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
[2]Vermont Complex Systems Center, Computational Story Lab, University of Vermont
ryan.gallagher@uvm.edu
{reing,kale,gregv}@isi.edu

Abstract

While generative models such as Latent Dirichlet Allocation (LDA) have proven fruitful in topic modeling, they often require detailed assumptions and careful specification of hyperparameters. Such model complexity issues only compound when trying to generalize generative models to incorporate human input. We introduce Correlation Explanation (CorEx), an alternative approach to topic modeling that does not assume an underlying generative model, and instead learns maximally informative topics through an information-theoretic framework. This framework naturally generalizes to hierarchical and semi-supervised extensions with no additional modeling assumptions. In particular, word-level domain knowledge can be flexibly incorporated within CorEx through anchor words, allowing topic separability and representation to be promoted with minimal human intervention. Across a variety of datasets, metrics, and experiments, we demonstrate that CorEx produces topics that are comparable in quality to those produced by unsupervised and semi-supervised variants of LDA.

1 Introduction

The majority of topic modeling approaches utilize probabilistic generative models, models which specify mechanisms for how documents are written in order to infer latent topics. These mechanisms may be explicitly stated, as in Latent Dirichlet Allocation (LDA) (Blei et al., 2003), or implicitly stated, as with matrix factorization techniques (Hofmann, 1999; Ding et al., 2008; Buntine and Jakulin, 2006). The core generative mechanisms of LDA, in particular, have inspired numerous generalizations that account for additional information, such as the authorship (Rosen-Zvi et al., 2004), document labels (McAuliffe and Blei, 2008), or hierarchical structure (Griffiths et al., 2004).

However, these generalizations come at the cost of increasingly elaborate and unwieldy generative assumptions. While these assumptions allow topic inference to be tractable in the face of additional metadata, they progressively constrain topics to a narrower view of what a topic can be. Such assumptions are undesirable in contexts where one wishes to minimize model complexity and learn topics without preexisting notions of how those topics originated.

For these reasons, we propose topic modeling by way of Correlation Explanation (CorEx),[1] an information-theoretic approach to learning latent topics over documents. Unlike LDA, CorEx does not assume a particular data generating model, and instead searches for topics that are "maximally informative" about a set of documents. By learning informative topics rather than generated topics, we avoid specifying the structure and nature of topics ahead of time.

In addition, the lightweight framework underlying CorEx is versatile and naturally extends to hierarchical and semi-supervised variants with no additional modeling assumptions. More specifically, we

[1]Open source, documented code for the CorEx topic model available at https://github.com/gregversteeg/corex_topic.

Transactions of the Association for Computational Linguistics, vol. 5, pp. 529–542, 2017. Action Editor: Diana McCarthy.
Submission batch: 7/2017 Published 12/2017.

may flexibly incorporate word-level domain knowledge within the CorEx topic model. 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 of importance to the user. By incorporating relevant domain words, we might encourage our topic model to recognize a rare disease that would otherwise be missed in clinical health notes, focus more attention to topics from news articles that can guide relief workers in distributing aid more effectively, or disambiguate aspects of a complex social issue.

Our contributions are as follows: first, we frame CorEx as a topic model and derive an efficient alteration to the CorEx algorithm to exploit sparse data, such as word counts in documents, for dramatic speedups. Second, we show how domain knowledge can be naturally integrated into CorEx through "anchor words" and the information bottleneck. Third, we demonstrate that CorEx and anchored CorEx produce topics of comparable quality to unsupervised and semi-supervised variants of LDA over several datasets and metrics. Finally, we carefully detail several anchoring strategies that highlight the versatility of anchored CorEx on a variety of tasks.

2 Methods

2.1 CorEx: Correlation Explanation

Here we review the fundamentals of Correlation Explanation (CorEx), and adopt 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, indicated with lowercase, x. Furthermore, if we have n such random variables, let X_G denote a sub-collection of them, where $G \subseteq \{1, \ldots, n\}$. The probability of observing $X_G = x_G$ is written as $p(X_G = x_G)$, which is typically abbreviated to $p(x_G)$. 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

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

$$= D_{KL}\left(p(x_G) || \prod_{i \in G} p(x_i)\right). \quad (2)$$

We see that Eq. 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, Eq. 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.

$$TC(X_G; Y) = TC(X_G) - TC(X_G \mid Y) \quad (3)$$

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

The quantity expressed in Eq. 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 $TC(X_G \mid Y) = 0$. If this is the case, then $TC(X_G; Y)$ is maximized, and Y explains all of the dependencies in X_G.

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

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

As we describe in the following section, this objective can be efficiently approximated, despite the search occurring over an exponentially large probability space (Ver Steeg and Galstyan, 2014).

Since each topic explains a certain portion of the overall total correlation, we may choose the number of topics by observing diminishing returns to the objective. Furthermore, since the CorEx implementation depends on a random initialization (as described shortly), one may restart the CorEx topic model several times and choose the one that explains the most total correlation.

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.

