Revisiting Fightin' Words: Feature Selection Using an Informed Dirichlet Model

MATTHEW DENNY THURSDAY 14TH JULY, 2016

Selecting textual features that distinguish between documents written by different authors or groups of authors is an important task in the analysis of social and political texts. The natural language processing literature is rich with methods for feature selection (Manning et al., 2008), but not all of these are specifically tailored to social science applications. Monroe et al. (2008) introduce a set of feature selection methods that are tailored to the political science domain, but gloss over a number of important details, and do not provide a working implementation. This document provides an annotated description of the informed Dirichlet model for lexical feature selection presented in Monroe et al. (2008, sections 3.3.1–3.5.1), and accompanies a working implementation of several feature selection methods[1].

The goal of the informed Dirichlet model is to identify meaningful words that distinguish between documents written/spoken by two or more groups. The authors go through a lot of different approaches to try to get at the meaningful words that (in their example) distinguish between Democrat and Republican views on abortion. They settle on a Dirichlet model for selecting these top words, and consider two priors: and informed Dirichlet, and a Laplace prior. The Laplace prior model is quite complex and difficult to fit, and does not provide noticeably improved performance over the informed Dirichlet model according to the authors, so I focus on the informed Dirichlet model in this document. Below I describe the generative process for corpus term counts from Monroe et al. (2008, section 3.3), discuss feature evaluation under this model, and illustrate with some output.

1 Generating Term Counts

Let a corpus have a vocabulary of W unique terms, define $\mathbf{y} = \{y_w\}_{w=1}^W$ as the vector of term counts in the corpus, and let $n = \sum_{w=1}^W y_w$ be the total number of tokens in the corpus. The authors model $\mathbf{y}$ as a draw from a multinomial distribution with with multinomial probability vector $\boldsymbol{\pi}$:

$$\mathbf{y} \sim \text{Multinomial}(n, \boldsymbol{\pi}) \tag{1}$$

This corpus may contain documents about a number of different "topics". For example, in the U.S. congressional bills corpus, there are bills about health care, energy policy, civil rights, defense appropriations, etc. Let these topics $\mathbf{t} = \{t_k\}_{k=1}^K$ be indexed by k, then we can analogously define the counts of words associated with topic k as $\mathbf{y}_k = \{y_{kw}\}_{w=1}^W$ and the total number of tokens associated with a topic $n_k = \sum_{w=1}^W y_{kw}$. The authors are not clear on this point, but it seems that topics may either be unique labels for each document (such as the categorical labels given by the Congressional Bills project (see Purpura and Hillard, 2006)), or a set of terms associated with a topic inferred using LDA. What the authors do not make clear is whether their approach depends on the assumption that each token in a document is uniquely associated with a topic ($n = \sum_k \sum_w \mathbf{y}_{kw}$). For all of the analyses in this document, we will be following the assumption that documents are uniquely assigned to topics.

Monroe et al. also model the distribution of terms in a topic as a draw from a multinomial distribution with with multinomial probability vector $\boldsymbol{\pi}_k$:

$$\mathbf{y}_k \sim \text{Multinomial}(n_k, \boldsymbol{\pi}_k) \tag{2}$$

Finally, each document may be written by a member of one of I groups. In our setting we can think of these groups as political parties (Democrat and Republican), but they could be men and women,

[1]Available as part of the SpeedReader R package (beta): https://github.com/matthewjdenny/SpeedReader.

or legislators from different states, for example. Thus we can also define term counts for documents written by members of a particular group similarly to the way that we define term counts for topics. Let $\mathbf{y}^{(i)} = \{y_w^{(i)}\}_{w=1}^W$ be the vector of term counts in documents written by members of group i, and the total number of tokens in documents written by members of group i is thus $n^{(i)} = \sum_{w=1}^W y_w^{(i)}$. The authors model the distribution of terms in a documents written by members of group i as a draw from a multinomial distribution with multinomial probability vector $\boldsymbol{\pi}^{(i)}$:

$$\mathbf{y}^{(i)} \sim \text{Multinomial}(n^{(i)}, \boldsymbol{\pi}^{(i)}) \tag{3}$$

Combining all of the ideas described above, the author's main goal is to model the counts of terms written about topic k by members of group i (the counts of terms in speeches about reproductive health given by Democrats). Thus the authors model the distribution of terms in a documents written by members of group i about topic k as a draw from a multinomial distribution with multinomial probability vector $\pi_k^{(i)}$:

$$\mathbf{y}_k^{(i)} \sim \text{Multinomial}(n_k^{(i)}, \boldsymbol{\pi}_k^{(i)}) \tag{4}$$

1.1 Placing a Prior on $\mathbf{y}_k^{(i)}$

In their preferred approach described in Monroe et al. (2008, section 3.5.1), the authors place an informative prior on the distributions over terms in documents written by members of group i about topic k. They select a Dirichlet prior on $\pi_k^{(i)}$ as it is conjugate to the Multinomial distribution. Thus,

$$\boldsymbol{\pi}_k^{(i)} \sim \text{Dirichlet}(\alpha, \mathbf{m}) \tag{5}$$

the authors want to induce shrinkage in their estimates of the degree to which particular terms are associated with documents written by group i about topic k, relative to documents written by other groups. In order to do this, they select a particular form for $\alpha, \mathbf{m}$. The authors select α equal to the average number of tokens in a document, across the entire corpus, and set $\mathbf{m}$ proportional to the frequency of a term in all documents in the corpus.

$$m_w = \frac{y_w}{n} \tag{6}$$

I will discuss the implications of selecting a prior of this form in section 2.2.

2 Evaluating Features

Due to Dirichlet-multinomial conjugacy and the lack of any other covariates in the model, it is possible possible to form a posterior point estimate of $\boldsymbol{\pi}_k^{(i)}$ analytically. This point estimate takes the following form:

$$\widehat{\pi}_{kw}^{(i)} = \frac{y_{kw}^{(i)} + \alpha m_w}{n_k^{(i)} + \alpha} \tag{7}$$

(where we note that $\alpha = \sum_{w=1}^W \alpha m_w$). In words, we have a point estimate of the posterior probability of observing a particular term in a document about topic k, written by a member of group i. But what we really want to know is the odds of observing that particular term in a document about topic k, written by a member of group i, *relative* to observing it in a document about topic k, written by a member of any other group. We start by denoting the odds of observing term w in topic k as:

$$\Omega_{kw} = \frac{\pi_{kw}}{1 - \pi_{kw}} \tag{8}$$

Now we can form the log-odds ratio of observing a particular term w in a document about topic k, written by a member of group i, relative to observing it in a document about topic k written by a member of any other group as $\delta_{kw}^{(i)} = \log(\Omega_{kw}^{(i)}/\Omega_{kw})$. Expanding and substituting in our point estimates for π_{kw}, $\pi_{kw}^{(i)}$:

