Learning and Visualizing Political Issues from Voting Records

Erik Goldman, Evan Cox, Mikhail Kerzhner

Abstract

For our project, we analyze data from US Congress voting records, a dataset that consists of "yes," "no," or "not present" votes on various bills for each active congressperson. By scraping and parsing this data, we are able to model a congressperson as a list of bills for which they voted. After doing so, we have a data model perfectly suited to the "bag of words" model from information retrieval-- an unordered set of "terms" (or bills) instantiated in multiple "documents" (congresspeople). We use several information retrieval techniques on this dataset, focusing on the use of latent factor models. These models assume that a fixed number of latent "topics" are responsible for creating each bill, and that congresspeople themselves can be represented as a mixture of these topics corresponding to the topics they are likely to vote for. We then decompose our documents and bills into topic vectors, a process that yielded several interesting results. First, we are able to compute the document similarity between any two congresspeople, which we use to create a 3D visualization of congress in which the distance between congresspeople represents their voting record dissimilarity. Second, we have a topic representation of all the bills in congress, which we use to cluster and visualize legislation. To perform the actual dimensionality reduction, we use research LSI techniques such as the singular value decomposition, pLSI, and latent Dirichlet allocation. We used LDA to visualize political issues viewed as topics, and to gain a measure of bipartisanship in congress. Our dataset is readily available from http://www.govtrack.us/congress/votes.xpd.

Pre-Processing and Setup

For our project, we use the data available at http://clerk.house.gov/evs/2007/index.asp. Vote information for each bill is in the convenient XML format, and we were able to build a scraper that extracted the needed information for our data analysis. We created a numerical encoding for each congressperson (c_i) as well as a numerical encoding for each bill (b_i). With these encodings, we produce our training data set in two formats. The first format is a mapping from b_i to a list of (c_i, vote type), where vote type is simply Yes, No, or Not Present. The second format is the reverse mapping from c_i to a list of (b_i, vote type).

This data is then sent to a Python script, which prepares our extracted information for LDA (see next section) by turning it into a series of "documents," with each document as a congressperson and each word as a bill that they voted for.

