# Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge

Ryan J. Gallagher[1,2], Kyle Reing[1], David Kale[1], and Greg Ver Steeg[1]

[1]**Information Sciences Institute**
University of Southern California
Marina del Rey, CA 90292,
[2]**Vermont Complex Systems Center and Computational Story Lab**
University of Vermont
Burlington, VT 05405

## Abstract

Popular approaches to topic modeling often invoke the use of probabilistic generative models, such as Latent Dirichlet Allocation (LDA). While such models have enjoyed widespread use and proven fruitful, specifying these models or generalizing them to incorporate human input requires detailed and often unrealistic assumptions about the data generating process. We introduce a new approach to topic modeling via Correlation Explanation (CorEx), which leverages an information-theoretic framework to bypass typical topic modeling assumptions. Using two challenging, real-world datasets, we demonstrate that CorEx yields results that are comparable to LDA in terms of semantic coherence and document classification. We then devise a flexible methodology for incorporating word-level domain knowledge into CorEx by introducing anchor words in a manner reminiscent of the information bottleneck. Augmenting CorEx with anchor words allows the topic model to be guided with minimal human intervention towards topics that do not naturally emerge. Furthermore, we show that these new topics are often highly coherent and act as better predictors in document classification.

## 1 Introduction

Unsupervised extraction of themes from documents, such as books, articles, and microblogs, is a common challenge in many fields of research. This challenge has largely been met by topic modeling, the process of grouping semantically coherent sets of words into "topics" in order to facilitate text summarization and document classification. The adaptability of topic modeling algorithms to different domains has prompted rich research investigations into the evolution of scientific knowledge (Griffiths and Steyvers, 2004), change in discourse surrounding climate change (Cody et al., 2016), and authorial anomalies in fictional works (Steyvers et al., 2004). As it stands, topic modeling is one of the most popular ways of extracting information from unstructured textual data.

Two methodologies largely dominate topic modeling: matrix factorization, such as Latent Semantic Indexing (LSI) (Deerwester et al., 1990; Landauer et al., 1998), and probabilistic generative models, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Generative models, and LDA in particular, have eclipsed topic modeling research and applications. LDA specifies a document generation process: it is assumed that for each document a topic is randomly chosen from a specified distribution, and then a word is randomly chosen according to a distribution specified by the chosen topic. The document-topic and topic-word distributions that generate the document are unknown, but can be inferred using Bayesian inference.

Of course, researchers do not believe that documents are written by iteratively drawing random words from random topics. However, the specification of distributions and parameters, a problem that is compounded when domain knowledge comes into play, is perhaps undesirable in contexts where one wishes to naturally uncover word and document relationships with minimal human input. For this reason, we propose topic modeling using Correlation Explanation (CorEx)[1], an information-theoretic approach to learning latent topics over documents. Unlike LDA, or even matrix factorization techniques which have subtle ties to generative models (Hofmann, 1999; Ding et al., 2008), CorEx does not assume a particular data generating model.

The information-theoretic framework behind CorEx also naturally allows for flexible incorporation of word-level domain knowledge. Topic models are often susceptible to portraying only dominant themes of documents. Injecting a topic model, such as CorEx, with domain knowledge can help guide it towards otherwise underrepresented topics that are pertinent to the user. This can be useful, for example, if we wish to learn to automatically diagnose patients from medical notes written by their doctor. By incorporating word level domain knowledge, we might encourage our topic model to recognize a rare disease that would otherwise be missed. Alternately, if we have documents that relate to some natural disaster, we may want to focus our attention on topics that could guide relief workers to distribute aid more effectively.

Our contributions are as follows. First, we frame CorEx as a topic model and derive an efficient adaptation to the CorEx algorithm to exploit sparse data, such as word counts in documents, for dramatic speedups. Second, we demonstrate that CorEx is competitive with LDA in topic model quality according to several measures. Finally, we show that domain knowledge can be naturally integrated into CorEx using "anchor words". Using two challenging, real-world problem domains, we perform an extensive analysis detailing the effects and benefits of using anchored CorEx.

## 2 Related Work

With respect to integrating domain knowledge into topic models, we draw inspiration from Arora et al., who used anchor words in the context of non-negative matrix factorization (2012). Using an assumption of separability, these anchor words act as high precision markers of particular topics and, thus, help discern the topics from one another. Although the original algorithm proposed by Arora et. al and subsequent improvements to the algorithm find these anchor words automatically (Arora et al., 2013; Lee and Mimno, 2014), recent adaptations allow manual insertion of anchor words and other metadata (Nguyen et al., 2014; Nguyen et al., 2015). Our work is similar to the latter, where we treat anchor words as fuzzy logic markers and embed them into the topic model in a semi-supervised fashion. In this sense, our work is closest to Halpern et al., who have also made use of domain expertise and semi-supervised anchored words in devising topic models (2014; 2015).

There is an adjacent line of work that has focused on incorporating word-level information into LDA-based models. Andrezejewski and Zhu have presented two flavors of such models. One allows specification of Must-Link and Cannot-Link relationships between words that help partition otherwise muddled topics (Andrzejewski et al., 2009). The other model makes use of "$z$-labels," words that are known to pertain to a specific topics and that are restricted to appearing in some subset of all the possible topics (Andrzejewski and Zhu, 2009). Similarly, Jagarlamudi et. al proposed SeededLDA, a model that seeds words into given topics and guides, but does not force, these topics towards these integrated words (2012). While we also seek to guide our model towards topics containing user-provided words, our model naturally extends to incorporating such information, while the LDA-based models require involved and careful construction of new assumptions. Thus, our framework is more lightweight and flexible than LDA-based models.

Mathematically, CorEx topic models most closely resemble topic models based on latent tree reconstruction (Chen et al., 2015). In Chen et. al.'s analysis, their own latent tree approach and CorEx both report significantly better perplexity than hierarchical topic models based on the hierarchical Dirichlet process and the Chinese restaurant process but they showed that their method was much faster than CorEx. We revisit this comparison after introducing our new formulation exploiting sparsity in Sec. 3.3. CorEx has also been investigated as a way to find "surprising" documents (Hodas et al., 2015).

[1]Open source code to run the CorEx topic model can be found at `https://github.com/gregversteeg/corex_topic`

## 3 Methods

### 3.1 Correlation Explanation

Here we review the fundamentals of Correlation Explanation (CorEx), largely adopting the notation used by Ver Steeg and Galstyan in their original presentation of the model (2014). Let $X$ be a discrete random variable that takes on a finite number of values. Furthermore, if we have $n$ such random variables, let $X_G$ denote a subcollection of them, where $G \subseteq \{1, \dots, n\}$. The entropy of $X$ is written as $H(X)$ and the mutual information of two random variables $X_1$ and $X_2$ is given by $I(X_1 : X_2) = H(X_1) + H(X_2) - H(X_1, X_2)$.