$$\begin{aligned}
\widehat{\delta}_{kw}^{(i)} &= \log(\Omega_{kw}^{(i)}/\Omega_{kw}) && (9)\\
&= \log\left(\frac{\frac{\pi_{kw}^{(i)}}{1-\pi_{kw}^{(i)}}}{\frac{\pi_{kw}}{1-\pi_{kw}}}\right) && (10)\\
&= \log\left(\frac{\frac{y_{kw}^{(i)}+\alpha m_w}{n_k^{(i)}+\alpha}}{1-\frac{y_{kw}^{(i)}+\alpha m_w}{n_k^{(i)}+\alpha}}\right) - \log\left(\frac{\frac{y_{kw}+\alpha m_w}{n_k+\alpha}}{1-\frac{y_{kw}+\alpha m_w}{n_k+\alpha}}\right) && (11)\\
&= \log\left(\frac{\frac{y_{kw}^{(i)}+\alpha m_w}{n_k^{(i)}+\alpha}}{\frac{n_k^{(i)}+\alpha}{n_k^{(i)}+\alpha}-\frac{y_{kw}^{(i)}+\alpha m_w}{n_k^{(i)}+\alpha}}\right) - \log\left(\frac{\frac{y_{kw}+\alpha m_w}{n_k+\alpha}}{\frac{n_k+\alpha}{n_k+\alpha}-\frac{y_{kw}+\alpha m_w}{n_k+\alpha}}\right) && (12)\\
&= \log\left(\frac{\frac{y_{kw}^{(i)}+\alpha m_w}{n_k^{(i)}+\alpha}}{\frac{n_k^{(i)}+\alpha-y_{kw}^{(i)}+\alpha m_w}{n_k^{(i)}+\alpha}}\right) - \log\left(\frac{\frac{y_{kw}+\alpha m_w}{n_k+\alpha}}{\frac{n_k+\alpha-y_{kw}+\alpha m_w}{n_k+\alpha}}\right) && (13)\\
&= \log\left(\frac{\left(y_{kw}^{(i)}+\alpha m_w\right)\left(n_k^{(i)}+\alpha\right)}{\left(n_k^{(i)}+\alpha\right)\left(n_k^{(i)}+\alpha-y_{kw}^{(i)}+\alpha m_w\right)}\right)\\
&\quad - \log\left(\frac{(y_{kw}+\alpha m_w)(n_k+\alpha)}{(n_k+\alpha)(n_k+\alpha-y_{kw}+\alpha m_w)}\right) && (14)\\
&= \log\left(\frac{y_{kw}^{(i)}+\alpha m_w}{n_k^{(i)}+\alpha-y_{kw}^{(i)}+\alpha m_w}\right) - \log\left(\frac{y_{kw}+\alpha m_w}{n_k+\alpha-y_{kw}+\alpha m_w}\right) && (15)
\end{aligned}$$

which is equivalent to the result in equation (15) in Monroe et al. (2008). From here we can finally capture the usage difference of term w in documents about topic k between two groups i and j as a log odds ratio:

$$\begin{aligned}
\widehat{\delta}_{kw}^{(i-j)} &= \left[\log\left(\frac{y_{kw}^{(i)}+\alpha m_w}{n_k^{(i)}+\alpha-y_{kw}^{(i)}+\alpha m_w}\right) - \log\left(\frac{y_{kw}+\alpha m_w}{n_k+\alpha-y_{kw}+\alpha m_w}\right)\right]\\
&\quad - \left[\log\left(\frac{y_{kw}^{(j)}+\alpha m_w}{n_k^{(j)}+\alpha-y_{kw}^{(j)}+\alpha m_w}\right) - \log\left(\frac{y_{kw}+\alpha m_w}{n_k+\alpha-y_{kw}+\alpha m_w}\right)\right] && (16)\\
&= \log\left(\frac{y_{kw}^{(i)}+\alpha m_w}{n_k^{(i)}+\alpha-y_{kw}^{(i)}+\alpha m_w}\right) - \log\left(\frac{y_{kw}^{(j)}+\alpha m_w}{n_{kw}^{(j)}+\alpha-y_{kw}^{(j)}+\alpha m_w}\right) && (17)
\end{aligned}$$

We have arrived at a log odds ratio expressing the differential odds we see a particular term w used by members of groups i and j in documents about topic k. So a large positive value would indicate that members of group i tend to use the word much more frequently, and a large negative value would indicate that members of group j use the word much more frequently. As Monroe et al. (2008) note, this point estimate doesn't necessarily get us anywhere if we are looking for meaningful words that will distinguish between the two group's views on topic k. That is because (similar to pointwise mutual information), these point estimates will be dominated by obscure (infrequent) words. This is where using a model based approach is helpful, because our point estimates for these infrequent words will also have high variance. The authors point out that because we are using a "model", it is possible to calculate standard errors for the point estimates of the log odds-ratios.

2.1 Calculating Standard Errors

If we calculate standard errors for our point estimates, we can then calculate z-scores for each term, and rank terms by these z-scores. Intuitively, using z-scores for ranking terms should provide better performance, because they will balance the desire for a large (proportional) difference in term use between groups with a penalty for infrequent (high variance) terms. We can calculate standard errors for log odds-ratios using the "logit approximation", as described in Morris and Gardner (1988). For a log odds ratio of the form $\log(a/b) - log(c/d)$, this approximation is:

$$\widehat{\sigma}^2 = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \tag{18}$$

This is a large sample (normal) approximation that should be reasonably accurate as long as $a, b, c, d \gg 0$. It is possible to calculate the exact variance (Breslow and Day, 1980), but this involves working with a non-central hypergeometric distribution which is generally very complex, so the standard approach in the literature is to use the normal approximation when working with contingency tables. However, it is possible to have one of a, b, c, d close to zero when we are using text (something not discussed by Monroe et al. (2008)). In this case, we should expect the approximation of the variance to be inflated leading to a smaller z-score and a lower ranking. Looking at the counts for both groups is therefore quite important while using this approximation. However, the use of a strong informative prior should help address this issue somewhat. Plugging in the terms in equation 17 into equation 18 yields:

$$\text{Var}\left(\widehat{\delta}_{kw}^{(i-j)}\right) = \frac{1}{y_{kw}^{(i)} + \alpha m_w} + \frac{1}{n_k^{(i)} + \alpha - y_{kw}^{(i)} + \alpha m_w} + \frac{1}{y_{kw}^{(j)} + \alpha m_w} + \frac{1}{n_{kw}^{(j)} + \alpha - y_{kw}^{(j)} + \alpha m_w} \tag{19}$$

With this variance approximation in hand, we can finally calculate z-scores for $\widehat{\delta}_{kw}^{(i-j)}$, which Monroe et al. (2008) denote $\widehat{\zeta}_{kw}^{(i-j)}$ using the following standard formula:

$$\widehat{\zeta}_{kw}^{(i-j)} = \frac{\widehat{\delta}_{kw}^{(i-j)}}{\sqrt{\text{Var}\left(\widehat{\delta}_{kw}^{(i-j)}\right)}} \tag{20}$$

2.2 The Impact of an Informative Prior

Having derived the formulas for the point estimate and variance of $\widehat{\delta}_{kw}^{(i-j)}$, we can get a better sense of why an informative prior might be helpful in feature selection. We can see that as α increases, it will tend to shrink the point estimates of $\widehat{\zeta}_{kw}^{(i-j)}$ for terms that occur very frequently in the corpus (like function words) towards zero. This can improve interpretability of the top words. Coupled with tendency for infrequent words to be penalized via larger variance leaves us with top words which are somewhere in the middle in terms of frequency. The real question for a given application is what α we should select, and how it will affect our top words. A particular corpus may differ significantly from the floor speeches corpus used in Monroe et al. (2008) in that the documents may be much longer, and may comprise a much larger vocabulary[2]