2.2 CorEx Implementation

We summarize the implementation of CorEx as presented by Ver Steeg and Galstyan (2014) in preparation for innovations introduced in the subsequent sections. The numerical optimization for CorEx begins with a random initialization of parameters and then proceeds via an iterative update scheme similar to EM. For computational tractability, we subject the optimization in Eq. 5 to the constraint that the groups, G_j, do not overlap, i.e. we enforce single-membership of words within topics. The optimization entails a combinatorial search over groups, so instead we look for a form that is more amenable to smooth optimization. We rewrite the objective using the alternate form in Eq. 4 while 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. $i \in G_j$).

$$\max_{\alpha_{i,j}, p(y_j|x)} \quad \sum_{j=1}^{m} \left(\sum_{i=1}^{n} \alpha_{i,j} I(X_i : Y_j) - I(X : Y_j) \right)$$
$$\text{s.t.} \qquad \alpha_{i,j} = \mathbb{I}[j = \arg\max_{\bar{j}} I(X_i : Y_{\bar{j}})]. \tag{6}$$

Note that the constraint on non-overlapping groups now becomes a constraint on α. To make the optimization smooth we should relax the constraint so that $\alpha_{i,j} \in [0, 1]$. To do so, we replace the second line with a softmax function. The update for α at iteration t becomes,

$$\alpha_{i,j}^t = \exp\left(\lambda^t (I(X_i : Y_j) - \max_{\bar{j}} I(X_i : Y_{\bar{j}})) \right).$$

Now $\alpha \in [0, 1]$ and the parameter λ controls the sharpness of the softmax function. Early in the optimization we use a small value of λ, then increase it later in the optimization to enforce a hard constraint. The objective in Eq. 6 only lower bounds total correlation in the hard max limit. The constraint on α forces competition among latent factors to explain certain words, while setting $\lambda = 0$ results in all factors learning the same thing. Holding α fixed, taking the derivative of the objective (with respect to the variables $p(y_j|x)$, and setting it equal to zero leads to a fixed point equation. We use this fixed point to define update equations at iteration t.

$$p_t(y_j) = \sum_{\bar{x}} p_t(y_j|\bar{x}) p(\bar{x}) \tag{7}$$

$$p_t(x_i|y_j) = \sum_{\bar{x}} p_t(y_j|\bar{x}) p(\bar{x}) \mathbb{I}[\bar{x}_i = x_i] / p_t(y_j)$$

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

The first two lines just define the marginals in terms of the optimization parameter, $p_t(y_j|x)$. We take $p(x)$ to be the empirical distribution defined by some observed samples, $x^\ell, \ell = 1, \ldots, N$. The third line updates $p_t(y_j|x^\ell)$, the probabilistic labels for each latent factor, Y_j, for a given sample, x^ℓ. Note that an easily calculated constant, $\mathcal{Z}_j(x^\ell)$, appears to ensure the normalization of $p_t(y_j|x^\ell)$ for each sample. We iterate through these updates until convergence.

After convergence, we use the mutual information terms $I(X_i : Y_j)$ to rank which words are most informative for each factor. The objective is a sum of terms for each latent factor and this allows us to rank the contribution of each factor toward our lower bound on the total correlation. The expected log of the normalization constant, often called the free energy, $\mathbb{E}[\log \mathcal{Z}_j(x)]$, plays an important role since its expectation provides a free estimate of the i-th term in the objective (Ver Steeg and Galstyan, 2015), as

can be seen by taking the expectation of Eq. 8 at convergence and comparing it to Eq. 6. Because our sample estimate of the objective is just the mean of contributions from individual sample points, x^ℓ, we refer to $\log \mathcal{Z}_j(x^\ell)$ as the pointwise total correlation explained by factor j for sample ℓ. Pointwise TC can be used to localize which samples are particularly informative about specific latent factors.

2.3 Sparsity Optimization

2.3.1 Derivation

To alter the CorEx optimization procedure to exploit sparsity in the data, we now assume that all variables, x_i, y_j, are binary and x is a binary vector where $X_i^\ell = 1$ if word i occurs in document ℓ and $X_i^\ell = 0$ otherwise. Since all variables are binary, the marginal distribution, $p(x_i|y_j)$, 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 in Eq. 8 for each document ℓ. 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.

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

Note, when the word does not appear in the document, only the leading term of Eq. 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 assume by default that a word does not appear in the document, and then correct the contribution to the update if the word does appear.

Thus, when substituting back into Eq. 8, the sum becomes a matrix multiplication between a matrix with dimensions of the number of variables by the number of documents and entries x_i^ℓ that is assumed to be sparse and a dense matrix with dimensions of the number of variables by the number of latent factors. Given n variables, N samples, and ρ 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.