The total correlation, or multivariate mutual information, of a group of random variables $X_G$ is expressed as

$$TC(X_G) = \sum_{i \in G} H(X_i) - H(X_G) \tag{1}$$

$$= D_{KL}\left(p(X_G) || \prod_{i \in G} p(X_i)\right). \tag{2}$$

We see that Eqn. 1 does not quantify "correlation" in the modern sense of the word, and so it can be helpful to conceptualize total correlation as a measure of total dependence. Indeed, Eqn. 2 shows that total correlation can be expressed using the Kullback-Leibler Divergence and, therefore, it is zero if and only if the joint distribution of $X_G$ factorizes, or, in other words, there is no dependence between the random variables.

The total correlation can be written when conditioning on another random variable $Y$, $TC(X_G \mid Y) = \sum_{i \in G} H(X_i \mid Y) - H(X_G \mid Y)$. So, we can consider the reduction in the total correlation when conditioning on $Y$.

$$TC(X_G; Y) = TC(X_G) - TC(X_G \mid Y) \tag{3}$$

$$= \sum_{i \in G} I(X_i : Y) - I(X_G : Y) \tag{4}$$

This measures how much $Y$ explains the dependencies in $X_G$. The quantity expressed in Eqn. 3 acts as a lower bound of $TC(X_G)$ (Ver Steeg and Galstyan, 2015), as readily verified by noting that $TC(X_G)$ and $TC(X_G|Y)$ are always non-negative. Also note, the joint distribution of $X_G$ factorizes conditional on $Y$ if and only if $T(X_G \mid Y) = 0$. If this is the case, then $TC(X_G; Y)$ is maximized.

In the context of topic modeling, $X_G$ represents a group of words and $Y$ represents a topic. Since we are always interested in grouping multiple sets of words into multiple topics, we will denote the latent topics as $Y_1, \dots Y_m$ and their corresponding groups of words as $X_{G_j}$ for $j = 1, \dots, m$ respectively. The CorEx topic model seeks to maximally explain the dependencies of words in documents through latent topics by maximizing $TC(X; Y_1, \dots, Y_m)$. Instead, we maximize the following lower bound on this expression:

$$\max_{G_j, p(y_j | x_{G_j})} \sum_{j=1}^{m} TC(X_{G_j}; Y_j). \tag{5}$$

This optimization is subject to the constraint that the groups, $G_j$, do not overlap and the conditional distribution is normalized. The solution to this objective can be efficiently approximated, despite the search occurring over an exponentially large probability space (Ver Steeg and Galstyan, 2014).

The latent factors, $Y_j$, are optimized to be informative about dependencies in the data and do not require generative modeling assumptions. Note that the discovered factors, $Y$, can be used as inputs to construct new latent factors, $Z$, and so on leading to a hierarchy of topics. Although this extension is quite natural, we focus our analysis on the first level of topic representations for easier interpretation and evaluation.

### 3.2 Anchor Words via the Bottleneck

The information bottleneck formulates a trade-off between compressing data $X$ into a representation $Y$, and preserving the information in $X$ that is relevant to $Z$ (typically labels in a supervised learning task) (Tishby et al., 2000; Friedman et al., 2001). More formally, the information bottleneck is expressed as

$$\max_{p(y|x)} \beta I(Z : Y) - I(X : Y), \tag{6}$$

where $\beta$ is a parameter controlling the trade-off between compressing $X$ and preserving information about $Z$.

To see the connection with CorEx, we rewrite the objective of Eqn. 5 by following the derivation of

Ver Steeg and Galstyan (Ver Steeg and Galstyan, 2014) and introducing indicator variables $\alpha_{i,j}$ which are equal to 1 if and only if word $X_i$ appears in topic $Y_j$ (i.e. $i \in G_j$).

$$\max_{\alpha_{i,j}, p(y_j|x)} \sum_{j=1}^{m} \left( \sum_{i=1}^{n} \alpha_{i,j} I(X_i : Y_j) - I(X : Y_j) \right) \tag{7}$$

Note that the constraint on non-overlapping groups now becomes a constraint on $\alpha$. Comparing the objective to Eqn. 6, we see that we have exactly the same compression term for each latent factor, $I(X : Y_j)$, but the relevance variables now correspond to $Z \equiv X_i$. Inspired by the success of the bottleneck, we suggest that if we want to learn representations that are more relevant to specific keywords, we can simply anchor a word $X_i$ to topic $Y_j$, by constraining our optimization so that $\alpha_{i,j} = \beta_{i,j}$, where $\beta_{i,j} \geq 1$ controls the anchor strength. Otherwise, the updates on $\alpha$ remain the same as in Ver Steeg and Galstyan's original presentation (2014). This schema is a natural extension of the CorEx objective and it is flexible, allowing for multiple words to be anchored to one topic, for one word to be anchored to multiple topics, or for any combination of these anchoring strategies. Furthermore, it combines supervised and unsupervised learning by allowing us to leave some topics without anchors.

### 3.3 CorEx Sparsity Optimization

We now assume that all variables, $x_i, y_j$, are binary and the vector $x$ is a binary bag of words vector where $X_i^\ell = 1$ if word $i$ occurs in document $\ell$ and $X_i^\ell = 0$ otherwise. We want to alter the CorEx optimization procedure to exploit the sparsity in the data. The numerical optimization for CorEx involves iteratively updating a fixed point equation until convergence. Similar to the EM algorithm, we start with a random soft labeling for each document and each latent factor at time $t = 0$, $p_{t=0}(y_j|x^\ell)$. Next we update the marginal distributions $p_t(x_i, y_j)$ and the $\alpha_{i,j}^t$ using the original CorEx procedure. Note that since all variables are binary, the marginal distribution is just a two by two table of probabilities and can be estimated efficiently. The time-consuming part of training is the subsequent update of the document labels.

$$\log p_{t+1}(y_j|x^\ell) = \tag{8}$$
$$\log p_t(y_j) + \sum_{i=1}^{n} \alpha_{i,j}^t \log \frac{p_t(x_i^\ell \mid y_j)}{p(x_i^\ell)} - \log Z_j(x^\ell)$$

for each document $\ell$ (Ver Steeg and Galstyan, 2014). The computation of the log likelihood ratio for all $n$ words over all documents is not efficient, as most words do not appear in a given document. We rewrite the logarithm in the interior of the sum.

$$\log \frac{p_t(x_i^\ell \mid y_j)}{p(x_i^\ell)} = \log \frac{p_t(X_i = 0 \mid y_j)}{p(X_i = 0)} + \tag{9}$$
$$x_i^l \log \left( \frac{p_t(X_i^\ell = 1 \mid y_j) p(X_i = 0)}{p_t(X_i = 0 \mid y_j) p(X_i^\ell = 1)} \right)$$

Note, when the word does not appear in the document, only the leading term of Eqn. 9 will be nonzero. However, when the word does appear, everything but $\log P(X_i^\ell = 1 \mid y_j)/p(X_i^\ell = 1)$ cancels out. So, we have taken advantage of the fact that the CorEx topic model binarizes documents to, by default, assume the word does not appear in the document, and correct the contribution to the update if the word does appear.