Critically, the vocabulary sizes (depending on the term-vector extraction method) for a corpus like the congressional bills corpus are orders magnitude larger than the vocabulary size in the examples used by Monroe et al. (2008) which was approximately 3,000 terms. The number of unique terms in the

[2]The Monroe et al. corpus has an average document length of approximate 500 words and a vocabulary size of only a few thousand.

vocabulary for the congressional bills corpus range from approximately eighty-thousand to over twenty-million depending on the term vector extraction method. If we examine the form of our point estimate $\widehat{\pi}_{kw}^{(i)}$, it becomes clear that three factors are important for the degree of smoothing in the model:

$$\widehat{\pi}_{kw}^{(i)} = \frac{y_{kw}^{(i)} + \alpha m_w}{n_k^{(i)} + \alpha} \tag{21}$$

The first of these is obviously α – as α increases, we get more smoothing. The second important factor is the size of the vocabulary. For a fixed $n_k^{(i)}$, increasing the vocabulary size by a factor of ten will effectively decrease the smoothing by a factor of ten as well. Finally, as the number of terms in the average document increases, so does the degree of smoothing. This can particularly make results for written and spoken text largely incomparable due to potentially large differences document length.

3 TF-IDF for Feature Selection

One obvious alternative to this complicated model is to use TF-IDF scoring. However, it seems like there are a lot fo different interpretations of what this means, and some of them are not appropriate for feature selection. The canonical formulation of TF-IDF taken from Manning et al. (2008) is :

$$\text{tf-idf}_{w,d} = \text{tf}_{w,d} \times \text{idf}_w \tag{22}$$

where documents are indexed by d and terms are indexed by w (note that Manning et al. (2008) index terms by t, but I am using w for consistency with the rest of our paper). Manning et al. (2008) suggest that the simplest version of term frequency is simply the count of term t in document d:

$$\text{tf}_{w,d} = \text{ The number of times term t appears in document d} \tag{23}$$

The authors also define the inverse document frequency to be:

$$\text{idf}_w = \log\left[\frac{N}{1 + df_w}\right] \tag{24}$$

where N is the total number of documents in the corpus, and the document frequency df_w is the number of documents where term w appears at least once. We add one to the denominator to prevent dividing by zero when $df_w = 1$, and as Manning et al. (2008) note this does not affect rankings since the 1 is just a constant multiplicative factor. The definition provided in Monroe et al. (2008) is essentially a straw man as it formulates the inverse document frequency term at the category level (so it is either 1 or 2). It provides terrible performance by construction because of the choice of formulation, and does not conform to any previously published definition of TF-IDF, so we will ignore it for the rest of this document. In the example application to the congressional bills corpus, lets ask ourselves what a reasonable formulation of TF-IDF might be, given that we want to identify words that are most highly associated with documents about topic k written by legislators that belong to group i?

It seems logical to keep the canonical formulation of idf_w given by Manning et al. (2008) in calculating our TF-IDF scores, because this best preserves the information we care about from the idf_w term. That would mean that while we may only be looking at documents about energy policy, we are going to use all of the information available to us (all documents in the corpus) in constructing the idf_w term. As for the $\text{tf}_{w,d}$ term, one option would be to aggregate these counts over all documents associated with topic k, written by group i. This would effectively combine these documents as one big document from which we could get $\text{tf}_{w,d,k}^{(i)}$. However, this would break the correspondence between the tf and idf terms in terms of their relative magnitude. The simple way to address this is to simply take the average of $\text{tf}_{w,d}$ over all documents associated with topic k and group i. Let $N_k^{(i)}$ be the total number of documents associated with topic k and group i, then the average term frequency in documents associated with topic k and group i is:

$$\text{average tf}^{(i)}_{w,k} = \frac{1}{N^{(i)}_k} \sum_{\text{topic}(d)=k} [\text{tf}_{w,d}] \tag{25}$$

Thus, I propose we define our TF-IDF measure as:

$$\text{tf-idf}^{(i)}_{w,k} = \text{average tf}^{(i)}_{w,k} \times \text{idf}_w \tag{26}$$

In words, we simply average the term frequency over all documents associated with topic k and group i and then multiply this by the normal inverse document frequency term to get our TF-IDF scores. One way to potentially improve on this measure is to make use of log term frequency counts, as suggested in Manning and Schütze (1999, p. 544). If we adopt this formulation then our TF-IDF scores become:

$$\text{tf-idf}^{(i)}_{w,k} = \left[1 + \log\left(\text{average tf}^{(i)}_{w,k}\right)\right] \times \text{idf}_w \tag{27}$$

The reason this might yield an improvement is that it will place a greater relative weight on the IDF term, which in the case of congressional texts seems to be important. This will tend to select for terms which appear in fewer documents than the version that uses natural term counts. In addition to the formulations discussed above, I have tested out a couple of others from the Wikipedia page for TF-IDF [link], and the "augmented" TF formulation from Manning and Schütze (1999, p. 544), but these alternative formulations tended to yield (qualitatively) worse performance in my testing in that the top terms are less interpretable. I test the formulations in equations 26 and 27 in the empirical evaluation in the next section and the results indicate that the log(term frequency) presented in equation 27 generally provides better qualitative performance.

4 Empirical Evaluation

The feature selection methods described above are implemented in the `feature_selection()` function in the SpeedReader[3] R package (beta). These methods are applied to the corpus of all bills introduced in the United States Congress between 1993 and 2014. In this application, I work with final versions of bills from the congressional bills corpus (as opposed to the original versions before the amendment process), of which there are 99,776. For the purpose of illustration, I begin by working with unigrams. I focus on bills that are coded as being mainly about "Healthcare", using the major topic labels generated by Purpura and Hillard (2006). For this analysis, I focus on all bills introduced in the House and Senate during the 113th session of Congress (2013–2014). This results in a total of 1,097 bills that were coded as mostly about health policy, of which 551 were sponsored by Democrats and 516 were sponsored by Republicans[4]. I selected health policy during the 2013–2014 session of Congress as a topical area to focus on because there are numerous media accounts of repeated efforts on the part of Republicans in Congress to weaken or repeal the Affordable Care Act during this time period. This makes health policy a place where we should see marked differences in language use during this period. Additional, Purpura and Hillard (2006) attained relatively high classification accuracy (88%) in their validation experiments with this topic compared to many others.

Tables 2 and 4 present the top unigrams associated with Democrat and Republican sponsored bills respectively, using four different feature selection methods. The first of these is pointwise mutual information (PMI) ranking using a cutoff of words that appeared at least 50 times in bills sponsored by both Democrats and Republicans. I selected this relatively high threshold because it seemed to provide the best performance in testing by avoiding terms that appear almost exclusively in documents written by

[3] https://github.com/matthewjdenny/SpeedReader

[4] 30 bills in this category were sponsored by Independents, but are omitted from this analysis because the informed Dirichlet model was designed to be applied to the comparison of two categories.

one party. The second measure is the formulation of TF-IDF from equation 26, while the third measure is the formulation of TF-IDF with logged term frequency from equation 27. The final column displays top words as ranked by the informed Dirichlet model described above. Following Monroe et al. (2008), I set $\alpha = 2,547$, the mean number of unigrams per document across the entire corpus, and $\boldsymbol{m}$ proportional to the relative frequency of a term in the entire corpus.

