisconsin
richmond county	south dakota	us	wisconsins
river	south lyon	usa	worcester
riverside	southeast	ut	wv
roanoke	southeastern	utah	wy
rochester	southwest	utahans	wyo
rockford	southwestern	utahns	wyoming
rocky	spartanburg	utica	x
roseville	spokane	v	y
round lake beach	springdale	va	yakima
s	springfield	vallejo	yonkers
sacramento	st. louis	valley	york
saginaw	st. paul	vancouver	youngstown
saint louis	st. petersburg	vermont	z
saint paul	stamford	vermonters	
saint petersburg	state	vero beach	
salem	sterling heights	victorville	
salinas	stockton	virginia	
salt lake city	sunnyvale	virginia beach	
san antonio	syracuse	virginians	
san bernardino	system	virginias	
san buenaventura	systems	virignia	
san diego	t	visalia	
san francisco	tacoma	vt	
san jose	tallahassee	w	
santa ana	tampa	wa	
santa barbara	temecula	waco	
santa clara	tempe	warren	
santa clarita	tenneessee	washing	
santa cruz	tennesee	washingt	
santa maria	tennesseans	washington	
santa rosa	tennessee	washington	
sarasota	tex	washington2	
savannah	texans	washingtonys	
sc	texas	washinton	
scottsdale	the	wasington	
scranton	thornton	waterbury	
sd	thousand oaks	waterloo	