https://dx.doi.org/10.1016/j.bdr.2020.100173

Agreeing to Disagree: Choosing among Eight Topic-Modeling Methods

Qiang Fu
Department of Sociology
The University of British Columbia
Vancouver, BC, Canada
Email: qiang.fu@ubc.ca

Yufan Zhuang
IBM Research
Yorktown Heights, NY, USA
Email: yz3453@columbia.edu

Jiaxin Gu
Department of Sociology
The University of British Columbia
Vancouver, BC, Canada
Email: gujiaxinsoci@gmail.com

Yushu Zhu
Urban Studies Program
and School of Public Policy
Simon Fraser University
Vancouver, BC, Canada
Email: yushu_zhu@sfu.ca

Xin Guo
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hong Kong; and
School of Mathematics and Physics
The University of Queensland
Brisbane, QLD 4072, Australia
Email: xin.guo@uq.edu.au

Abstract—Topic modeling is a key research area in natural language processing and has inspired innovative studies in a wide array of social-science disciplines. Yet, the use of topic modeling in computational social science has been hampered by two critical issues. First, social scientists tend to focus on a few standard ways of topic modeling. Our understanding of semantic patterns has not been informed by rapid methodological advances in topic modeling. Moreover, a systematic comparison of the performance of different methods in this field is warranted. Second, the choice of the optimal number of topics remains a challenging task. A comparison of topic-modeling techniques has rarely been situated in a social-science context and the choice appears to be arbitrary for most social scientists. Based on about 120,000 Canadian newspaper articles since 1977, we review and compare eight traditional, generative, and neural methods for topic modeling (Latent Semantic Analysis, Principal Component Analysis, Factor Analysis, Non-negative Matrix Factorization, Latent Dirichlet Allocation, Neural Autoregressive Topic Model, Neural Variational Document Model, and Hierarchical Dirichlet Process). Three measures (coherence statistics, held-out likelihood, and graph-based dimensionality selection) are then used to assess the performance of these methods. Findings are presented and discussed to guide the choice of topic-modeling methods, especially in social science research.

Index Terms—Topic Modeling, Natural Language Processing, Computational Social Science, Optimal Number of Topics

I. Introduction

Topic modeling refers to a group of statistical and machine-learning methods which are used to extract meaningful topics and explore semantic patterns of digitized text data. In recent years, the topic-modeling methods have increasingly been adopted by scholars from different disciplines who are interested in big data research and computational social science [1]–[4]. Most noteworthy among these research efforts is the wide use of statistical generative models, such as the latent Dirichlet allocation (LDA), in social science research.

Yet, computational social science has not been well informed by an explosion in methods and algorithms for topic modeling in the past two decades [1], [5], [6]. As suggested by DiMaggio [7], most topic-modeling techniques require various methodological decisions that many social scientists are unfamiliar with, have not considered, or lack experience with. Moreover, a systematic comparison of traditional and novel methods has not been explicitly conducted to guide the application of topic modeling in social sciences. In particular, although the choice of the optimal number of topics (also referred as the choice of K in the statistics and machine-learning literature) is a critical decision for researchers to explore semantic patterns and specify the abstraction of meaningful components in a text corpus, the choice of K still appears to be a black box for most social scientists and is subject to their subjective, if not arbitrary, assessment.

Drawing on 119,480 articles published by three mainstream Canadian newspapers (The Globe and Mail, The Toronto Star, and National Post) from January 1st 1977 to June 30th 2019, we review and use eight topic-modeling methods (Latent Semantic Analysis, Principal Component Analysis, Factor Analysis, Non-negative Matrix Factorization, Latent Dirichlet Allocation, Neural Autoregressive Topic Model, Neural Variational Document Model, and Hierarchical Dirichlet Process) to assess the choice of K in topic modeling. By adopting various data reduction, statistical generative, and neural variational modeling techniques, our investigation aims to provide a more holistic view of the application of topic modeling and offer practical guidance for the choice of K, especially for computational social scientists.