Starting by examining the unigram results, the four methods for feature selection seem to offer qualitatively similar performance, with PMI perhaps providing somewhat less interpretable results. Thus we cannot make any clear statements about which method should be preferred based on this qualitative analysis alone. The top terms associated with Democrat-sponsored bills seem to deal more with diseases and treatments, while the top terms associated with Republican-sponsored bills seem to deal more heavily with insurance. This general finding was corroborated by a manual examination of a sample of bills that contain a high count of the top terms in each category, and fits with the popular narrative that the Republicans in Congress spent more energy on insurance (Affordable Care Act) related issues than Democrats. However, these results are far from conclusive. For example, to rigorously verify that the insurance related top terms associated with Republican sponsored bills are in fact dealing with the Affordable Care Act, a much more exhaustive manual investigation would be necessary.

This ambiguity comes from examining unigrams out of their context in longer n-grams. One potential solution to this problem is to instead consider syntactically coherent phrases as the units of analysis instead of unigrams. In particular, I apply these methods to a set of phrase extractions detailed in Denny et al. (2015). Tables 6 and 7 present the top phrases associated with Democrat and Republican sponsored bills respectively, and were compiled in a similar manner to the unigram tables ($\alpha = 2,470$). As we can see, the increased context provided considering longer phrases as the units of analysis tends to disambiguate the meaning of a particular unigram, and improve the overal interpretability of the top terms associated with each party.

Turning to a qualitative comparison of the different methods for feature selection, the informed Dirichlet model arguably selects features which are more interpretable than those selected by any of the other methods. In particular, the other methods tend to include a number of "boilerplate" phrases such as references to the U.S. code or parts of a bill, which are not informative about policy differences in the legislation sponsored by members of different parties. However, one general issues across all methods is that a number of different phrases with the same meaning a captured in the top terms. We can clearly see that some of these phrases subsume each other, such as "patient protection and affordable care act", "protection and affordable care act", "patient protection and affordable care", "protection and affordable care", etc. To deal with these phrases that should be subsumed, I present a correlation-based algorithm for term subsumption from a ranked list of terms.

5 Correlation-Based Term Subsumption

The output from the feature selection methods described above is a ranked list of terms with the largest association scores with the particular category (in this case Democrat or Republican sponsored bills) of interest. As mentioned in the previous section, when these methods are applied to longer n-grams as the unit of analysis, we find that a number of terms which share a sub-string relationship are represented in the top terms. In order to aid in interpretability, we would like to automatically subsume these terms and present a list of top terms that represent a unique meaning in that list. To do so, I propose a correlation-based term subsumption algorithm for automatically clustering terms which share a sub-string relationship based on high correlations among their document-frequencies. Pseudocode for this algorithm is presented in Algorithm 1.

In words this algorithm proceeds as follows: We begin with a ranked list of terms and an associated document-term matrix as input. We then loop over this ranked list of terms, generating a specified number of term clusters one by one and removing the terms that are included in the current cluster from the input list after each iteration. At the beginning of each iteration, we select a **focal term** which is the highest

Algorithm 1: Correlation-Based Phrase Subsumption

```
Data: ranked_term_list,
      document_term_matrix,
      term_clusters_to_output,
      top_terms_to_search,
      correlation_threshold

# create a blank list to fill with term clusters, of length: term_clusters_to_output.
ranked_term_clusters = List(term_clusters_to_output)
for i ∈ 1:term_clusters_to_output do
    # 0. get the first term in ranked_term_list which will be our focal term for this iteration.
    focal_term = ranked_term_list[1]  # 1. Find terms of which the focal_term is a sub-string.
    current_term_cluster = List()  #  List to hold candidate terms.
    # only search the remaining top_terms_to_search of the ranked_term_list (cuts down on computational costs).
    for j ∈ 1:top_terms_to_search do
        if grep(focal_term, ranked_term_list[j]) then
            append(current_term_cluster, ranked_term_list[j])
        end
    end
    # Loop over all terms of which the focal term is a sub-string (currently stored in current_term_cluster) and find
    all sub-strings of those terms, and add them to current_term_cluster.
    for k ∈ 1:length(current_term_cluster) do
        for j ∈ 1:top_terms_to_search do
            if grep(ranked_term_list[j], current_term_cluster[k]) then
                append(current_term_cluster, ranked_term_list[j])
            end
        end
    end
    # get the unique terms in current_term_cluster, which is now the list of candidate terms to be subsumed.
    # 2. calculate correlations between the focal term and all other terms in current_term_cluster.
    correlation(focal_term,current_term_cluster)
    # 3. remove all terms from current_term_cluster whose correlation with the focal term is less than
    correlation_threshold.
    # 4. remove all terms remaining in current_term_cluster from ranked_term_list.
    # 5. We can now select the longest term (largest number of characters) in current_term_cluster to represent that
    cluster, and combine it with the metadata (z-score, variance, counts in both cateogries, etc.) associated with the
    focal term. There are now two pieces of information about the current term cluster: a list of terms that are included
    in it, and a "representative term" paired with the term-level metadata associated with the focal term. Both of these
    peice of information can now be stored in ranked_term_clusters.
    ranked_term_clusters[i] = current_term_cluster
end
return (ranked_term_clusters)
```

ranked term remaining in the input list. We then find all terms in the top (200-500) remaining terms in the ranked list of which the **focal term** is a sub-string. Because these terms are longer (more characters), they may convey more meaning, and may link the **focal term** to other terms which are both fragments of a common longer term. Once we have found all terms of which the **focal term** is a sub-string, we then find all terms that are sub-strings of those terms. In this way we may end up with some terms that do not overlap with the **focal term**, but are substrings of longer terms that do overlap with the focal term. We then get the unique terms out of this list of candidate terms before proceeding to the next step.

Next, we calculate the correlation coefficient of the raw document term frequencies (in the subset of documents associated with the groups being compared – so in our running example, the 1,097 health care related bills introduced in Congress from 2013-2014) between the **focal term** and the other candidate terms identified through sub-string relationships as described above. The reason we only calculate pairwise correlations with the **focal term** and do not look at correlations between all terms is that we really do only want to subsume terms that are specifically highly correlated with the **focal term**. Otherwise, it could be the case that we end up subsuming terms which are only related to the focal term through a chain of correlations but are not highly (directly) correlated with the **focal term**. I believe adopting this approach is more conservative than a "connected components in the correlation graph" approach because it will tend to subsume fewer terms at each iteration, leading to an increased possibility of terms that share a sub-string relationship being included in the resulting list of top terms, but a decreased possibility of subsuming terms that really represent a distinct concept.