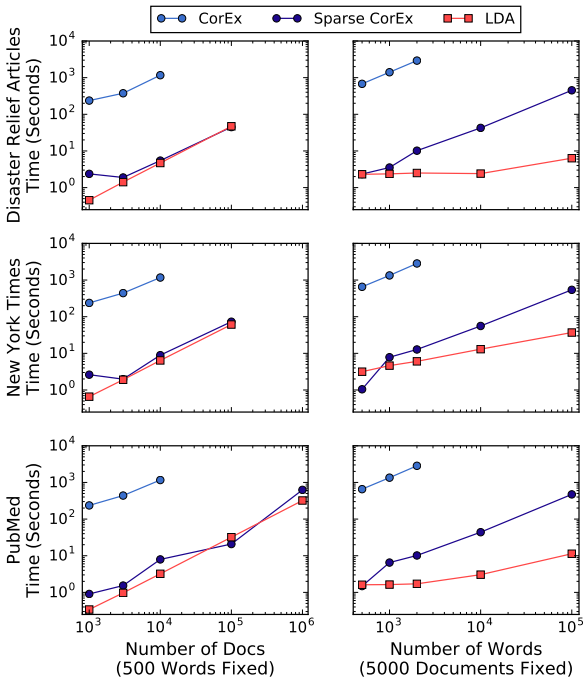

Figure 1: Speed comparisons to a fixed number of iterations as the number of documents and words vary. New York Times articles and PubMed abstracts were collected from the UCI Machine Learning Repository (Lichman, 2013). The disaster relief articles are described in section 4.1, and represented simply as bags of words, not phrases.

2.3.2 Optimization Evaluation

We perform experiments comparing the running time of CorEx before and after implementing the improvements which exploit sparsity. We also compare with Scikit-Learn's simple batch implementation of LDA using the variational Bayes algorithm (Hoffman et al., 2013). Experiments were performed on a four core, Intel i5 chip running at 4 GHz with 32 GB RAM. We show run time when varying the data size in terms of the number of word types and the number of documents. We used 50 topics for all runs and set the number of iterations for each run to 10 iterations for LDA and 50 iterations for CorEx. Results are shown in Figure 1. We see that CorEx exploiting sparsity is orders of magnitude faster than the

naive version and is generally comparable to LDA as the number of documents scales. The slope on the log-log plot suggests a linear dependence of running time on the dataset size, as expected.

2.4 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., 1999; Friedman et al., 2001). More formally, the information bottleneck is expressed as

$$\max_{p(y|x)} \beta I(Z:Y) - I(X:Y), \tag{10}$$

where β is a parameter controlling the trade-off between compressing X and preserving information about the relevance variable, Z.

To see the connection with CorEx, we compare the CorEx objective as written in Eq. 6 with the bottleneck in Eq. 10. 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 $Z \equiv X_i$. If we want to learn representations that are more relevant to specific keywords, we can simply anchor a word type 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 α remain the same. This schema is a natural extension of the CorEx optimization and it is flexible, allowing for multiple word types to be anchored to one topic, for one word type to be anchored to multiple topics, or for any combination of these semi-supervised anchoring strategies.

3 Related Work

With respect to integrating domain knowledge into topic models, we draw inspiration from Arora et al. (2012), who used anchor words in the context of non-negative matrix factorization. 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. (2012), and subsequent improvements to their approach, 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. (2014; 2015), who have also made use of domain expertise and semi-supervised anchored words in devising topic models.

There is an adjacent line of work that has focused on incorporating word-level information into LDA-based models. Jagarlamudi et al. (2012) proposed SeededLDA, a model that seeds words into given topics and guides, but does not force, these topics towards these integrated words. Andrzejewski and Zhu (2009) presented a model that makes use of "z-labels," words that are known to pertain to specific topics and that are restricted to appearing in some subset of all the possible topics. Although the z-labels can be leveraged to place different senses of a word into different topics, it requires additional effort to determine when these different senses occur. Our anchoring approach allows a user to more easily anchor one word to multiple topics, allowing CorEx to naturally find topics that revolve around different senses of a word.

Andrzejewski et al. (2009) presented a second model which allows specification of Must-Link and Cannot-Link relationships between words that help partition otherwise muddled topics. These logical constraints help enforce topic separability, though these mechanisms less directly address how to anchor a single word or set of words to help a topic emerge. More generally, the Must/Cannot link and z-label topic models have been expressed in a powerful first-order-logic framework that allows the specification of arbitrary domain knowledge through logical rules (Andrzejewski et al., 2011). Others have built off this first-order-logic approach to automatically learn rule weights (Mei et al., 2014) and incorporate additional latent variable information (Foulds et al., 2015).

Mathematically, CorEx topic models most closely resemble topic models based on latent tree reconstruction (Chen et al., 2016). In Chen et al.'s (2016) analysis, their own latent tree approach and CorEx both report significantly better perplexity than hi-

erarchical topic models based on the hierarchical Dirichlet process and the Chinese restaurant process. CorEx has also been investigated as a way to find "surprising" documents (Hodas et al., 2015).

4 Data and Evaluation Methods

4.1 Data

We use two challenging datasets with corresponding domain knowledge lexicons to evaluate anchored CorEx. Our first dataset consists of 504,000 humanitarian assistance and disaster relief (HA/DR) articles covering 21 disaster types collected from ReliefWeb, an HA/DR news article aggregator sponsored by the United Nations. To mitigate overwhelming label imbalances during anchoring, we both restrict ourselves to documents in English with one label, and randomly subsample 2,000 articles from each of the largest disaster type labels. This leaves us with a corpus of 18,943 articles.[2]

We accompany these articles with an HA/DR lexicon of approximately 34,000 words and phrases. The lexicon was curated by first gathering 40–60 seed terms per disaster type from HA/DR domain experts and CrisisLex. This term list was then expanded by creating word embeddings for each disaster type, and taking terms within a specified cosine similarity of the seed words. These lists were then filtered by removing names, places, non-ASCII characters, and terms with fewer than three characters. 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 types appear in the HA/DR articles.