II. Preprocessing and Word Representation

We apply the following procedures to the text corpus before it is processed by topic modeling. Common stop words in English [8] such as "the", "a", and "an" are removed. Next we apply the RAKE (Rapid Automatic Keyword Extraction) algorithm [9] to identify key phrases in the corpus, and then combine words into phrases such that words like "united" and "kingdom" are combined into "united kingdom". After the data-cleaning procedure, we represent the text corpus using a document-word matrix X: each column of the matrix corresponds to a document and each row of the matrix corresponds to a word [10]. To indicate a word's relative importance in the corpus, elements of the matrix are also weighted by the conventional term frequency-inverse document frequency (tf-idf) [11]. One way to calculate the tf-idf weight $w_{t,d}$ of a term (word) t and a document d is as follows [12],

$$w_{t,d} = \mathsf{tf}_{t,d} \times \log \frac{N}{\mathsf{df}_t},$$

where $\mathsf{tf}_{t,d}$ is a term t's frequency in a document d, N is the total number of documents, and df_t represents the total number of documents in a text corpus containing the term t. $w_{t,d}$ increases if a term has a higher frequency in a document but such increase in the term-frequency weight is offset by this term's popularity across all documents in a corpus. This tf-idf weight thus filters out popular common words in a text corpus.

III. Topic Modeling Methods Investigated

We next briefly review these eight topic-modeling methods to be investigated in this study.

A. Latent Semantic Analysis

Latent semantic analysis (LSA), which draws on singular value decomposition and a low-rank approximation of a document-word matrix, has long been adopted by researchers from different fields to identify meaningful themes in a text corpus [13], [14]. To illustrate how LSA works, we have the singular value decomposition (SVD) of a document-word matrix X as:

$$X = U\Sigma V^T,$$

where both U and V are orthogonal matrices and Σ is a diagonal matrix. To understand the three matrices, we note that the square matrix XX^T contains all dot products denoting the correlation between any two word vectors across all documents, and X^TX contains all dot products denoting the correlation between any two documents. We have:

$$U^TXX^TU = \Sigma\Sigma^T \text{and } V^TX^TXV = \Sigma^T\Sigma, \text{or}$$
$$XX^T = U\Sigma\Sigma^TU^T \text{and } X^TX = V\Sigma^T\Sigma V^T.$$

XX^T and X^TX have the same non-zero eigenvalues expressed by $\Sigma\Sigma^T$ (or, equally by $\Sigma^T\Sigma$), and their corresponding eigenvectors are contained in U and V, respectively.

The number of positive singular values in Σ corresponds to the rank of X, or the number of topics in topic modeling, while the values of singular values indicate the relative importance of these topics. For a space spanned by singular vectors associated with these singular values, the coordinates of a word i across different topics are denoted by the i^{th} row of U and the coordinates of a document j across all topics are denoted by the j^{th} column of V^T. The corresponding loadings of all words on the k^{th} topic are denoted by elements in the k^{th} columns of U; and the corresponding loadings of all documents on the k^{th} topic are denoted by elements in the k^{th} rows of V^T. While topics identified by the LSA method can be expressed by clusters of words and/or documents once they are projected to a semantic space, we use columns of U to denote topics and their corresponding relations with words. If the values of singular values are below a specific threshold, researchers can remove these small singular values to achieve a low-rank approximation of the document-word matrix X [15].

B. Principal Component Analysis

Principal component analysis (PCA) can be viewed as an extension of SVD [16]. To identify distinctive features of its covariance matrix XX^T, a document-word matrix X is projected into orthogonal directions. PCA is looking for a projection matrix P so that, after the projection, the covariance matrix YY^T of the new document-word matrix $Y = PX$ has the largest variance in these projection directions. In this process, these projection directions suggested by the projection matrix P correspond to the basis vectors, which are orthogonal to each other. Otherwise, for example, the direction of the eigenvector associated with the second largest eigenvalue (variance) can be parallel to or even overlap with that associated with the largest eigenvalue (and so forth for the directions of the remaining eigenvectors), which cannot suggest distinctive features of the data. The off-diagonal elements (i.e., covariance) of YY^T should consequently be zero. We have:

$$YY^T = (PX)(PX)^T = PXX^TP^T = D,$$

where D should be a diagonal matrix. Now we rank the normalized eigenvectors $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_p$ of XX^T to have a new orthogonal matrix $Z = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_p)$, and let:

$$Z^TXX^TZ = \Sigma^T\Sigma = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix}. \quad (1)$$