We then keep candidate terms in a cluster with the **focal term** if their document frequencies are correlated with those of the **focal term** at above some threshold (in the examples here: 0.9). It will likely be application dependent what the optimal threshold should be, and I intend to investigate this further in the future. Finally, we remove the terms we are including in a cluster with the **focal term** from the ranked list of terms we use as input before proceeding to the next iteration – ths ensuring that those terms are not included in any of the later term clusters. To select the representative term for each cluster (to present to the user) we choose the term that consists of the largest number of characters within each cluster. An example of terms that were considered for inclusion in a term cluster with "health insurance" as the **focal term** is provided in Table 1. As we can see, some terms are included that do not overlap at all with the **focal term**, such as "coverage offered", but are extremely highly correlated via a common parent term, which is "health insurance coverage offered". I feel that this approach strikes the right balance between specificity and coverage when considering which terms to cluster together, leading to a highly interpretable output list of ranked terms with distinct meanings.

I applied this method to the top phrases associated with Democratic and Republican healthcare bills generated by the informed Dirichlet model, presented in Tables 8 and 9. The results with term clustering presented in Tables 10 and 11 show that this method highlights meaningful top phrase clusters, each of which has a distinct meaning. This improves the interpretability of these lists by eliminating large numbers of closely related terms. For example, a number of sub-strings of term "patient protection and affordable care act" in the top Republican phrases are now combined together. The top twenty Republican phrase clusters now present much more unique information about top Republican terms instead of simply repeating sub-strings of one term.

6 Comparison Between Unigrams and Phrases

Having addressed the issue of duplication in the top phrases associated with Democrat and Republican sponsored bills related to health care introduced between 2013 and 2014, we can now attempt to interpret the phrase results, and seek to compare phrases and unigrams in this domain. Monroe et al. (2008) compare top terms associated with Democrats and Republicans using funnel plots (Spiegelhalter, 2005), and I provide a similar comparison of phrases associated with Democrat and Republican sponsored bills related to health care in Figure 1. In this plot, each term that appears at least once in the 1,097 health care bills use in the analysis is plotted as a dot with the x-coordinate representing its total count in

Table 1: Example terms associated with **focal term** *health insurance*. Terms that were included in a cluster with *health insurance* based on a correlation threshold of 0.9 are highlighed in blue.

Term	Correlation with **focal term**	Included in Cluster
health insurance	1.00000000	Yes
health insurance coverage	0.98720755	Yes
health insurance issuer	0.98079061	Yes
individual health insurance	0.88703801	No
individual health insurance coverage	0.88516437	No
health insurance coverage offered	0.95494809	Yes
group health insurance	0.30958607	No
health insurance issuers	0.75753589	No
group health insurance coverage	0.29272403	No
health insurance mandate	0.02245245	No
insurance coverage	0.98721909	Yes
insurance issuer	0.98083056	Yes
individual health	0.88713052	No
coverage offered	0.96951616	Yes
insurance coverage offered	0.95507076	Yes
group health	0.89240273	No
insurance issuers	0.76651233	No
insurance mandate	0.02245245	No

those bills, and its y-coordinate representing its z-score. Terms in gray have z-scores whose absolute value is less than 1.96, while terms in black have z-scores whose absolute value is greater than or equal 1.96. The dots highlighted in blue (red) are associated with the top 20 Democrat (Republican) term clusters displayed in the right margin, where the top (bottom) terms have the largest magnitude z-scores.

Before drawing any conclusions from these lists of terms, I looked at the titles of bills associated with high counts of each of the top twenty phrase clusters for Democrats and Republicans to verify my interpretations. The resulting substantive conclusions we can draw from examination of Figure 1 are much clearer than in the case of unigrams: Republicans were much more focussed on the financial aspects of the health care system (particularly as they relate to repealing the Affordable Care Act), while Democrats were more focussed on introducing legislation related to actual healthcare provision and public health. One particularly interesting term: "*acting through the director of the centers for disease control*" came up repeatedly in the context of Democrat sponsored bills directing the CDC to study some public health issue. These issues were incredibly numerous and included everything from surveillance of the West-African Ebola outbreak, to breast-cancer studies, to monitoring the effects of drinking water quality on health.

7 Conclusion

The informed Dirichlet model for feature selection introduced by Monroe et al. (2008) is an effective but poorly understood method for finding terms that distinguish between two sets of documents. In this document, I re-derive the entire model, and explore its functionality in much greater depth than Monroe et al. do in their original paper. I find that the performance of competing methods such as PMI and TF-IDF based ranking is much more similar to that of the informed Dirichlet model than the authors of the original paper would have us believe, but that the informed Dirichlet model seems to offer performance that is at least as good as these methods in most applications, and significantly better in some settings. I apply this method to a corpus of congressional texts and find that the key innovation associated with highly interpretable results is the use of phrases as the unit of analysis instead of unigrams. In order to deal with the duplication issues associated phrases that overlap, I introduce a novel correlation based

Figure 1: Funnel plot of top phrase clusters (after applying correlation based term subsumption) in health care legislation introduced by Democrats and Republicans between 2013 and 2014 as ranked by the informed Dirichlet model.