Thus, when substituting back into Eqn. 8, the sum becomes a matrix multiplication between a matrix with dimensions of number of variables by number of documents and entries $x_i^l$ that is assumed to be sparse and a dense matrix with dimensions of number of variables by number of latent factors. This results in significant speedups for the CorEx algorithm.

**Running time evaluation** Chen et. al. compared running times of CorEx without sparsity speedups versus several hierarchical topic models (2015). The largest comparison used the twenty newsgroups dataset with a 5000 word vocabulary (Table 6 in their paper) and found that CorEx took about 3 days while their method took over 7 hours. Exploiting sparsity as described, we were able to run CorEx on the same experiment in about 45 minutes. Even accounting for minor variations in systems (our experiment was on a mid-2012 Macbook Pro while they used an unspecified desktop computer), it is clear that exploiting sparsity confers large computational benefits. Given $n$ variables, $N$ samples, and $\rho$ nonzero entries in the data matrix, the asymptotic scaling for CorEx goes from $O(Nn)$ to $O(n) + O(N) + O(\rho)$ exploiting sparsity. Latent tree modeling approaches are quadratic in $n$ or worse, so we expect CorEx's computational advantage to increase for larger datasets.

## 4 Data and Evaluation Methods

### 4.1 Data

Our first data set consists of 504,000 humanitarian assistance and disaster relief (HA/DR) articles collected from ReliefWeb, an HA/DR news article aggregator sponsored by the United Nations. Of these articles, about 111,000 of them are in English and contain a label indicating at least one of 21 disaster types, such as Flood, Earthquake, or Wild Fire. To mitigate overwhelming label imbalances, we both restrict the documents to those with one label, and randomly subsample 2000 articles from each of the largest disaster type labels. This leaves us with a corpus of 18,943 articles.

These articles are accompanied by an HA/DR lexicon of approximately 34,000 words and phrases. The lexicon was curated by first gathering seed terms from HA/DR domain experts and CrisisLex, resulting in approximately 40-60 terms per disaster type. This term list was then expanded through the use of several word2vec models per each set of seeds words, and then filtered by removing names, places, non-ASCII characters, terms with fewer than three characters, and words deemed too "semantically distant" from the seeds words by the word2vec models. Finally, the extracted terms were audited using CrowdFlower, where users rated the relevance of the terms on a Likert scale. Low relevance terms were dropped from the lexicon. Of these terms 11,891 appear in the HA/DR articles.

Our second set of data consists of deidentified clinical discharge summaries from the Informatics for Integrating Biology and the Bedside (i2b2) 2008 Obesity Challenge. These summaries are labeled by clinical experts with conditions frequently associated with obesity, such as Coronary Artery Disease, Depression, and Obstructive Sleep Apnea. For these documents, we leverage a text pipeline that extracts common medical terms and phrases (Dai et al., 2008; Chapman et al., 2001). There are 4,114 such terms that appear in the i2b2 clinical health notes. For both sets of data, we use their respective lexicons to parse the documents.

### 4.2 Evaluation

It is well-known that traditional methods for evaluating topic models, such as perplexity and held-out log-likelihood do not necessarily correlate with human evaluation of semantic topic quality (Chang et al., 2009). Therefore, we measure the semantic quality of the topic models using Mimno et. al's UMass automatic topic coherence score (2011). This measure has been shown to correlate well with human evaluation of topic coherence. Suppose there are $n$ topics, and that the $k$ most probable words of topic $t$ are given by the list $(w_1^t, \ldots, w_k^t)$. Then the coherence of topic $t$ is given by

$$\sum_{i=2}^{k} \sum_{j=1}^{i} \log \frac{D(w_i^t, w_j^t) + 1}{D(w_j^t)} \tag{10}$$

where $D(w_i^t)$ is the number of documents in which word $w_i$ appears, and $D(w_i^t, w_j^t)$ is the number of documents in which $w_i$ and $w_j$ appear together.

Second, in the case of the disaster relief documents, we make use of the HA/DR lexicon word labels to report the purity of the topic word lists, the highest fractional count of the word labels. For example, given a topic list with $k$ words, the purity of a list with words all of the same label is 1, while that of a list with words all different labels is $1/k$. Since the HA/DR lexicon labels are the result of expert knowledge and crowd-sourcing, the purity provides us with a measure of semantic topic consistency similar to word intrusion tests (Chang et al., 2009; Lau et al., 2014).

Finally, we evaluate the models in terms of document classification, where the feature set of each document is its topic distribution. The classification is carried out using multiclass logistic regression as implemented by the Scikit-Learn library (Pedregosa et al., 2011), where one binary regression is trained for each label and the label with the highest probability of appearing is selected. While more sophisticated machine learning algorithms may produce better predictive scores, their complex frameworks have the potential to obfuscate differences between topic models. We also leverage the interpretability of logistic regression in our analysis of anchored CorEx. We perform all document classification tasks using a 60/40 split for training and testing.

### 4.3 Choosing Anchor Words

In analyzing anchored CorEx, we wish to systematically test the effect of anchor words given the domain-specific lexicons. To do so, we follow the approach used by Jagarlamudi et. al: for each label in a data set, we find the words that have the highest mutual information, or information gain, with the label (2012). For word $w$ and label $L$, this is computed as

$$I(L : w) = H(L) - H(L \mid w), \tag{11}$$

where for each document of label $L$ we consider if the word $w$ appears or not.

## 5 Results

### 5.1 LDA Baseline Comparison

CorEx takes binarized documents as input for its topic model, so we compare it to LDA giving LDA two different inputs: binarized document-word counts and standard document-word counts. In doing these comparisons, we use the Gensim implementation of LDA (Řehůřek and Sojka, 2010). The results of comparing CorEx to LDA as a function of the number of topics are presented in Figure 1.

On the disaster relief articles, we see that CorEx is competitive with LDA in terms of document classification, and even outperforms LDA in terms of document classification on the clinical health notes. This is despite the fact that CorEx leverages only binary word counts, and LDA uses regular count data. So, with less information than LDA, CorEx produces topics that are as good as or better than the topics produced by LDA when used for document classification.