Our second dataset consists of 1,237 deidentified clinical discharge summaries from the Informatics for Integrating Biology and the Bedside (i2b2) 2008 Obesity Challenge.[3] These summaries are labeled by clinical experts with 15 conditions frequently associated with obesity. For these documents, we leverage a text pipeline that extracts common medical terms and phrases (Dai et al., 2008; Chapman et al., 2001), which yields 3,231 such term types. For both sets of documents, we use their respective lexicons to break the documents down into bags of words and phrases.

We also make use of the 20 Newsgroups dataset, as provided and preprocessed in the Scikit-Learn library (Pedregosa et al., 2011).

4.2 Evaluation

CorEx does not explicitly attempt to learn a generative model and, thus, traditional measures such as perplexity are not appropriate for model comparison against LDA. Furthermore, it is well-known that 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 topic quality using Mimno et al.'s (2011) UMass automatic topic coherence score, which correlates with human judgments.

We also evaluate the models in terms of multiclass logistic regression document classification (Pedregosa et al., 2011), where the feature set of each document is its topic distribution. We perform all document classification tasks using a 60/40 training-test split.

Finally, we measure how well each topic model does at clustering documents. We obtain a clustering by assigning each document to the topic that occurs with the highest probability. We then measure the quality within clusters (homogeneity) and across clusters (adjusted mutual information). The highest possible value for both measures is one. We do not report clustering metrics on the clinical health notes because the documents are multi-label and, in that case, the metrics are not well-defined.

4.3 Choosing Anchor Words

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. (2012) to automatically generate anchor words: for each label in a data set, we find the words that have the highest mutual information with the label. For word w and label L, this is computed as

$$I(L:w) = H(L) - H(L \mid w), \qquad (11)$$

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

[2] HA/DR articles and accompanying lexicon available at http://dx.doi.org/10.7910/DVN/TGOPRU

[3] Data available upon data use agreement at https://www.i2b2.org/NLP/Obesity/

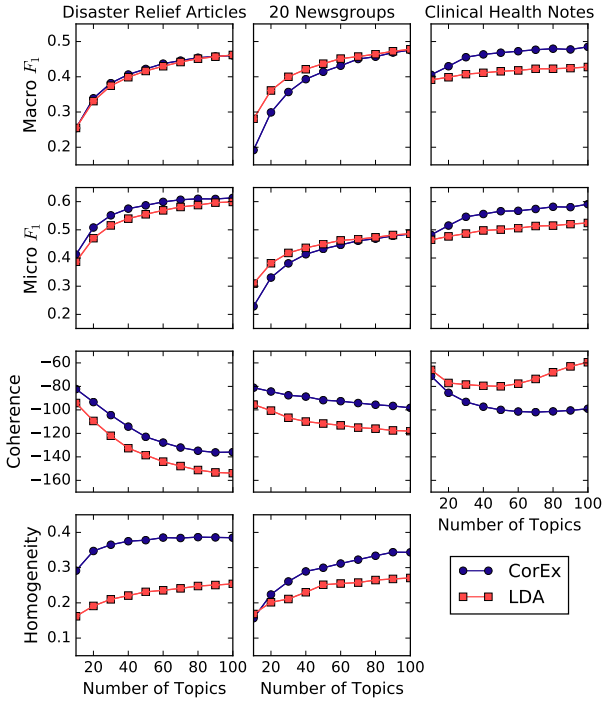

Figure 2: Baseline comparison of CorEx to LDA with respect to topic coherence and document classification and clustering on three different datasets as the number of topics vary. Points are the average of 30 runs of a topic model. Confidence intervals are plotted but are so small that they are not distinguishable. CorEx is trained using binary data, while LDA is trained on count data. Homogeneity is not well-defined on the multi-label clinical health notes, so it is omitted.

5 Results

5.1 LDA Baseline Comparison

We compare CorEx to LDA in terms of topic coherence, document classification, and document clustering across three datasets. CorEx is trained on binary data, while LDA is trained on count data. While not reported here, CorEx consistently outperformed LDA trained on binary data. 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 2.

Across all three datasets, we find that the topics produced by CorEx yield document classification results that are on par with or better than those produced by LDA topics. In terms of clustering, CorEx consistently produces document clusters of higher homogeneity than LDA. On the disaster relief articles, the CorEx clusters are nearly twice as homogeneous as the LDA clusters.