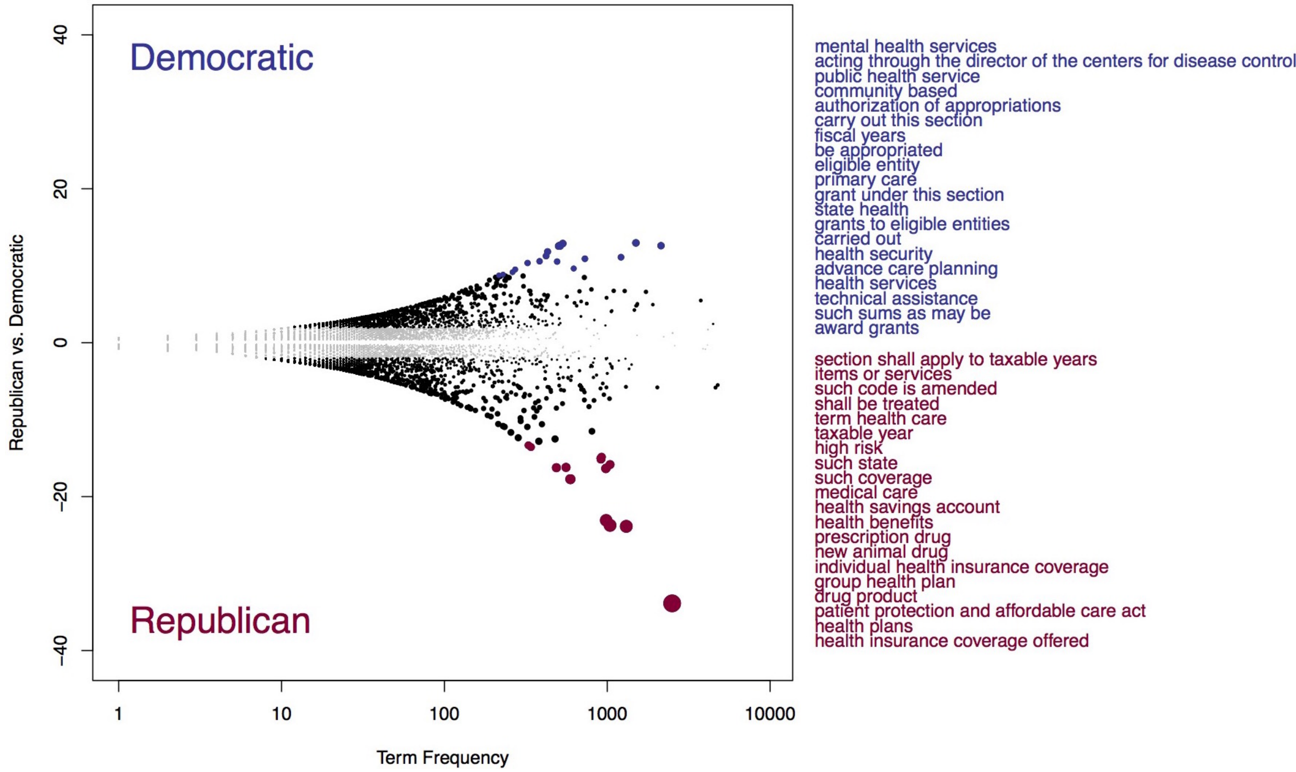
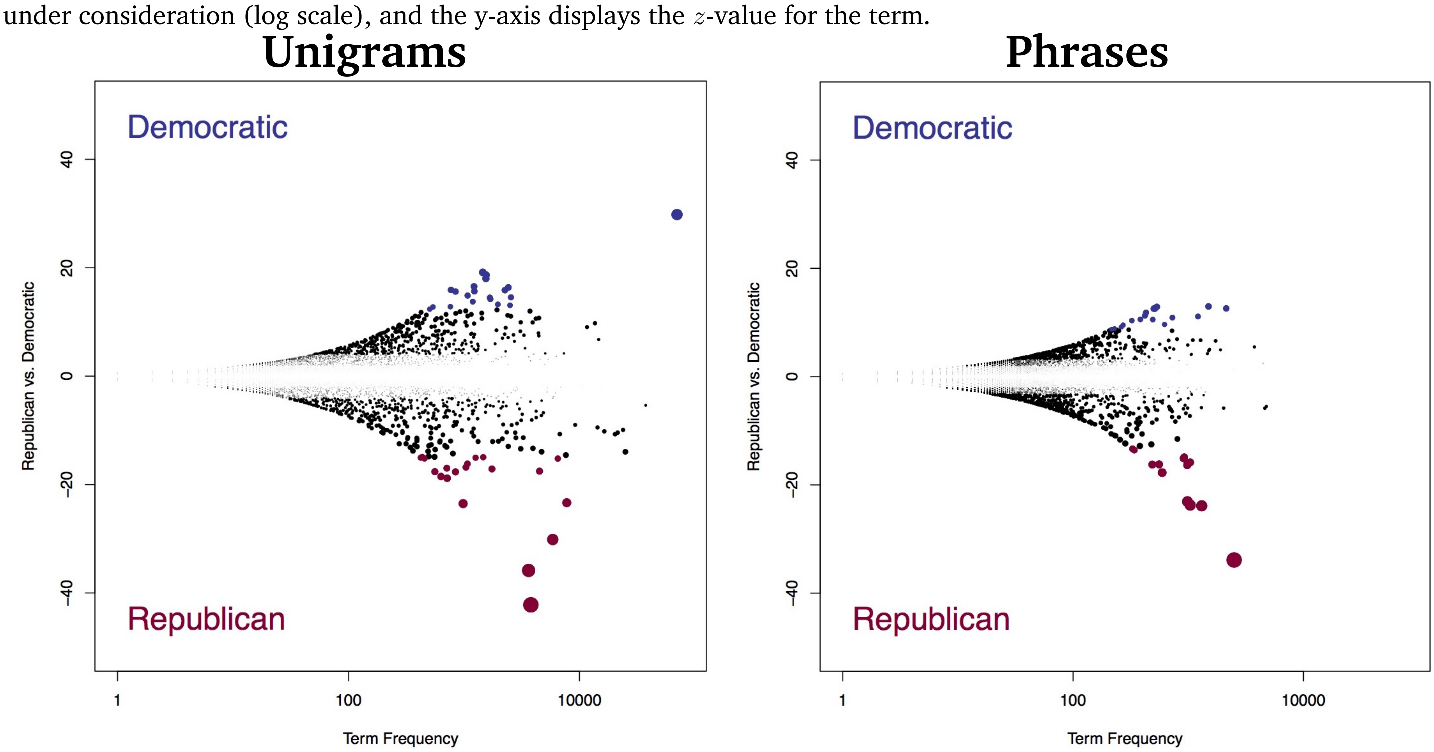


Figure 2: Funnel plots comparing unigrams and phrases associated with Democrat and Republican sponsored bills about health care introduced between 2013-2014. The x-axis in these plots is the number of times a term appeared in the 1097 bills under consideration (log scale), and the y-axis displays the z-value for the term.



phrase subsumption algorithm, which I apply to top phrases associated with healthcare bills introduced by Democrats and Republicans during the 2013-2014 legislative session. My results indicate that Republicans tend to focus much more heavily on repealing Obamacare during this time period, while Democrats are focused more heavily on standard healthcare issues, which is consistent with the popular accounts of the parties healthcare policy during this period. Future work could apply this method in other domains and extend the model by considering other priors, but I feel that this work still makes a contribution to the literature on feature selection, particularly for political texts.

References

Breslow, N.E. and N.E. Day. Classical Methods of Analysis of Grouped Data. *Statistical Methods in Cancer Research*, pages 121–159, 1980.

Denny, Matthew J., Brendan O. Connor, and Hanna Wallach. A Little Bit of NLP Goes A Long Way: Finding Meaning in Legislative Texts with Phrase Extraction. In *Midwest Political Science Association Anual Meeting*, 2015.

Manning, Christopher D and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008. http://www-nlp.stanford.edu/IR-book/.

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

Morris, J a and M J Gardner. Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British medical journal (Clinical research ed.)*, 296(6632):1313–1316, 1988.

Purpura, Stephen and Dustin Hillard. Automated Classification of Congressional Legislation. *Proceedings of the 2006 international conference on Digital government research*, pages 219–225, 2006. http://www.purpuras.net/dgo2006PurpuraHillardClassifyingCongressionalLegislation.pdf.

Spiegelhalter, David J. Funnel plots for comparing institutional performance. *Statistics in Medicine*, 24 (8):1185–1202, 2005.

Table 2: Top unigrams in bills about health care policy sponsored by **Democrats** (2013–2014) under three different ranking methods.

Rank	PMI	TF-IDF	TF-IDF with log(TF)	Dirichlet
1	school	health	mips	and
2	minority	care	patient	deleted
3	local	services	drug	prevention
4	planning	drug	cancer	grant
5	indian	mips	mental	grants
6	nursing	medical	care	programs
7	populations	patient	hospital	school
8	students	secretary	physician	research
9	work	medicare	medicaid	training
10	grants	social	medicare	local
11	guidance	mental	diabetes	cancer
12	prevention	such	veterans	national
13	living	program	medical	disease
14	youth	veterans	disease	centers
15	women	hospital	clinical	community
16	about	under	health	activities
17	diabetes	eligible	patients	education
18	grant	professional	professional	diabetes
19	cancer	data	deleted	tobacco
20	carried	physician	social	minority

Table 3: Descriptive statistics associated with top unigrams in bills about health care policy sponsored by **Democrats** (2013–2014) using informed Dirichlet Ranking.