Inspecting the last two rows of Figure 1, we find that LDA performs better than CorEx in terms of topic coherence, while CorEx performs better than LDA in terms of topic purity. While this appears to yield seemingly conflicting information about the semantic quality of these topic models, it is important to acknowledge that the UMass topic coherence measures assumes that the topic words are the most

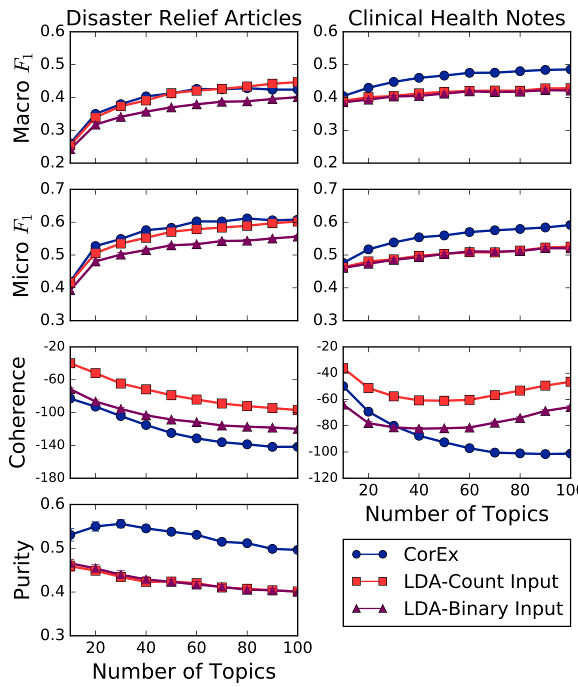


Figure 1: Baseline comparison of CorEx to LDA with respect to document classification and topic quality on disaster relief articles and clinical health notes as the number of topics vary. Points are the average of 30 runs of a topic model. Confidence intervals are plotted but are so small that they are not distinguishable. CorEx uses binarized documents, so we compare CorEx to LDA with binarized input and standard count input.

probable words per each topic. CorEx does not output the most probable words, but rather the words of highest mutual information with the topic. This provides a possible explanation for why CorEx does not perform as well as LDA in terms of coherence, but significantly outperforms in terms of purity. Although topic coherence correlates well with human evaluation of semantic quality, it appears important to apply the measure only *within* models and not *across* models if the topic words are ordered according to different criteria.

### 5.2 Anchored CorEx

To discern the effects of anchoring words to CorEx and simulate domain knowledge injection, we devise the following experiment: first, we determine the top five anchor words for each document label using the methodology described in Section 4.3.

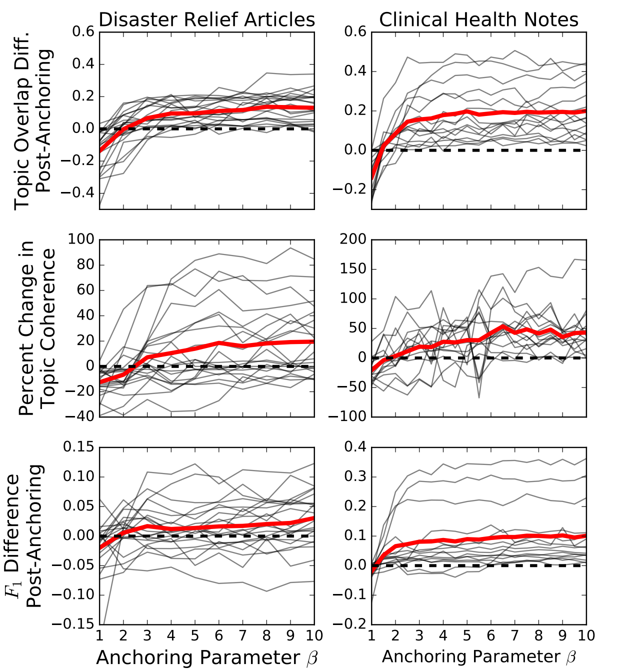


Figure 2: Effect of anchoring words to a single topic for one document label at a time as a function of the anchoring parameter $\beta$. Light gray lines indicate the trajectory of the metric for a given disaster or disease label. Thick red lines indicate the pointwise average across all labels for fixed value of $\beta$.

Second, for each document label, we run an anchored CorEx topic model with that label's anchor words anchored to exactly one topic. We compare this anchored topic model to an unsupervised CorEx topic model using the same random seeds, thus creating a matched pair where the only difference is the treatment of anchor words. Finally, this matched pairs process is repeated 30 times, yielding a distribution for each metric over each label.

We use 50 topics when modeling the ReliefWeb articles and 30 topics when modeling the i2b2 clinical health notes. These values were chosen by observing diminishing returns to the total correlation explained by additional topics. In Figure 2 we show how the results of this experiment vary as a function of the anchoring parameter $\beta$ for each disaster and disease type in the two data sets. We examine a more detailed cross section of these results in Fig 3, where we set $\beta = 5$ for the clinical health notes and set $\beta = 10$ for the disaster relief articles.

A priori we do not know that anchoring will cause the anchor words to appear at the top of topics. So, we first measure how the topic overlap, the proportion of the top ten mutual information words that appear within the top ten words of the topics, changes before and after anchoring. From Figure 2 we see that as $\beta$ increases, more of these relevant words consistently appear within the topics. For the disaster relief articles, many disaster types see about two more words introduced, while in the clinical health notes the overlap increases by up to four words. Analyzing the cross section in Figure 3, we see many of these gains come from disaster and disease types that appeared less in the topics pre-anchoring. Thus, we can sway the topic model towards less dominant themes through anchoring. Document labels that were already well represented are those where the topic overlap changes the least.

Next, we examine whether these anchored topics are more coherent topics. To do so, we compare the coherence of the anchored topic with that of the most predictive topic pre-anchoring, the topic with the largest corresponding coefficient in magnitude of the logistic regression, when the anchored topic itself is most predictive. From Figure 2, we see these results have more variance, but largely the anchored topics are more coherent. In some cases, the coherence is 1.5 to 2 times that of pre-anchoring. Furthermore, by Figure 3, we find that the anchored topics are, indeed, often the most predictive topics for each document label. Similar to topic overlap, the labels that see the least improvement are those that appear the most and are already well-represented in the topic model.

Finally, we find that the anchored, more coherent topics can lead to modest gains in document classification. For the disaster relief articles, Figure 2 shows that there are mixed results in terms of $F_1$ score improvement, with some disaster types performing consistently better, and others performing consistently worse. The results are more consistent for the clinical health notes, where there is an average increase of about 0.1 in the $F_1$ score, and some disease types see an increase of up to 0.3 in $F_1$. Given that we are only anchoring 5 words to the topic model, these are significant gains in predictive power.