Low-rank Approximation of Congresspeople and Issues

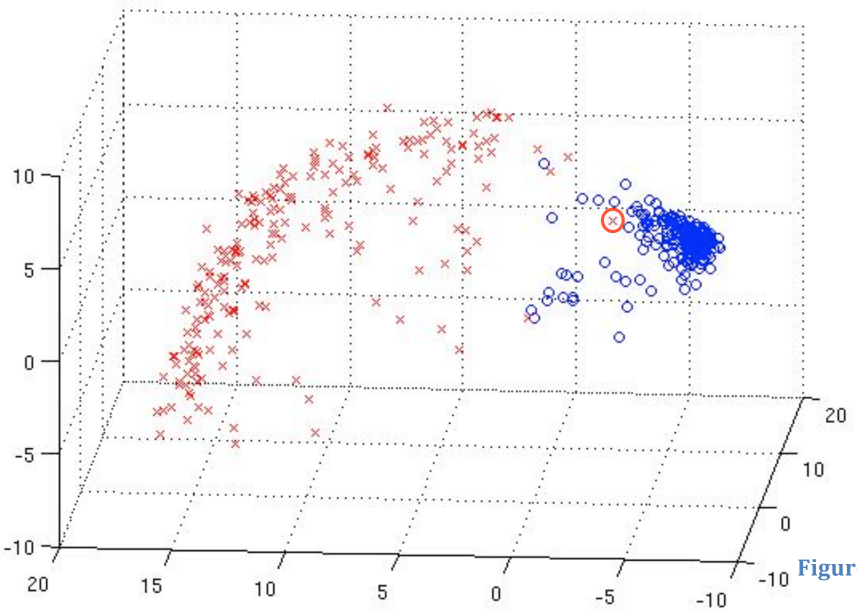


Figure 1

As mentioned before, one of the goals of the project is to create a simple visual representation of congresspeople. In order to do this, we need to reduce the matrix *M* to three dimensions, where each row of *M* represents a bill, each column represents a congressperson, and the matrix entry represents a yes, no, or not present vote on the bill. This is accomplished using LSI techniques of SVD decomposition and low-rank approximation. Using Latent Semantic Indexing, we can reduce *M* to a k-rank matrix *M'*, and *M'* is a product of three matrices U,S, and V. *M'* has the property that it is a matrix of rank k with the smallest Frobenius error. Consequently, *M'* captures the orthogonal (and thus "information dense") axes from our high-dimensional data. In our case, we produce *M'* of rank three, and, using S and V matrices from the SVD decomposition of *M'*, we can represent each congressperson as a point in three dimensional space, where each dimension corresponds to a linear combination of various bills (or topics). By

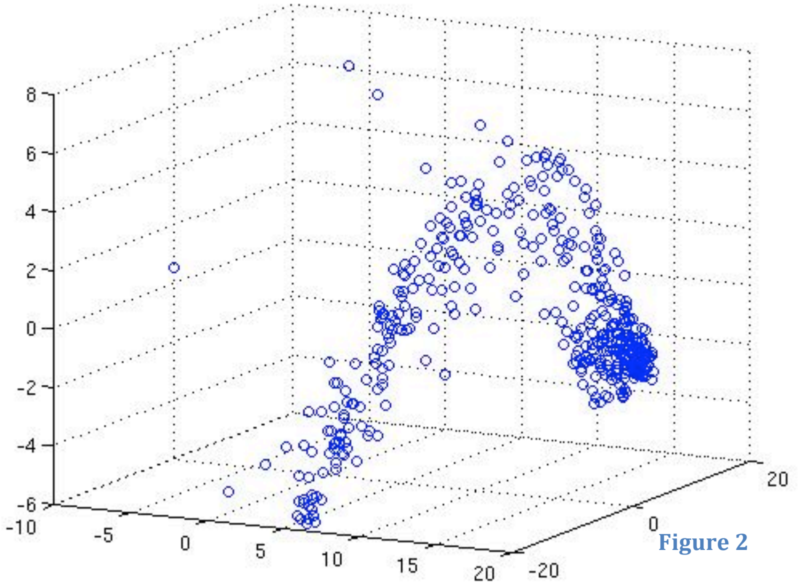


Figure 2

the Johnson-Lindenstrauss theorem, the distances between two congresspeople in this 3D space should be a reflection of their distance in the much higher dimension space of congresspeople x bills. Because the nature of LSI is such that dimension reduction combines "related" axes in vector space, in our graphs, congresspeople with similar opinions on various issues appear close to each other, and congresspeople with differing opinions will be far apart.

After applying the above methods to our data, we were able to generate the graph in Figure 1. In the graph, each blue data point represents one democratic congressperson, and each red point represents one republican congressperson. The Latent Semantic Analysis technique works as predicted. In the figure, it is clear that democratic congresspeople belong to one clear cluster, whereas republican congresspeople are in the other clear cluster. When analyzing the graph, we saw a few republican congressmen who were close or in the cluster of democrats. We investigated these curious data points and found that these congresspeople were well known to be moderate republicans with a liberal voting record. In particular, in figure 1, the data point with the red circle around it represents Wayne Gilchrest, a congressman from Maryland, who according to Wikipedia is commonly known to be a "republican-in-the-name-only" and, in fact, "was ranked as the House's most liberal Republican in 2008…by the *National Journal*" (Wikipedia).

In addition, our graph shows a much tighter cluster of Democrats than Republicans, perhaps indicating that Republicans in the House were more ideologically independent than Democrats, who tended to vote as a more cohesive block during this period.

Since the results of applying SVD and low rank approximation to the bill-congressman matrix M were so successful, our next strategy was to apply SVD to the transpose of M. The idea is to now reduce every bill to a point in three- dimensional space, and graph the resulting points. The results are plotted in Figure 2. There are no two or three obvious clusters in the figure, and the points form somewhat of a continuous surface. Consequently, this strategy is not successful in finding few, obvious bill clusters, and we used a modification of the above technique to achieve interesting results.