CorEx outperforms LDA in terms of topic coherence on two out of three of the datasets. While LDA

Rank	Disaster Relief Topic
1	drought, farmers, harvest, crop, livestock, planting, grain, maize, rainfall, irrigation
3	eruption, volcanic, lava, crater, eruptions, volcanos, slopes, volcanic activity, evacuated, lava flows
8	winter, snow, snowfall, temperatures, heavy snow, heating, freezing, warm clothing, severe winter, avalanches
23	military, armed, civilians, soldiers, aircraft, weapons, rebel, planes, bombs, military personnel
Rank	**20 Newsgroups Topic**
3	team, game, season, player, league, hockey, play, teams, nhl
14	car, bike, cars, engine, miles, road, ride, riding, bikes, ground
26	nasa, launch, orbit, shuttle, mission, satellite, gov, jpl, orbital, solar
39	medical, disease, doctor, patients, treatment, medicine, health, hospital, doctors, pain
Rank	**Clinical Health Notes Topic**
12	vomiting, nausea, abdominal pain, diarrhea, fever, dehydration, chill, clostridium difficile, intravenous fluid, compazine
19	anxiety state, insomnia, ativan, neurontin, depression, lorazepam, gabapentin, trazodone, fluoxetine, headache
27	pain, oxycodone, tylenol, percocet, ibuprofen, morphine, osteoarthritis, hernia, motrin, bleeding

Table 1: Examples of topics learned by the CorEx topic model. Words are ranked according to mutual information with the topic, and topics are ranked according to the amount of total correlation they explain. Topic models were run with 50 topics on the Reliefweb and 20 Newsgroups datasets, and 30 topics on the clinical health notes.

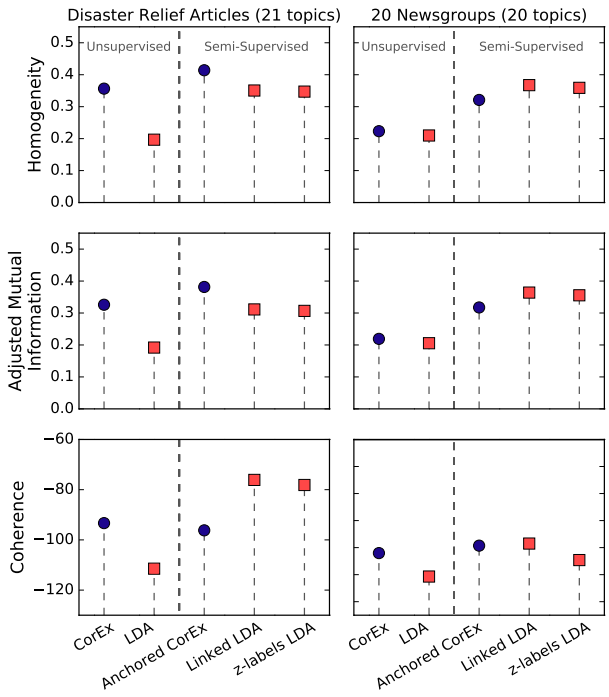

Figure 3: Comparison of anchored CorEx to other semi-supervised topic models in terms of document clustering and topic coherence. For each dataset, the number of topics is fixed to the number of document labels. Each dot is the average of 30 runs. Confidence intervals are plotted but are so small that they are not distinguishable.

Rank	Anchored Disaster Relief Topic
1	**harvest**, **locus**, drought, **food crisis**, farmers, crops, crop, malnutrition, food aid, livestock
4	**tents**, **quake**, international federation, red crescent, red cross, blankets, earthquake, **richter scale**, societies, aftershocks
12	**climate**, **impacts**, **warming**, climate change, irrigation, consumption, household, droughts, livelihoods, interventions
19	**storms**, weather, winds, coastal, **tornado**, meteorological, **tornadoes**, strong winds, tropical, roofs
Rank	**Anchored 20 Newsgroups Topic**
5	government, **congress**, **clinton**, state, national, economic, general, states, united, order
6	**bible**, **christian**, god, jesus, christians, believe, life, faith, world, man
15	use, used, high, **circuit**, power, work, voltage, need, low, end
20	**baseball**, **pitching**, **braves**, **mets**, **hitter**, pitcher, cubs, dl, sox, jays

Table 2: Examples of topics learned by CorEx when simultaneously anchoring many topics with anchoring parameter $\beta = 2$. Anchor words are shown in **bold**. Words are ranked according to mutual information with the topic, and topics are ranked according to the amount of total correlation they explain. Topic models were run with 21 topics on the Reliefweb articles and 20 topics on the 20 Newsgroups dataset.

produces more coherent topics for the clinical health notes, it is particularly striking that CorEx is able to produce high quality topics while only leveraging binary count data. Examples of these topics are shown in Table 1. Despite the binary counts limitation, CorEx still finds meaningfully coherent and competitive structure in the data.

5.2 Anchored CorEx Analysis

We now examine the effects and benefits of guiding CorEx through anchor words. In doing so, we also compare anchored CorEx to other semi-supervised topic models.

5.2.1 Anchoring for Topic Separability

We are first interested in how anchoring can be used to encourage topic separability so that documents cluster well. We focus on the HA/DR articles and 20 newsgroups datasets, since traditional clustering metrics are not well-defined on the multi-label clinical health notes. For both datasets, we fix the number of topics to be equal to the number of document labels. It is in this context that we compare anchored CorEx to two other semi-supervised topic models: z-labels LDA and must/cannot link LDA.