Unlike the gains in topic overlap and coherence, the $F_1$ score increases do not simply correlate with

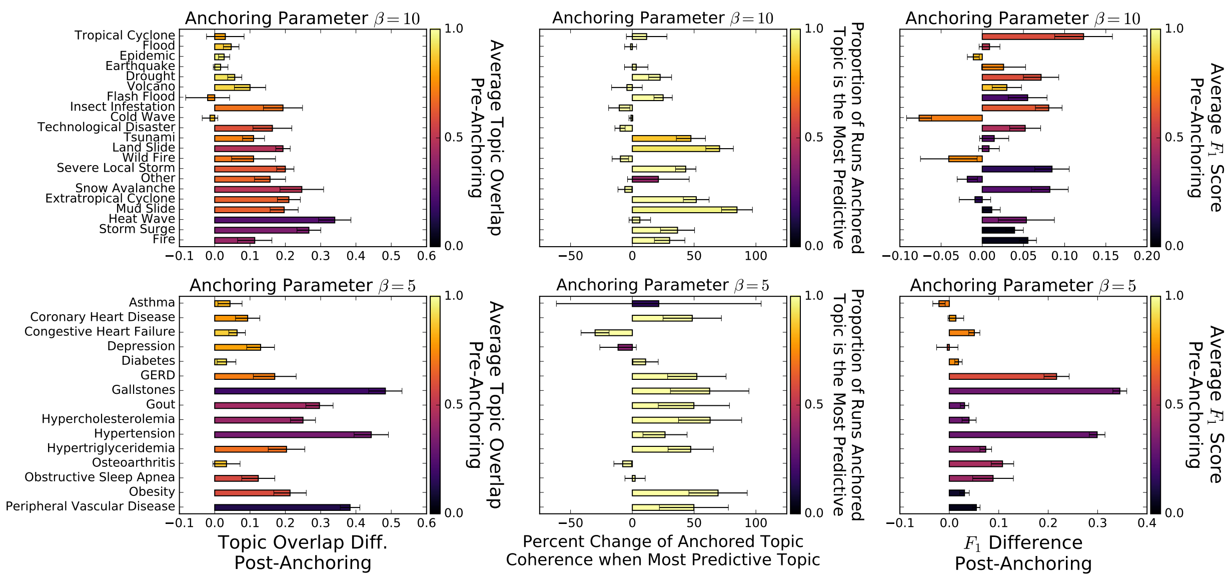


Figure 3: Cross-section results of the anchoring metrics from fixing $\beta = 5$ for the clinical health notes, and $\beta = 10$ for the disaster relief articles. Disaster and disease types are sorted by frequency, with the most frequent document labels appearing at the top. Error bars indicate 95% confidence intervals. The color bars provide baselines for each metric: topic overlap pre-anchoring, proportion of topic model runs where the anchored topic was the most predictive topic, and $F_1$ score pre-anchoring.

which document labels appeared most frequently. For example, we see in Figure 3 that Tropical Cyclone exhibits the largest increase in predictive performance, even though it is also one of the most frequently appearing document labels. Similarly, some of the major gains in $F_1$ for the disease types, and major losses in $F_1$ for the disaster types, do not come from the most or least frequent document labels. Thus, if using anchored CorEx for document classification, it is important to examine how the anchoring affects prediction for individual document labels.

We hypothesize that the results of topic overlap, topic coherence, and $F_1$ score are more muted and have higher variance on the disaster relief articles because there is higher lexical overlap between disaster types than the disease types in the clinical health notes. For example, documents discussing Floods and Flash Foods share many common themes, as do documents discussing Landslides and Mudslides. So again, we emphasize that in applying anchored CorEx, the user should pay attention to how the topics change with the introduction of anchoring, and that the user should experiment with different values of the anchoring parameter $\beta$ to see how these topics are affected.

## 6 Discussion

In this paper, we have introduced an information-theoretic topic model, CorEx, that does not rely on any of the generative assumptions of LDA-based topic models. CorEx is competitive with LDA in terms of producing semantically coherent topics that aid document classification. We also derived a flexible method for anchoring word-level domain knowledge in the CorEx topic model through the information bottleneck. Anchored CorEx guides the topic model towards themes that do not naturally emerge, and often produces more coherent and predictive topics.

Anchored CorEx is more flexible than previous attempts at integrating word-level information into topic models, allowing multiple anchor words per topic, multiple topics per anchor word, and/or a mixture of anchored and unsupervised topics. Our primary goal in this paper was to demonstrate that this anchoring could sway the topic model towards specified, underrepresented topics, and so we largely explored the effect of anchoring words to encourage a single topic. However, the flexibility of anchoring words through the information bottleneck lends itself to many possible creative anchoring strategies that could guide the topic model in different ways. Different goals may call for different anchoring strategies, and future work will explore the effect of alternate strategies.

While we have demonstrated several advantages of the CorEx topic model to LDA, it does have some shortcomings. Most notably, CorEx relies on binary count data, rather than the standard count data that is used as input into LDA and other topic models. Our sparse implementation also requires that each word appears in only one topic. These are not fundamental limitations of the theory, but a matter of computational efficiency. In future work, we hope to remove these restrictions while preserving the speed of the sparse CorEx topic modeling algorithm. As we have demonstrated, this information-theoretic approach has rich potential for finding structure in documents in a new way, and helping domain experts guide topic models with minimal intervention to capture otherwise eclipsed themes.

## Acknowledgments

We would like to thank the Machine Learning and Data Science group at the Information Sciences Institute for their help and insight during the course of this research. RJG was a visiting research assistant at the Information Sciences Institute while performing this research. RJG and GVS were supported by DARPA award HR0011-15-C-0115 and DK was supported by the Alfred E. Mann Innovation in Engineering Doctoral Fellowship.

## References

David Andrzejewski and Xiaojin Zhu. 2009. Latent dirichlet allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 43–48. Association for Computational Linguistics.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32. ACM.

Sanjeev Arora, Rong Ge, and Ankur Moitra. 2012. Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 1–10. IEEE.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David M Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *ICML (2)*, pages 280–288.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Peixian Chen, Nevin L Zhang, Leonard KM Poon, and Zhourong Chen. 2015. Progressive em for latent tree models and hierarchical topic detection. *arXiv preprint arXiv:1508.00973*.

Emily M Cody, Jennie C Stephens, James P Bagrow, Peter Sheridan Dodds, and Christopher M Danforth. 2016. Transitions in climate and energy discourse between hurricanes katrina and sandy. *Journal of Environmental Studies and Sciences*, pages 1–15.

Manhong Dai, Nigam H Shah, Wei Xuan, Mark A Musen, Stanley J Watson, Brian D Athey, Fan Meng, et al. 2008. An efficient solution for mapping free text to ontology terms. *AMIA Summit on Translational Bioinformatics*, 21.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Chris Ding, Tao Li, and Wei Peng. 2008. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis*, 52(8):3913–3927.

Nir Friedman, Ori Mosenzon, Noam Slonim, and Naftali Tishby. 2001. Multivariate information bottleneck. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 152–161. Morgan Kaufmann Publishers Inc.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

Yoni Halpern, Youngduck Choi, Steven Horng, and David Sontag. 2014. Using anchors to estimate clinical state without labeled data. In *AMIA Annual Symposium Proceedings*, volume 2014, page 606. American Medical Informatics Association.

Yoni Halpern, Steven Horng, and David Sontag. 2015. Anchored discrete factor analysis. *arXiv preprint arXiv:1511.03299*.

Nathan Hodas, Greg Ver Steeg, Joshua Harrison, Satish Chikkagoudar, Eric Bell, and Courtney Corley. 2015. Disentangling the lexicons of disaster response in twitter. In *The 3rd International Workshop on Social Web for Disaster Management (SWDM'15)*.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, pages 530–539.