Table 1

Land Development 1. To provide for the continuation of agricultural and other programs of the Department of Agriculture through the fiscal year 2012. 2. Making supplemental appropriations for agricultural and other emergency assistance. 3. Water Resources Development Act.	National Security 1. Ensuring Military Readiness Through Stability and Predictability Deployment Policy Act. 2. Comprehensive American Energy Security and Consumer Protection Act. 3. To provide for the redeployment of United States Armed Forces and defense contractors from Iraq.
Child safety 1. Enhancing the Effective Prosecution of Child Pornography Act of 2007. 2. PROTECT Our Children Act of 2007. 3. KIDS Act of 2007.	Unemployment Relief 1. Emergency Extended Unemployment Compensation Act. 2. Making supplemental appropriations for job creation and preservation, infrastructure investment, and economic and energy assistance for the fiscal year ending September 30. 3. Emergency Extended Unemployment Compensation Act.

Discovery of Bill Issues

In order to find a small number of bill clusters, we perform a low rank approximation of the transpose of *M*. In this case, our approximation is of rank twenty-five, which we determined experimentally. Consequently, every bill is now a point in twenty-five dimensional space. Once we have this representation, we are able to collect the "top" bills for every dimension. A bill is considered a top bill for a dimension if the value of the coordinate of the bill for this dimension is large, which indicates that this dimension had a significant contribution to the reconstruction of the bill in our higher order space, and informally that the latent topic plays a large role in the perception of this bill. Upon analysis of the top bills, in almost every dimension, we saw that top bills corresponded to one particular issue. As an example, the top three bills for four out of twenty-five dimensions are shown in Table 1. Each group is labeled with the real-world political issue that the dimension represents. Consequently, this bill clustering technique is successful in finding bills concerned with the same topic using unsupervised learning.

A latent topic approach to political issue discovery

Another one of our goals for this project was to develop a generative topic model for congress, and analyze the implications of such a model, where the topics represent political issues that drive politicians votes. A topic in this model is simply a distribution over votes, "yes", "no", "not present" on bills.
We designed the following generative topic model for the voting record of a congressperson.

1. Choose N ~ [number of bills, which turns out to be irrelevant]
2. Choose θ~ Dirichlet(α) corresponding to the "congressperson topic multinomial," the mixing proportions of topics for c
3. For each vote N
    a. Choose the latent topic z_n ~ Mult(θ)
    b. Vote on the bill b with vote v, (absent, yes, or no) $b_v \sim p(b_v \mid z_n, \beta)$, which is a multinomial conditioned on the topic z_n.

Essentially, if a congressperson is pro-choice, then they have a high probability of voting yes on bills that are also considered pro-choice. Once we condition on the topic of pro-choice, the congressperson and the bill become conditionally independent. Thus, we draw a topic out of a congressperson's topic multinomial, draw a bill out of that topic's distribution, and the congressperson votes the way given by the topic's distribution over bill votes. Accordingly a pro-choice congressperson c's multinomial θ^c will have high values for topics Z whose distribution place more weight on yea votes on pro-choice bills.

This topic model is an exact analogy to latent Dirichlet allocation (LDA), which supposes the same distribution over documents and words[1]. We are thus able to model a congressperson as a document and their votes on bills as the words that comprise the document representing them. By processing this data with LDA, we can recover the probabilities of each bill given a topic as well as the congressperson topic multinomial for each voting member of the U.S. House of Representatives. LDA is a widely studied generative model of corpora and we were able to take advantage of existing algorithms for finding a distribution for each of the k topics that maximizes the likelihood of this generative model. We used the Gibbs sampling algorithm presented in [1] that finds the distribution for each topic that maximizes the likelihood of this generative model. The Markov chain state update rule for the assignment of w_i to a topic z_i is given by the formula.