Term	Log-Odds Ratio	Variance	z-Scores	Democrat Count	Republican Count
and	0.24	0.00	29.81	44706	25204
deleted	3.54	0.03	19.15	1434	20
prevention	1.27	0.00	18.64	1295	258
grant	1.20	0.00	17.99	1276	272
grants	1.27	0.01	16.54	1019	202
programs	0.77	0.00	16.35	1823	598
school	2.14	0.02	15.90	714	58
research	0.77	0.00	15.86	1703	557
training	1.15	0.01	15.66	1005	224
local	1.64	0.01	15.61	745	100
cancer	1.18	0.01	14.87	883	192
national	0.65	0.00	14.54	1861	689
disease	0.83	0.00	14.49	1280	395
centers	0.81	0.00	14.20	1285	407
community	0.97	0.01	13.75	940	251
activities	0.67	0.00	13.23	1444	520
education	0.58	0.00	13.07	1791	712
diabetes	1.23	0.01	12.81	631	131
tobacco	2.94	0.05	12.76	520	19
minority	1.78	0.02	12.36	451	54

Table 4: Top unigrams in bills about health care policy sponsored by **Republicans** (2013–2014) under three different ranking methods.

Rank	PMI	TF-IDF	TF-IDF with log(TF)	Dirichlet
1	issuer	health	coverage	insurance
2	sponsor	care	patient	coverage
3	claim	coverage	insurance	plan
4	court	insurance	issuer	issuer
5	premium	plan	drug	any
6	arrangement	patient	physician	association
7	met	medical	medicare	claim
8	association	medicare	hospital	claims
9	contribution	services	care	sponsor
10	insurance	drug	medical	individual
11	taxpayer	such	health	benefits
12	market	social	prescription	employer
13	loss	hospital	affordable	affordable
14	party	payment	professional	authority
15	connection	issuer	plan	which
16	employer	under	provider	arrangement
17	offered	physician	clinical	protection
18	liability	individual	medicaid	damages
19	employers	secretary	social	group
20	spending	eligible	mips	premium

Table 5: Descriptive statistics associated with top unigrams in bills about health care policy sponsored by **Republicans** (2013–2014) using informed Dirichlet Ranking.

Term	Log-Odds Ratio	Variance	z-Scores	Republican Count	Democrat Count
insurance	1.69	0.00	42.19	3013	782
coverage	1.34	0.00	35.84	2650	977
plan	0.81	0.00	30.14	3604	2265
issuer	2.17	0.01	23.51	847	136
any	0.53	0.00	23.35	4246	3516
association	1.80	0.01	18.84	580	135
claim	2.02	0.01	18.52	533	99
claims	1.38	0.01	17.66	624	220
sponsor	2.12	0.01	17.63	478	81
individual	0.52	0.00	17.53	2458	2050
benefits	0.84	0.00	17.12	1088	658
employer	1.49	0.01	16.98	539	171
affordable	1.12	0.00	16.80	713	327
authority	1.05	0.00	16.18	718	353
which	0.38	0.00	15.21	3302	3191
arrangement	1.84	0.01	15.18	372	83
protection	0.88	0.00	15.05	795	465
damages	3.14	0.04	15.04	399	24
group	0.80	0.00	14.95	899	568
premium	1.86	0.02	14.91	357	78

Table 6: Top phrases in bills about health care policy sponsored by **Democrats** (1993–2014) under three different ranking methods.

Rank	PMI	Dirichlet
1	centers for disease control and prevention	mental health
2	control and prevention	disease control
3	disease control and prevention	control and prevention
4	disease control	public health
5	centers for disease	centers for disease control
6	centers for disease control	centers for disease
7	authorization of appropriations	centers for disease control and prevention
8	carry out this section	disease control and prevention
9	carried out	community based
10	be appropriated	authorization of appropriations
11	are authorized	carry out
12	services administration	fiscal years
13	fiscal years	carry out this section
14	primary care	be appropriated
15	shall develop	eligible entity
16	institutes of health	primary care
17	national institutes of health	grant under this section
18	national institutes	state health
19	substance abuse	director of the centers
20	evidence based	eligible entities
Rank	TF-IDF	TF-IDF with log(TF)
1	act u.s.c.	social security act u.s.c.
2	social security act u.s.c.	security act u.s.c.
3	security act u.s.c.	act u.s.c.
4	health care	health care
5	social security act	mental health
6	social security	social security act
7	security act	subparagraph a
8	mental health	social security
9	public health	health service act u.s.c.
10	subparagraph a	public health service act u.s.c.
11	health service	service act u.s.c.
12	health service act u.s.c.	security act
13	public health service act u.s.c.	public health
14	service act u.s.c.	health service
15	public health service	public health service
16	health service act	health service act
17	public health service act	public health service act
18	veterans affairs	veterans affairs
19	u.s.c. w	u.s.c. w
20	act u.s.c. w	act u.s.c. w

Table 7: Top phrases in bills about health care policy sponsored by **Republicans** (1993–2014) under three different ranking methods.

Rank	PMI	Dirichlet
1	insurance coverage	health insurance
2	health insurance coverage	health plan
3	health insurance issuer	insurance coverage
4	insurance issuer	health insurance coverage
5	group health	insurance issuer
6	health insurance	health insurance issuer
7	health plan	affordable care act
8	group health plan	affordable care
9	health benefits	drug product
10	such state	group health
11	high risk	care act
12	health plans	patient protection and affordable care act
13	medical care	protection and affordable care act
14	affordable care act	patient protection and affordable care
15	affordable care	protection and affordable care
16	code is	patient protection and affordable
17	taxable year	protection and affordable
18	such code	patient protection
19	protection and affordable care act	individual health
20	patient protection and affordable care act	individual health insurance
Rank	**TF-IDF**	**TF-IDF with log(TF)**
1	health insurance	social security act u.s.c.
2	act u.s.c.	security act u.s.c.
3	social security act u.s.c.	act u.s.c.
4	security act u.s.c.	health insurance
5	health care	health insurance coverage
6	health plan	health plan
7	social security act	affordable care act
8	social security	affordable care
9	health insurance coverage	insurance coverage
10	security act	protection and affordable care act
11	insurance coverage	patient protection and affordable care act
12	affordable care act	protection and affordable care
13	affordable care	patient protection and affordable care
14	prescription drug product	protection and affordable
15	protection and affordable care act	patient protection and affordable
16	patient protection and affordable care act	health care
17	protection and affordable care	patient protection
18	patient protection and affordable care	drug product
19	protection and affordable	prescription drug product
20	patient protection and affordable	care act

Table 8: Descriptive statistics associated with top phrases in bills about health care policy sponsored by **Democrats** (2013–2014) using informed Dirichlet Ranking.