Moontae Lee and David Mimno. 2014. Low-dimensional embeddings for interpretable anchor-based topic inference. In *Proceedings of Empirical Methods in Natural Language Processing*. Citeseer.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics.

Thang Nguyen, Yuening Hu, and Jordan L Boyd-Graber. 2014. Anchors regularized: Adding robustness and extensibility to scalable topic-modeling algorithms. In *ACL (1)*, pages 359–369.

Thang Nguyen, Jordan Boyd-Graber, Jeffrey Lund, Kevin Seppi, and Eric Ringger. 2015. Is your anchor going up or down? fast and accurate supervised topic models. In *North American Chapter of the Association for Computational Linguistics*. Citeseer.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA, May.

Kyle Reing, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2016. Toward interpretable topic discovery via anchored correlation explanation. *arXiv:1606.07043 (ICML Workshop on Human Interpretability in Machine Learning)*.

Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315. ACM.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Greg Ver Steeg and Aram Galstyan. 2014. Discovering structure in high-dimensional data through correlation explanation. In *Advances in Neural Information Processing Systems*, pages 577–585.

Greg Ver Steeg and Aram Galstyan. 2015. Maximally informative hierarchical representations of high-dimensional data. *in depth*, 13:14.

## A Supplemental Material: Anchor Words and Topic Examples

| Disease Type | Anchor Words |
|---|---|
| Asthma | asthma, albuterol, wheeze, advair, fluticasone |
| Coronary Artery Disease | coronary artery disease, aspirin, myocardial inarction, plavix |
| Congestive Heart Failure | congestive heart failure, lasix, diuresis, heart failure, cardiomyopathy |
| Depression | depression, prozac, celexa, seroquel, remeron |
| Diabetes | diabetes mellitus, diabetes, nph insulin, insulin, metformin |
| Gastroesophageal Reflux Disease | gastroesophageal refulx, no known drug allergy, protonix, not:, reflux |
| Gallstones | gallstone, cholecystitis, cholelithiasis, abdominal pain,vomiting |
| Gout | gout, allopurinol, colchicine, renal insufficiency, torsemide |
| Hypercholesterolemia | hypercholesterolemia, hyperlipidemia, aspirin, lipitor, dyslipidemia |
| Hypertension | hypertension, lisinopril, aspirin, diabetes mellitus, atorvastatin |
| Hypertriglyceridemia | hypertriglyceridemia, gemfibrozil, citrate, orphenadrine, hydroxymethylglutaryl coa reductase inhibitors |
| Osteoarthritis | osteoarthritis, degenerative joint disease, arthritis, naproxen, fibromyalgia |
| Obstructive Sleep Apnea | sleep apnea, obstructive sleep apnea, morbid obese, obesity, ipratropium |
| Obesity | obesity, morbid obesity, obese, sleep apnea, coronary artery disease |
| Peripheral Vascular Disease | cellulitis, erythema, ulcer, swelling, word finding difficulty |

Table A1: Words that have the highest mutual information with each disease type.

| Disaster Type | Anchor Words |
|---|---|
| Cold Wave | winter, snow, cold, temperatures, heavy snow |
| Drought | drought, taliban, wheat, refugees, severe drought |
| Earthquake | earthquake, quake, richter scale, tents, injured |
| Epidemic | virus, ebola outbreak, transmission, ebola virus, disaster |
| Extratropical Cyclone | typhoon, storm, farmland, houses, storm coincided |
| Fire | fire, hospitals, blaze, water crisis, firefighters |
| Flash Flood | flood, floods, flash floods, monitoring stations, muhuri |
| Flood | floods, flood, flooding, flood victims, rains |
| Heat Waves | heat, temperatures, heat wave, heatstroke, sunstroke |
| Insect Infestation | locust, food crisis, infestations, millet, harvest |
| Land Slide | landslides, houses, hunza river, search, village |
| Mud Slide | mudslides, rains, mudslide, torrential rains, houses |
| Other | climate, ocean, drought, impacts, warming |
| Severe Local Storm | tornado, storm, tornadoes, houses, storms |
| Snow Avalanche | avalanches, avalanche, snow, snowfall, an avalanche |
| Storm Surge | king tides, tropical storm, ocean, cyclone season, flooded |
| Technological Disaster | environmental, toxic waste, pollution, tanker, sludge |
| Tropical Cyclone | hurricane, cyclone, storm, tropical storm, national hurricane |
| Tsunami | earthquake, disaster, tsunamis, wave, rains |
| Volcano | eruption, lava, volcanic, crater, eruptions |
| Wild Fire | fires, fire, forest fires, firefighters, burning |

Table A2: Words that have the highest mutual information with each disaster type.


Figure A1: Hierarchical CorEx topic model of the disaster relief articles. Edge widths are proportional to the mutual information with the latent representation.

| Rank | Topic |
|---|---|
| 1 | drought, farmers, harvest, crop, livestock, planting, grain, maize, rainfall, irrigation |
| 2 | floods, flooding, flood, rains, flooded, landslides, inundated, rivers, submerged, flash floods |
| 3 | eruption, volcanic, lava, crater, eruptions, volcanos, slopes, volcanic activity, evacuated, lava flows |
| 4 | storm, winds, coast, hurricane, weather, tropical storm, national hurricane, coastal, storms, meteorological |
| 5 | virus, ebola outbreak, transmission, health workers, vaccination, ebola virus, suspected cases, fluids, ebola virus disease, ebola patients |
| 6 | malnutrition, refugees, food aid, nutrition, feeding, refugees in, hunger, nutritional, refugee, food crisis |
| 7 | international federation, red cross, red crescent, societies, volunteers, disaster relief emergency, national societies, disaster preparedness, information bulletin, relief operation |
| 8 | winter, snow, snowfall, temperatures, heavy snow, heating, freezing, warm clothing, severe winter, avalanches |
| 9 | support, assistance, appeal, funds, assist, contributions, fund, cash, contribution, organizations |
| 10 | taliban, repatriation, elections, militia, convoy, ruling, talibans, islamic, convoys, vote |
| 11 | ngos, donors, humanitarian, un agencies, mission, funding, unicef, conduct, humanitarian assistance, inter-agency |
| 12 | fires, fire, forest fires, burning, firefighters, wildfires, blaze, flames, fire fighting, forests |
| 13 | earthquake, quake, richter scale, aftershocks, earthquakes, magnitude earthquake, magnitude, devastating earthquake, an earthquake, earthquake struck |
| 14 | blankets, tents, families, clothing, utensils, plastic sheeting, clothes, tarpaulins, schools, shelters |
| 15 | rescue, search, injured, helicopters, death toll, rescue operations, rescue teams, police, rescuers, stranded |
| 16 | crops, cereal, cereals, millet, food shortages, sorghum, harvests, shortage, ration, rainy |
| 17 | medical, patients, hospital, hospitals, nurses, clinics, clinic, doctor, medical team, beds |
| 18 | water, water supply, drinking water, pumps, drinking, water supplies, potable water, water distribution, installed, constructed |
| 19 | locust, attacks, fighting, infestations, pesticides, opposition, attack, reform, dialogue, governance |
| 20 | environmental, pollution, contamination, fish, impacts, water quality, polluted, pollutants, chemicals, tanker |
| 21 | malaria, diarrhoea, diseases, oral, rehydration, salts, contaminated, epidemics, borne diseases, respiratory infections, clean |
| 22 | emergency, emergencies, ocha, disaster response, coordinating, emergency response, coordinated, coordinators, transportation, rapid assessment |
| 23 | military, armed, civilians, soldiers, aircraft, weapons, rebel, planes, bombs, military personnel |
| 24 | united nations, humanitarian affairs, agencies, agency, governmental, united nations childrens fund, relief coordinator, general assembly, international cooperation, donor community |
| 25 | transport, flights, trucks, airport, transported, flight, truck, airlift, cargo, route |