Using the method described in Section 4.3, we automatically retrieve the top five anchors for each disaster type and newsgroup. We then filter these lists of any words that are ambiguous, i.e. words that are anchor words for more than one document label. For anchored CorEx and z-labels LDA we simultaneously assign each set of anchor words to exactly one topic each. For must/cannot link LDA, we create must-links within the words of the same anchor

group, and create cannot-links between words of different anchor groups.

Since we are simultaneously anchoring to many topics, we use a weak anchoring parameter $\beta = 2$ for anchored CorEx. Using the notation from their original papers, we use $\eta = 1$ for z-labels LDA, and $\eta = 1000$ for must/cannot link LDA. For both LDA variants, we use $\alpha = 0.5$, $\beta = 0.1$ and take 2,000 samples, and estimate the models using code implemented by the original authors.

The results of this comparison are shown in Figure 3, and examples of anchored CorEx topics are shown in Table 2. Across all measures CorEx and anchored CorEx outperform LDA. We find that anchored CorEx always improves cluster quality versus CorEx in terms of homogeneity and adjusted mutual information. Compared to CorEx, multiple simultaneous anchoring neither harms nor benefits the topic coherence of anchored CorEx. Together these metrics suggest that anchored CorEx is finding topics that are of equivalent coherence to CorEx, but more relevant to the document labels since gains are seen in terms of document clustering.

Against the other semi-supervised topic models, anchored CorEx compares favorably. The document clustering of anchored CorEx is similar to, or better than, that of z-labels LDA and must/cannot link LDA. Across the disaster relief articles, anchored CorEx finds less coherent topics than the two LDA variants, while it finds similarly coherent topics as must/cannot link LDA on the 20 newsgroup dataset.

5.2.2 Anchoring for Topic Representation

We now turn to studying how domain knowledge can be anchored to a single topic to help an otherwise dominated topic emerge, and how the anchoring parameter β affects that emergence. To discern this effect, we focus just on anchored CorEx along with the HA/DR articles and clinical health notes, datasets for which we have a domain expert lexicon.

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. Unlike in the previous section, we do not filter these lists of ambiguous anchor words. 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.

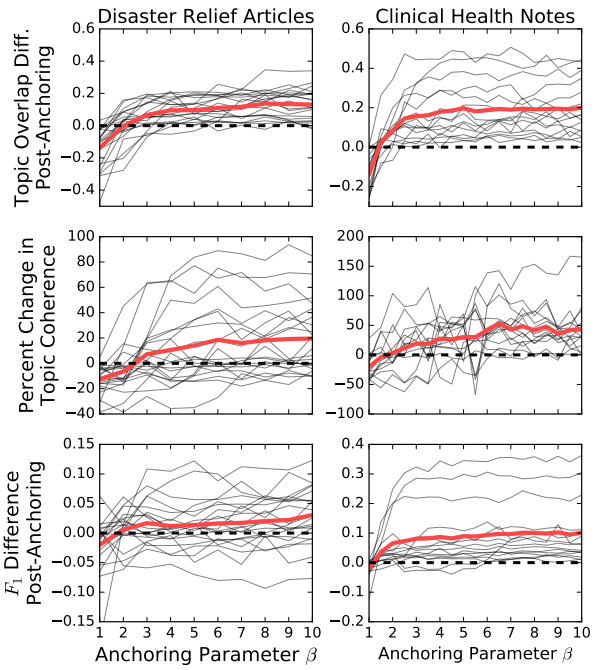

Figure 4: Effect of anchoring words to a single topic for one document label at a time as a function of the anchoring parameter β. Light gray lines indicate the trajectory of the metric for a given disaster or disease label. Thick red lines indicate the pointwise average across all labels for fixed value of β.

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 4 we show how the results of this experiment vary as a function of the anchoring parameter β for each disaster and disease type in the two data sets. Since there is heavy variance across document labels for each metric, we also examine a more detailed cross section of these results in Figure 5, where we set $\beta = 5$ for the clinical health notes and set $\beta = 10$ for the disaster relief articles. As we show momentarily, disaster and disease types that benefit the most from anchoring were un-