$$P(z_i = j \mid z_{-i,} w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\bullet)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{j}^{(d_i)} + T\alpha}$$

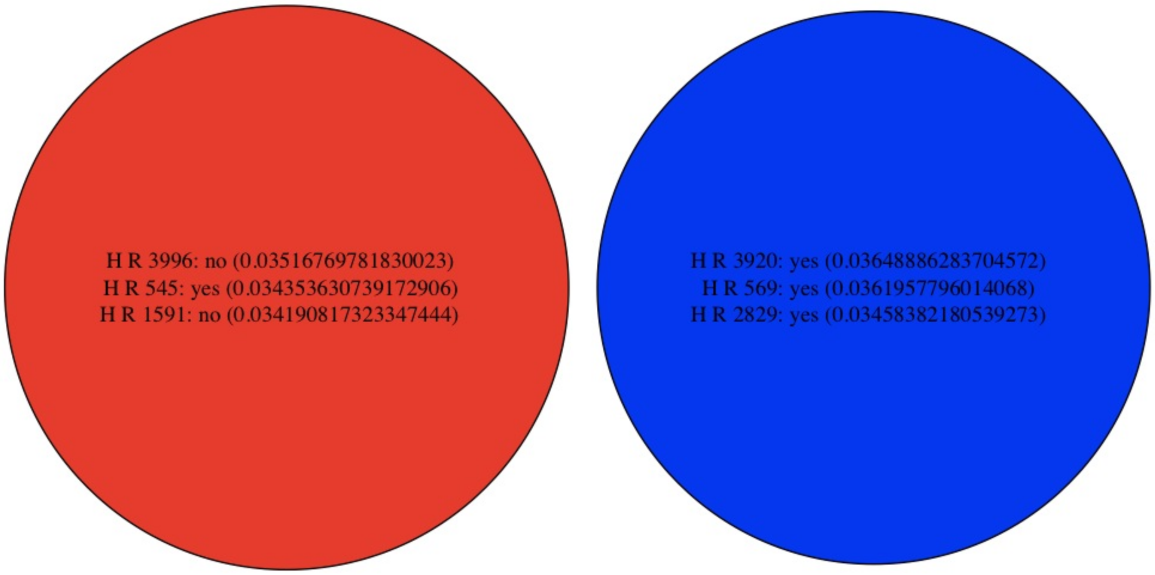
Where z_{-i} is the assignment of all z_k such that k ≠ i, $n_{-i,j}^{(wi)}$ is the number of times words the same as w_i have been assigned to topic j, $n_{-i,j}^{(\cdot)}$ is the total number of words assigned to topic j, $n_{-i,j}^{(di)}$ is the number of words in the document containing w_i that have been assigned to topic j, and $n_j^{(di)}$ is the total number of words in document d, T is the total number of topics, W is the size of the vocabulary, all not counting the assignment of the w_i, and α and β are free parameters that control how heavily the distribution is smoothed [2]. We determined α and β experimentally, but their choice did not have a visible effect on our results.

For convergence we tested the cosine distance of the vector for each topic where the i^{th} entry is $p(w_i|z)$, where w_i is a distinct word in the vocabulary, not the word in the i^{th} index, with the analogous vector previous iteration. We let the sampler burn in for 25 iterations and then repeatedly took the cosine similarity between two subsequent iterations. Once the cosine similarity of each vector with the previous iteration's was above a threshold *t* for 10 iterations, an approximate maximizing stationary distribution had been reached for each of the topics. Once we had obtained the stationary distribution for each topic we then calculated the probability of each congressperson d_i given a topic z_j, given by

$$\prod_{w_k \in d_i} p(z_k = j)$$

since a word and a document are conditionally independent given a topic. This gives a distribution over topics, the "congressperson multinomial" for each congressperson θ^i , where a $p(d_i|z_j)$ represents the portion of the mixture for d_i that is composed of topic z_j Given the multinomial distribution over topics for every congressperson, we then split them into democratic and republican congressmen, and averaged their distributions over topics, to find the multinomial distribution over topics for the "average" democratic congressman, and the multinomial distribution over topics for the "average" republican congressman. Accordingly θ^D, average democratic distribution is θ^R.