Table A3: Topics 1–25 resulting from the best of 10 CorEx topic models run on the disaster relief articles. Topics are ranked by total correlation explained.

| Rank | Topic |
|---|---|
| 26 | basin, monitoring stations, basins, muhuri, flood forecasting, significant rainfall, moderate rainfall, upstream, light, sludge |
| 27 | criminal, detained, parliament, protest, crime, protests, protesters, suspects, firing, incident |
| 28 | public health, organization, ministry of, efforts, outbreaks, building, leaders, civil society, minister of, facility |
| 29 | housing, reconstruction, construction, repair, rebuilding, repairs, temporary housing, corrugated, permanent housing, debris removal |
| 30 | houses, killed, village, were killed, buildings, swept, debris, roofs, roof, collapse |
| 31 | training, partners, protection, interventions, delivery establishment, violence, benefit, unfpa, pilt |
| 32 | sanitation, provision, safe, drinking water, latrine, hygiene education, implementing partners, diarrhoeal diseases, rehabilitated, dispaced persons, sanitation services |
| 33 | flour, wheat, sugar, vegetable, beans, rations, food rations, bread, lentils, needy |
| 34 | camps, living, army, troops, resettlement, relocated, relocation, relocate, flee, settlement |
| 35 | disaster, disasters, disaster relief, cyclone, coordinating council, cyclones, aftermath, devastation, devastated, natural disaster |
| 36 | relief, relief supplies, relief efforts, relief operations, relief assistance, relief goods, relief materials, relief agencies, donate, providing relief |
| 37 | household, procurement, vulnerable groups, beneficiary, pipeline, rehabilitate, local ngos, iodised salt, rainfed areas, water harvesting |
| 38 | staff, supplies, personnel, deployed, staff members, airlifted specialists, flown, logistical support, airlifting |
| 39 | facilities, soap, medical supplies, clean water, sanitation facilities, emergency medical, international organization, psychosocial, tent, migration iom |
| 40 | fuel, supply, diesel energy, nitrate, diesel fuel, orphanages, grid, hydroelectric, storage, facilities |
| 41 | cold, cold weather, wave, warm clothes, extreme temperatures, firewood, severe cold weather, severe cold wave, average temperature |
| 42 | cholera outbreak, cholera epidemic, poor sanitation, cholera outbreaks, wash, poor hygiene, dirty water, disinfect, hygiene awareness, good hygiene practices |
| 43 | government, governments, prime minister, administration, national disaster management, corporation, dollars, bilateral donors, disburse, telecom |
| 44 | famine, severe drought, crises, prolonged drought, devastating, mortality rate, degradation, catastrophic, famine relief, agricultural practices |
| 45 | vegetation, ecological, threat, mosquitoes, insect, insecticides, lakes, prolonged, habitation, adverse weather |
| 46 | latrines, water tanks, water containers, affected communities, chlorine tablets, household kits, solid waste, reception centre, local organisations, piped water |
| 47 | survivors, relief effort, relief workers, survivor, clean drinking water, outlying areas, devastating cyclone, cyclone struck, cyclone survivors, medic |
| 48 | perished, water storage, caused extensive damage, soil erosion, total loss, sewage systems, salt water, soup, water purifying tablets, electric power |
| 49 | canal, disruption, rehabilitating, infrastructures, vulnerable areas, uninterrupted, power plants, stagnant, inaccessible areas, distress |
| 50 | voluntary, basic needs, rehabilitation phase, blankets mattresses, raised, freight, humanitarian organizations, government agency, delta region, persons displaced |

Table A4: Topics 26–50 resulting from the best of 10 CorEx topic models run on the disaster relief articles. Topics are ranked by total correlation explained.