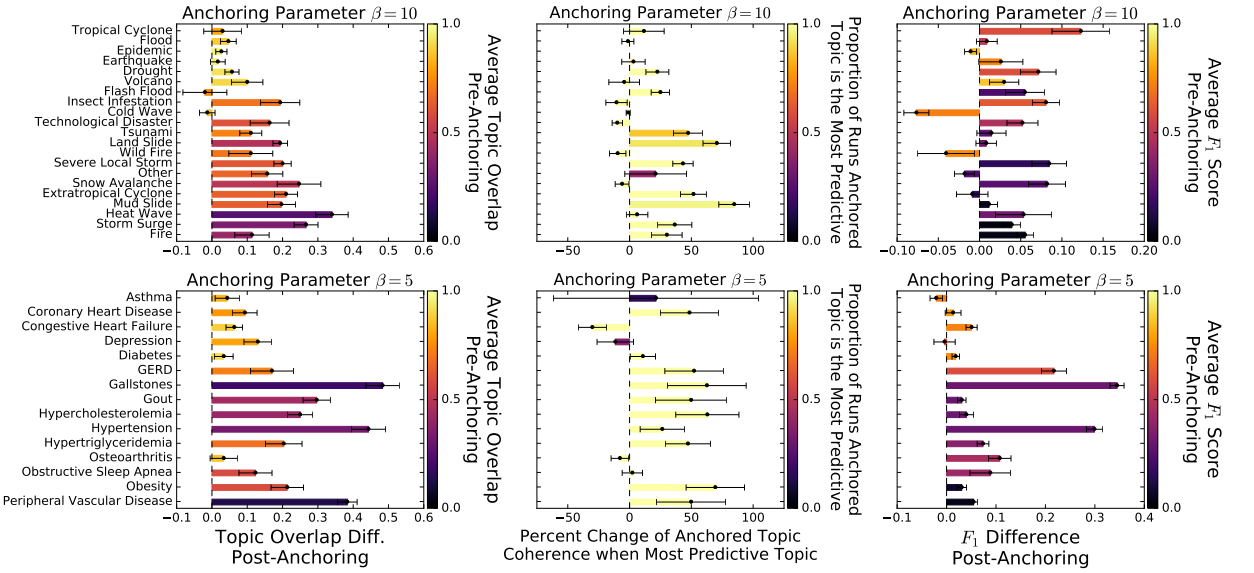

Figure 5: Cross-section results of the anchoring metrics from fixing $\beta = 5$ for the clinical health notes, and $\beta = 10$ for the disaster relief articles. Disaster and disease types are sorted by frequency, with the most frequent document labels appearing at the top. Error bars indicate 95% confidence intervals. The color bars provide context 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.

derrepresented pre-anchoring. Document labels that were well-represented prior to anchoring achieve only marginal gain. This results in the variance seen in Figure 4.

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 4 (row 1) we see that as β 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 5 (column 1), 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 occur the most frequently are those for which 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, i.e. the topic with the largest corresponding coefficient in magnitude of the logistic regression, when the anchored topic itself is most predictive. From Figure 4 (row 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 colors of the central panel of Figure 5, 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 4 (row 3) 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 which document labels appeared most frequently. For example, we see in Figure 5 (column 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 anchoring single topics within CorEx for document classification, it is important to examine how the anchoring affects prediction for individual document labels.

5.2.3 Anchoring for Topic Aspects

Finding topics that revolve around a word, such as a name or location, or a group of words can aid in understanding how a particular subject or event has been framed. We finish with a qualitative experiment where we disambiguate aspects of a topic by anchoring a set of words to multiple topics within the CorEx topic model. Note, must/cannot link LDA cannot be used in this manner, and z-labels LDA would require us to know these aspects beforehand.

We consider tweets containing #Ferguson (case-insensitive), which detail reactions to the shooting of Black teenager Michael Brown by White police officer Darren Wilson on August 9th, 2014 in Ferguson, Missouri. These tweets were collected from the Twitter Gardenhose, a 10% random sample of all tweets, over the period August 9th, 2014 to November 30th, 2014. Since CorEx will seek maximally informative topics by exploiting redundancies, we remove duplicates of retweets, leaving us with 869,091 tweets. We filter these tweets of punctuation, stop words, hyperlinks, usernames, and the 'RT' retweet symbol, and use the top 20,000 word types.

In the wake of both the shooting and the eventual non-indictment of Darren Wilson, several protests occurred. Some onlookers supported and encouraged such protests, while others characterized the protests as violent "riots." To disambiguate these different depictions, we train a CorEx topic model with 55 topics, anchoring "protest" and "protests" together to five topics, and "riot" and "riots" together to five topics with $\beta = 2$. These anchored topics are presented in Table 3.

	Topic Aspects of "protest"
1	**protest**, **protests**, peaceful, violent, continue, night, island, photos, staten, nights
2	**protest**, **protests**, #hiphopmoves, #cole, hiphop, nationwide, moves, fo, anheuser, boeing
3	**protest**, **protests**, st, louis, guard, national, county, patrol, highway, city
4	**protest**, **protests**, paddy, covering, beverly, walmart, wagon, hills, passionately, including
5	**protest**, **protests**, solidarity, march, square, rally, #oakland, downtown, nyc, #nyc
	Topic Aspects of "riot"
6	**riot**, **riots**, unheard, language, inciting, accidentally, jokingly, watts, waving, dies
7	**riot**, black, **riots**, white, #tcot, blacks, men, whites, race, #pjnet
8	**riot**, **riots**, looks, like, sounds, acting, act, animals, looked, treated
9	**riot**, **riots**, store, looting, businesses, burning, fire, looted, stores, business
10	gas, **riot**, tear, **riots**, gear, rubber, bullets, military, molotov, armored

Table 3: Topic aspects around "protest" and "riot" from running a CorEx topic model with 55 topics and anchoring "protest" and "protests" together to five topics and "riot" and "riots" together to five topics with $\beta = 2$. Anchor words are shown in **bold**. Note, topics are not ordered by total correlation.

The anchored topics reflect different aspects of the framing of the "protests" and "riots," and are generally interpretable, despite the typical difficulty of extracting coherent topics from short documents using LDA (Tang et al., 2014). The "protest" topic aspects describe protests in St. Louis, Oakland, Beverly Hills, and parts of New York City (topics 1, 3, 4, 5), resistance by law enforcement (topics 3 and 4), and discussion of whether the protests were peaceful (topic 1). Topic 2 revolves around hip-hop artists who marched in solidarity with protesters.