Here p is the number of words. D becomes a diagonal matrix when we make $P = Z^T$. The projection also corresponds to the maximization of $\mathbf{z}_i^T XX^T \mathbf{z}_i$ when $\mathbf{z}_i^T \mathbf{z}_i = 1$. If we take the derivative of $\mathbf{z}_i^T XX^T \mathbf{z}_i - \lambda \mathbf{z}_i^T \mathbf{z}_i$ with respective to $\mathbf{z}_i$, $\mathbf{z}_i$ must be an eigenvector of XX^T since $(XX^T - \lambda I)\mathbf{z}_i = 0$. The matrix containing all the eigenvectors of XX^T provides the loadings of all words on any topic, which brings a PCA approach to topic modeling. We treat these principal components as topics, and can obtain words of a topic through loadings of a principal component. The LSA and PCA models are similar to each other in that they both extract "components" from a document-word matrix X. These components can contain both positive and negative values. Yet, the interpretation of negative values in a topic-modeling setting can be difficult.

C. Factor Analysis

While LSA and PCA aim to extract major components from the data matrix, factor analysis (FA) tries to represent the data matrix and its internal relations through latent variables (or factors) based on a parametric model and a series of assumptions. The idea of FA can be illustrated as follows. We obtain a new document-word matrix X_* by centering each row of X. That is, we center the weight of each word across the whole corpus of N documents. We next represent the p words using latent factors:

$$Y_{N\times p} = X_*^T = F_{N\times k} A_{k\times p} + \varepsilon_{N\times p},$$

where F is a matrix, with the i-th column containing N observations (assumed independent) of the i-th factor F_i (which is a real-valued random variable), A is a matrix representing the coefficients (called loadings) of all the words on each of the k factors, and ε is a matrix of random variables modeling the error, which is sometimes referred to as specific factors. The FA model is defined with the following assumptions.

1) The factors $F_1, \ldots, F_k$ are assumed mean-zero, and the covariance matrix of the vector $(F_1, \ldots, F_k)^T$ is assumed to be the $k \times k$ identity matrix;
2) Each row of ε is considered to be an independent replication of the random vector $\vec{\varepsilon}$, where $\vec{\varepsilon} \in \mathbb{R}^p$, $\mathbb{E}[\vec{\varepsilon}] = 0$, and $\mathrm{Cov}(\vec{\varepsilon}) = \Psi = \mathrm{diag}\{\Psi_1, \ldots, \Psi_p\}$;
3) $\mathrm{Cov}(\vec{\varepsilon}, (F_1, \ldots, F_p)^T) = 0$.

Let Y_i^T be the i-th row of Y. From these assumptions we have

$$\begin{aligned}\mathrm{Cov}(Y_i) &= \mathrm{Cov}\left(A^T (F_1, \ldots, F_k)^T + \vec{\varepsilon}\right) \\ &= A^T I A + \Psi = A^T A + \Psi.\end{aligned}$$

The identity $\mathrm{Cov}(Y_i) = A^T A + \Psi$ has two implications. First, it is possible for researchers to estimate the loading matrix A first, and then derive the latent factors. Second, for any $1 \leq i \leq N$, we write $Y_i = (Y_{i,1}, \ldots, Y_{i,p})^T$. We then consider the j-th word to obtain

$$\mathrm{Var}(Y_{i,j}) = (A^T A + \Psi)_{j,j} = \|a_j\|^2 + \Psi_j,$$

where $a_j \in \mathbb{R}^k$ is the j-th column of A, namely, $A = [a_1, \ldots, a_p]$. Also, $\mathrm{Cov}(Y_{i,j}, Y_{i,l}) = \langle a_j, a_l \rangle$ if $j \neq l$. The sum of squared loadings of $Y_{i,j}$ on all the k factors, $\|a_j\|^2$, denotes the extent to which Y_j is explained by all factors (the dependence of $Y_{i,j}$ on all factors).

We use the EM algorithm to implement factor analysis [17], [18]. To have better explanatory power, these independent factors are often rotated to achieve maximum variance.