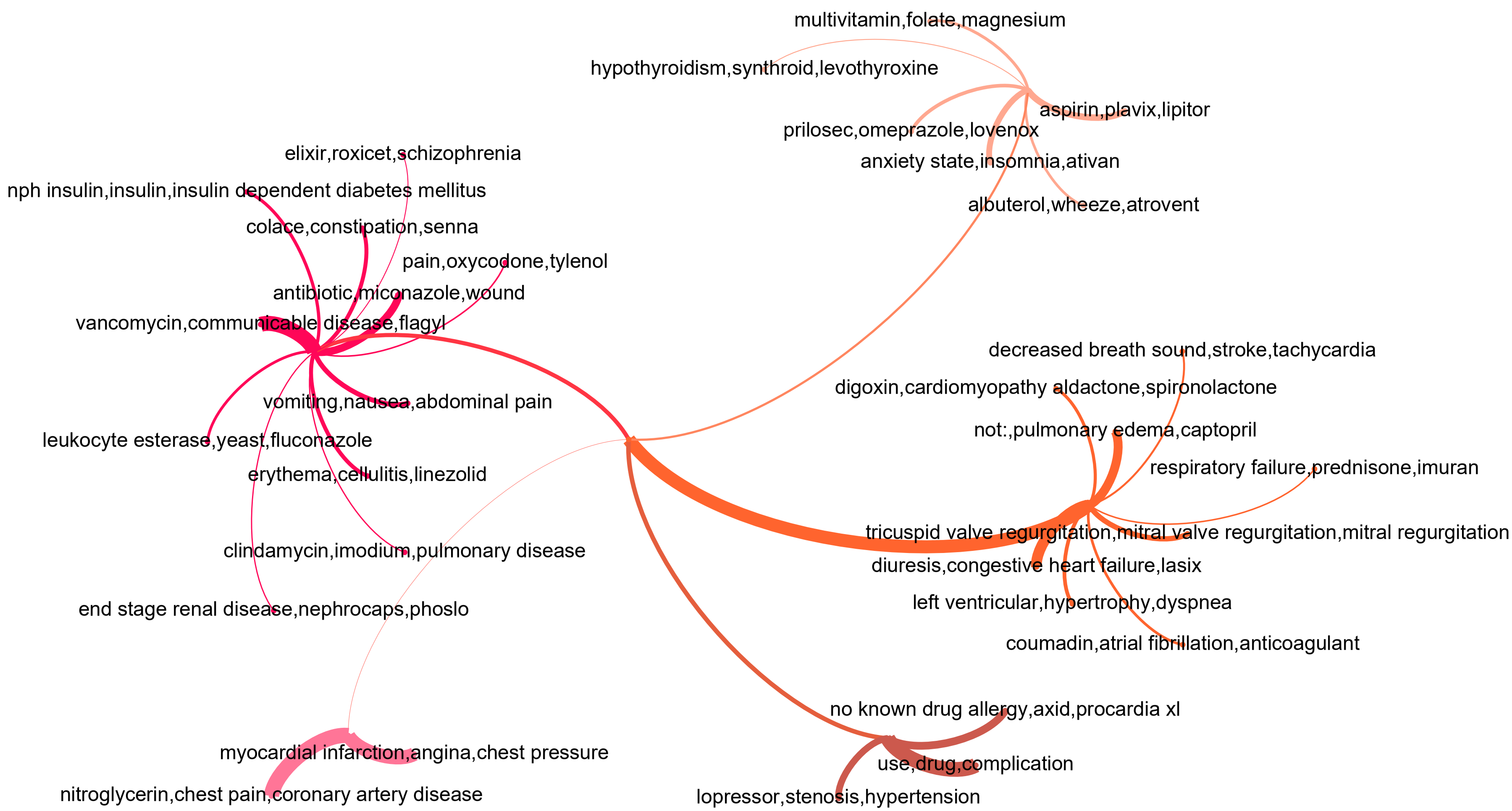


Figure A2: Hierarchical CorEx topic model of the clinical health notes. Edge widths are proportional to the mutual information with the latent representation.

| Rank | Topic |
|---|---|
| 1 | use, drug, complication, allergy, sodium, infection, furosemide, docusate, shortness of breath, potassium chloride |
| 2 | vancomycin, communicable disease, flagyl, levofloxacin, diabetes, renal failure, sepsis, ceftazidime, nutrition, gentamicin |
| 3 | aspirin, plavix, lipitor, toprol xl, lantus, hydroxymethylglutaryl coa reductase inhibitors, atorvastatin, nexium, novolog, disease |
| 4 | diuresis, congestive heart failure, lasix, edema, orthopnea, crackle, heart failure, dyspnea on exertion, oxygen, torsemide |
| 5 | albuterol, wheeze, atrovent, chronic obstructive pulmonary disease, asthma, flovent, ipratropium, fluticasone, advair, combivent |
| 6 | end stage renal disease, nephrocaps, phoslo, calcitriol, cellcept, kidney transplant, arteriovenous fistula, acetate, cyclosporine, neoral |
| 7 | nitroglycerin, chest pain, coronary artery disease, hypokinesia, st depression, lesion, unstable angina, akinesia, st elevation, diaphoresis |
| 8 | respiratory failure, prednisone, imuran, immunosuppression, necrosis, cyclosporin, sick, magnesium oxide, tachypnea, arteriovenous malformation |
| 9 | elixir, roxicet, schizophrenia, risperdal, zofran, crushed, valproic acid, promethazine, phenergan, prochlorperazine |
| 10 | leukocyte esterase, yeast, fluconazole, urosepsis, dysphagia, oxycontin, lidoderm, chemotherapy, adriamycin, medical problems |
| 11 | colace, constipation, senna, lactulose, dulcolax, milk of magnesia, sennoside, dilaudid, protonix, reglan |
| 12 | vomiting, nausea, abdominal pain, diarrhea, fever, dehydration, chill, clostridium difficile, intravenous fluid, compazine |
| 13 | coumadin, atrial fibrillation, anticoagulant, warfarin, k vitamin, amiodarone, atrial flutter, flutter, deep venous thrombosis, allopurinol |
| 14 | digoxin, cardiomyopathy, aldactone, spironolactone, carvedilol, dobutamine, alcohol, idiopathic cardiomyopathy, ventricular rate, addiction |
| 15 | clindamycin, imodium, pulmonary disease, erythromycin, defervesced, sweating, carafate, quinidine, cytomegalovirus, cepacol |
| 16 | lopressor, stenosis, hypertension, heparin, hypercholesterolemia, aortic valve insufficiency, mitral valve insufficiency, aortic valve stenosis, sinus rhythm, peripheral vascular disease |
| 17 | antibiotic, miconazole, wound, nitrate, morbid obese, fentanyl, sleep apnea, obesity, abscess, ampicillin |
| 18 | erythema, cellulitis, linezolid, swelling, erythematous, osteomyelitis, ancef, keflex, dicloxacillin, bacitracin |
| 19 | anxiety state, insomnia, ativan, neurontin, depression, lorazepam, gabapentin, trazodone, fluoxetine, headache |
| 20 | multivitamin, folate, magnesium, folic acid, mvi, maalox, thiamine, vitamin c, gluconate, dyspepsia |

Table A5: Topics 1–20 resulting from the best of 10 CorEx topic models run on the clinical health notes. Topics are ranked by total correlation explained.

| Rank | Topic |
|---|---|
| 21 | decreased breath sound, stroke, tachycardia, seizure disorder, lymphocyte, atelectasis, polymorphonuclear leukocytes, ecchymosis, seizure, cefotaxime |
| 22 | not: , pulmonary edema, captopril, pleural effusion, rales, beta blocker, fatigue, dead, q wave, dysfunction |
| 23 | hypothyroidism, synthroid, levothyroxine, levoxyl, diovan, valsartan, angioedema, bestrophinopathy, atherosclerosis, ursodiol |
| 24 | nph insulin, insulin, insulin dependent diabetes mellitus, anemia, humulin insulin, retinopathy, hyperglycemia, humulin, gastrointestinal bleeding, nephropathy |
| 25 | tricuspid valve regurgitation, mitral valve regurgitation, mitral regurgitation, left atrial enlargement, zaroxolyn, ectopy, right atrial enlargement, metolazone, deficit, regurgitant |
| 26 | prilosec, omeprazole, lovenox, pulmonary embolism, enoxaparin, xalatan, oxybutynin, helicopter pylori, flonase, ramipril |
| 27 | pain, oxycodone, tylenol, percocet, ibuprofen, morphine, osteoarthritis, hernia, motrin, bleeding |
| 28 | left ventricular hypertrophy, dyspnea, living alone, smokes, syndrome, hives, palpitation, elderly, left axis deviation, usual state of health |
| 29 | myocardial infarction, angina, chest pressure, patent ductus arteriosus, atenolol, micronase, adenosine, non-insulin dependent diabetes mellitus, ecotrin, caltrate |
| 30 | no known drug allergy, axid, procardia xl, vasotec, obese, mevacor, tissue plasminogen activator, middle-aged, nifedipine, procardia |

Table A6: Topics 21–30 resulting from the best of 10 CorEx topic models run on the clinical health notes. Topics are ranked by total correlation explained.