The "riot" topic aspects discuss racial dynamics of the protests (topic 7) and suggest the demonstrations are dangerous (topics 8 and 9). Topic 10 describes the "riot" gear used in the militarized response to the Ferguson protesters, and Topic 7 also hints at aspects of conservatism through the hashtags #tcot (Top Conservatives on Twitter) and #pjnet (Patriot Journalist Network).

As we see, anchored CorEx finds several interesting, non-trivial aspects around "protest" and "riot" that could spark additional qualitative investigation. Retrieving topic aspects through anchor words in this manner allows the user to explore different frames of complex issues, events, or discussions within documents. As with the other anchoring strategies, this has the potential to supplement qualitative research done by researchers within the social sciences and digital humanities.

6 Discussion

We have introduced an information-theoretic topic model, CorEx, that does not rely on any of the generative assumptions of LDA-based topic models. This topic model seeks maximally informative topics as encoded by their total correlation. 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. Both CorEx and anchored CorEx consistently produce topics that are of comparable quality to LDA-based methods, despite only making use of binarized word counts.

Anchored CorEx is more flexible than previous attempts at integrating word-level information into topic models. Topic separability can be enforced by lightly anchoring disjoint groups of words to separate topics, topic representation can be promoted by assertively anchoring a group of words to a single topic, and topic aspects can be unveiled by anchoring a single group of words to multiple topics. The flexibility of anchoring through the information bottleneck lends itself to many other possible creative anchoring strategies that could guide the topic model in different ways. Different goals may call for different anchoring strategies, and domain experts can shape these strategies to their needs.

While we have demonstrated several advantages of the CorEx topic model to LDA, it does have some technical shortcomings. Most notably, CorEx relies on binary count data in its sparsity optimization, rather than the standard count data that is used as input into LDA and other topic models. While we have demonstrated CorEx performs at the level of LDA despite this limitation, its effect would be more noticeable on longer documents. This can be partly overcome if one chunks such longer documents into shorter subdocuments prior to running the topic model. Our implementation also requires that each word appears in only one topic. These limitations 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, the information-theoretic approach provided via CorEx has rich potential for finding meaningful structure in documents, particularly in a way that can help domain experts guide topic models with minimal intervention to capture otherwise eclipsed themes. The lightweight and versatile framework of anchored CorEx leaves open possibilities for theoretical extensions and novel applications within the realm of topic modeling.

Acknowledgments

We would like to thank the Machine Intelligence and Data Science (MINDS) research group at the Information Sciences Institute for their help and insight during the course of this research. We also thank the Vermont Advanced Computing Core (VACC) for its computational resources. Finally, we thank the anonymous reviewers and the TACL action editors Diane McCarthy and Kristina Toutanova for their time and effort in helping us improve our work. Ryan J. Gallagher was a visiting research assistant at the Information Sciences Institute while performing this research. Ryan J. Gallagher and Greg Ver Steeg were supported by DARPA award HR0011-15-C-0115 and David Kale 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.

David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. 2011. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 22, page 1171.

Sanjeev Arora, Rong Ge, and Ankur Moitra. 2012. Learning topic models–going beyond SVD. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS)*, 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 *Proceedings of International Conference on Machine Learning*, 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.

Wray Buntine and Aleks Jakulin. 2006. Discrete component analysis. In *Subspace, Latent Structure and Feature Selection*, pages 1–33. Springer.

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 K. M. Poon, and Zhourong Chen. 2016. Progressive EM for latent tree models and hierarchical topic detection. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 1498–1504.

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.

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.

James Foulds, Shachi Kumar, and Lise Getoor. 2015. Latent topic networks: A versatile probabilistic programming framework for topic models. In *Proceedings of the International Conference on Machine Learning*, pages 777–786.

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.

Thomas L. Griffiths, Michael I. Jordan, Joshua B. Tenenbaum, and David M. Blei. 2004. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems*, pages 17–24.

Yoni Halpern, Youngduck Choi, Steven Horng, and David Sontag. 2014. Using anchors to estimate clinical state without labeled data. In *AMIA Annual Symposium Proceedings*. 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)*.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 289–296.

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.

Moontae Lee and David Mimno. 2014. Low-dimensional embeddings for interpretable anchor-based topic inference. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 1319–1328.

Moshe Lichman. 2013. UC Irvine Machine Learning Repository.

Jon D. McAuliffe and David M. Blei. 2008. Supervised topic models. In *Advances in Neural Information Processing Systems*, pages 121–128.

Shike Mei, Jun Zhu, and Jerry Zhu. 2014. Robust RegBayes: Selectively incorporating first-order logic domain knowledge into Bayesian models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 253–261.

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 *Proceedings of the Association of Computational Linguistics*, 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 *Proceedings of North American Chapter of the Association for Computational Linguistics*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Oliver Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard 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 Modeling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

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

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494.

Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. 2014. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the International Conference on Machine Learning*, pages 190–198.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. In *Proceedings of 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

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 *Artificial Intelligence and Statistics*, pages 1004–1012.