Given these average distributions over topics, we then visualized the difference of the distributions in the following way. We took the p(R|z), p(D|z) for each topic z, and mapped that to a value between 0 and 255. We then represented each topic z as a circle, colored with red proportional to p(R|z) and blue proportional to p(D|z). The color for a topic was given by (Red = 255 * p(R|z), Green = 0, Blue = 255 * p(D|z). A topic that is more republicans will be redder, and a topic that is democratic will be bluer. Given this visualization, the presence of only red or blue topics with varying brightness suggests that republican and democrats tend to generate their votes from mutually exclusive sets of topics, and accordingly tend to vote differently. The presence of purple topics suggests that there are issues that democrats and republicans tend to vote similarly on. So the presence of only red means that the congress at this time was more bipartisan, with congress-people having less mutual topics they agreed on, and accordingly drew their votes from. Here is the visualization for a topic model k = 2, with the cosine similarity threshold for each topic *t* = .9.


H R 3996: no (0.03516769781830023)
H R 545: yes (0.034353630739172906)
H R 1591: no (0.034190817323347444)
H R 3920: yes (0.03648886283704572)
H R 569: yes (0.0361957796014068)
H R 2829: yes (0.03458382180539273)

If z_1 is the left topic, and z_2 is the right topic, then $P(R|z_1) = 0.962$, $P(R|z_2) = 0.00963$, $P(D|z_1) = 0.947$, $P(D|z_2) = 0.0243$. So we can conclude that Republicans' votes were mostly generated by drawing votes from the z_1 whereas Democrats' votes were generated mostly by drawing votes from z_2. Accordingly we can conclude that congress at this time was more bipartisan according to this model, because Democrats and Republicans were mostly comprised of votes drawn from mutually exclusive sets of topics.

The 3 most likely votes for each topic are presented by their bill number, and likelihood given that topic. The top 3 for the more Republican topic were,

1. H R 3996: Tax Increase Prevention Act of 2007, no (0.035)
2. H R 545: Native American Methamphetamine Enforcement and Treatment Act of 2007, yes (0.034)
3. H R 1591: U.S. Troop Readiness, Veterans' Care, Katrina Recovery, and Iraq Accountability Appropriations Act, 2007, no (0.034)

The top 3 for the more Democratic topic were

1. H R 3920: Trade and Globalization Assistance Act of 2007, yes (0.036)
2. H R 569: Water Quality Investment Act of 2007, yes (0.036)
3. H R 2829: Financial Services and General Government Appropriations Act, 2008, yes (0.034)

We performed this experiment with k = 2, 3, 4, 5, 6, 7, 8, but the results are not reproduced here for lack of space.
The strong disparity in the mixing proportions for the average democrat and average republican suggests that congress was strongly divided, the democrats and republicans votes being generated by nearly mutually exclusive distributions over votes, or political issues.

Conclusion

Our goal for the project was to use unsupervised learning techniques to find clusters of politicians with the same opinions as well as clusters of bills on the similar issues. In the end, we were able to successfully apply techniques of latent semantic indexing to achieve both of these goals. Our most interesting results included discovery of republican congressmen with liberal voting records as well as discovery of a small set of topics that are currently most important to the congress.

[1] Blei, David. M, Ng, Andrew Y, and Jordan, Michael I. "Latent Dirichlet Allocation." *In Advances in Neural Information Processing Systems 14*. (2002)
[2] Griffiths, Thomas L., and Steyvers, Mark. "A probabilistic approach to semantic representation." *In Proceedings of the Twenty-Fourth Annual Conference of Cognitive Science Society* (2002)
[3] S. Deerwester, Susan Dumais, G. W. Furnas, T. K. Landauer, R. Harshman (1990). "Indexing by Latent Semantic Analysis". *Journal of the American Society for Information Science* **41** (6): 391–407.