Term	Log-Odds Ratio	Var.	z-Scores	Dem. Count	Rep. Count
mental health	0.76	0.00	12.96	1100	399
disease control	1.68	0.02	12.88	466	67
control and prevention	1.70	0.02	12.63	446	63
public health	0.60	0.00	12.60	1498	639
centers for disease control	1.66	0.02	12.56	447	66
centers for disease	1.66	0.02	12.56	447	66
centers for disease control and prevention	1.70	0.02	12.55	440	62
disease control and prevention	1.69	0.02	12.54	441	63
community based	1.79	0.02	11.82	381	49
authorization of appropriations	1.61	0.02	11.26	365	56
carry out	0.71	0.00	11.11	881	333
fiscal years	0.96	0.01	10.91	563	166
carry out this section	1.55	0.02	10.58	330	54
be appropriated	1.22	0.01	10.55	401	91
eligible entity	2.90	0.08	10.34	311	13
primary care	0.91	0.01	9.63	474	148
grant under this section	1.86	0.04	9.51	243	29
state health	1.75	0.04	9.16	231	31
director of the centers	1.98	0.05	8.85	207	22
eligible entities	2.26	0.07	8.74	200	16

Table 9: Descriptive statistics associated with top phrases in bills about health care policy sponsored by **Republicans** (2013–2014) using informed Dirichlet Ranking.

Term	Log-Odds Ratio	Var.	z-Scores	Rep. Count	Dem. Count
health insurance	1.75	0.00	33.88	2042	460
health plan	1.66	0.00	23.87	1049	259
insurance coverage	2.40	0.01	23.73	931	109
health insurance coverage	2.39	0.01	23.08	880	104
insurance issuer	2.18	0.02	17.74	519	76
health insurance issuer	2.20	0.02	17.73	518	74
affordable care act	1.15	0.00	16.36	695	284
affordable care	1.15	0.00	16.32	696	286
drug product	2.54	0.02	16.24	442	45
group health	1.80	0.01	16.21	460	98
care act	1.06	0.00	15.84	719	321
patient protection and affordable care act	1.09	0.01	15.14	637	277
protection and affordable care act	1.09	0.01	15.14	637	277
patient protection and affordable care	1.09	0.01	15.11	637	278
protection and affordable care	1.09	0.01	15.11	637	278
patient protection and affordable	1.08	0.01	15.03	634	278
protection and affordable	1.08	0.01	15.03	634	278
patient protection	1.06	0.01	14.86	637	286
individual health	2.39	0.03	13.56	304	36
individual health insurance	2.55	0.04	13.33	298	30

Table 10: Descriptive statistics associated with top phrases after the application of phrase subsumption in bills about health care policy sponsored by **Democrats** (2013–2014) using informed Dirichlet Ranking.

Term	Log-Odds Ratio	Var.	z-Scores	Dem. Count	Rep. Count	Terms in Cluster
mental health services	0.76	0.00	12.96	1100	399	2
acting through the director of the centers for disease control	1.68	0.02	12.88	466	67	14
public health service	0.60	0.00	12.60	1498	639	3
community based	1.79	0.02	11.83	381	49	1
authorization of appropriations	1.61	0.02	11.26	365	56	1
carry out this section	0.71	0.00	11.10	881	333	2
fiscal years	0.96	0.01	10.91	563	166	1
be appropriated	1.22	0.01	10.54	401	91	1
eligible entity	2.89	0.08	10.36	311	13	1
primary care	0.91	0.01	9.64	474	148	1
grant under this section	1.86	0.04	9.51	243	29	1
state health	1.75	0.04	9.16	231	31	1
grants to eligible entities	2.26	0.07	8.74	200	16	2
carried out	1.31	0.02	8.66	252	52	1
health security	3.09	0.13	8.64	230	8	1
advance care planning	2.15	0.06	8.50	189	17	3
health services	0.70	0.01	8.47	523	200	1
technical assistance	1.51	0.03	8.46	214	36	1
such sums as may be	1.53	0.03	8.46	212	35	3
award grants	1.74	0.05	8.22	186	25	1

Table 11: Descriptive statistics associated with top phrases after the application of phrase subsumption in bills about health care policy sponsored by **Republicans** (2013–2014) using informed Dirichlet Ranking.

Term	Log-Odds Ratio	Var.	z-Scores	Rep. Count	Dem. Count	Terms in Cluster
health insurance coverage offered	1.75	0.00	33.87	2042	460	8
health plans	1.65	0.00	23.86	1049	259	2
patient protection and affordable care act	1.15	0.00	16.35	695	284	10
drug product	2.54	0.02	16.24	442	45	1
group health plan	1.80	0.01	16.21	460	98	2
individual health insurance coverage	2.39	0.03	13.56	304	36	3
new animal drug	2.78	0.06	11.66	237	19	3
prescription drug	0.85	0.01	11.51	518	287	1
health benefits	1.43	0.02	10.94	247	76	1
health savings account	2.99	0.08	10.84	215	14	3
medical care	1.18	0.01	10.59	284	113	1
such coverage	2.11	0.04	10.56	185	29	1
such state	1.38	0.02	10.19	222	72	1
high risk	1.31	0.02	9.80	217	76	1
taxable year	1.11	0.01	9.64	251	106	1
term health care	1.54	0.03	9.24	166	46	2
shall be treated	0.98	0.01	8.87	249	120	2
such code is amended	1.09	0.02	8.81	215	93	5
items or services	1.95	0.05	8.80	131	24	1
section shall apply to taxable years	1.46	0.03	8.63	151	45	7

Table 12: Top terms positively associated with bill passage out of committee as ranked by impact score.

Unigrams		Phrases (2+ Tokens)	
Term	Impact	Term	Impact
repackage	0.00093	45d may be carried back to a taxable year	0.00111
natos	0.00092	low population	0.00070
trying	0.00082	purposes of payments	0.00054
carson	0.00049	president is	0.00040
olds	0.00046	in the house of representatives june 12	0.00039
vacation	0.00045	section shall terminate on december	0.00038
mccaskill	0.00043	42 u.s.c. 300k	0.00036
subassembly	0.00043	participating in the medicare program	0.00029
surges	0.00043	201 (b) of the federal	0.00027
intimidate	0.00038	section 45c	0.00024
proprietors	0.00035	u.s.c. 2135	0.00023
honoraria	0.00033	program under this section shall submit	0.00023
climb	0.00033	legislative day	0.00022
mack	0.00033	base period shall be the calendar	0.00019
accountability	0.00027	local services	0.00019

Table 13: Top terms negitively associated with bill passage out of committee as ranked by impact score.

Unigrams		Phrases (2+ Tokens)	
Term	Impact	Term	Impact
taxable	-0.00425	members appointed	-0.00081
closing	-0.00146	state law	-0.00064
offenders	-0.00083	environmental protection agency	-0.00060
provisions	-0.00073	shall hold	-0.00058
relationships	-0.00061	has not	-0.00038
jailed	-0.00052	fund to be known	-0.00024
fisheries	-0.00036	5 of the federal trade commission act (15 u.s.c. 45)	-0.00024
dues	-0.00035	july 25	-0.00021
liberty	-0.00035	redesignating paragraphs	-0.00019
technique	-0.00033	shall be made	-0.00018
hundreds	-0.00025	will be met	-0.00009
discovered	-0.00025	travel expense	-0.00009
nomination	-0.00025	fisheries research	-0.00009
school	-0.00020	product shall be	-0.00005
quantified	-0.00016	require federal agencies	-0.00004