The link between PCA and FA has been noted in existing literature [13], [19]. In particular, consider the SVD of the estimated covariance matrix,

$$\frac{1}{N-1} Y^T Y = U \Lambda U^T,$$

where $U = [U_1, \ldots, U_p] \in \mathbb{R}^{p\times p}$ is an orthogonal matrix, and $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$ is a diagonal matrix storing the ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. We have

$$\begin{aligned}A^T A \approx \mathrm{Cov}(Y_i) &\approx \frac{1}{N-1} Y^T Y = U \Lambda U^T \\ &= \left(\sqrt{\lambda_1} U_1, \ldots, \sqrt{\lambda_p} U_p\right) \begin{pmatrix} \sqrt{\lambda_1} U_1^T \\ \vdots \\ \sqrt{\lambda_p} U_p^T \end{pmatrix}.\end{aligned}$$

Therefore, the first k vectors can be used to form an estimate of A,

$$\hat{A} = \begin{pmatrix} \sqrt{\lambda_1} U_1^T \\ \vdots \\ \sqrt{\lambda_k} U_k^T \end{pmatrix}.$$

Alternatively, one may use this matrix as a starting point for estimation. The resulted factors are considered as weight vectors for each topic. We specify the top words in a topic according to the same principle as previously discussed for LSA and PCA. The words are sorted according to their factor values and only these with reasonably high values are retained.

D. Non-negative Matrix Factorization

Non-negative matrix factorization (NMF), or non-negative matrix approximation, factorizes a matrix V into two matrices W and H, where all elements of the three matrices are not negative [20]:

$$V_{n\times m} \approx W_{n\times r} H_{r\times m}.$$

The dimension of r is often much smaller than that of m and n. The advantage of NMF over other factorization algorithms can be illustrated as follows. By making every element in these matrices non-negative, any column vector v_i in V is represented by a weighted sum of column vectors in W, and the corresponding weights of column vectors are expressed by elements in the i-th column of H:

$$v_i \approx w_1 h_{1i} + w_2 h_{2i} + \cdots + w_r h_{ri} = W h_i.$$

Using this non-negative factorization technique, researchers can study how a whole system consists of different parts through these positive weights. The idea behind NMF is inherently connected to how the relations between a whole system and its different parts are perceived by human beings [20].

The relevance of NMF to topic modeling, especially the probabilistic latent semantic analysis (pLSA), has been discussed [21]. For a document-word matrix X, we define elements of W as $w_{ik} = P(\text{topic}_k)P(\text{word}_i|\text{topic}_k)$, define elements in H as $h_{kj} = P(\text{document}_j|\text{topic}_k)$, and write elements x_{ij} as:

$$\begin{aligned} x_{ij} &= \sum w_{ik}h_{kj} \\ &= \sum P(\text{topic}_k)P(\text{word}_i|\text{topic}_k)P(\text{document}_j|\text{topic}_k). \end{aligned}$$

The idea described here is in line with that of pLSA, where a probabilistic model is used to generate topics, and words/documents are then generated based on the distribution of topics.

E. Latent Dirichlet Allocation

Based on a generative statistical model, LDA uses latent factors to capture semantic similarities of words and documents [22]. The procedure of LDA can be summarized as follows. To begin with, researchers need to specify the optimal number of topics K. Let p be the total number of words we study. A specific document $\mathbf{w}$ is modeled as a sequence of words $\mathbf{w} = (w_1, \ldots, w_\ell)$ of length $\ell \sim \text{Poisson}(\xi)$, where ξ is pre-specified. For this document $\mathbf{w}$, a K-dimensional probability vector θ with its non-negative coordinates summed to one is used to model the topic mixture. To generate θ, one uses $\theta \sim \text{Dirichlet}(\alpha)$ with $\alpha \in \mathbb{R}_+^K$ left for estimation. For each $1 \leq n \leq \ell$, a random topic $z_n \in \{1, \ldots, K\}$ is assigned to the n'th word with $z_n \sim \text{Multinomial}(\theta)$. Eventually, the n'th word $w_n \in \{1, \ldots, p\}$ is drawn randomly from $\text{Multinomial}(\beta_{z_n})$. Here $\beta = [\beta_1, \ldots, \beta_K]$ is a $p \times K$ matrix to be estimated, of which the i'th column is a probability vector characterizing the distribution of the p words in the topic i. In [23], the likelihood

$$L(\alpha, \beta|\theta, \mathbf{z}, \mathbf{w}) = p(\theta|\alpha) \prod_{n=1}^{\ell} [p(z_n|\theta)p(w_n|\beta_{z_n})]$$

is multiplied through all the documents, and maximized with the technique of variational inference, for the estimation of α and β.

F. A Neural Autoregressive Topic Model: DocNade

As informed by the Replicated Softmax [24] and the Neural Autoregressive Distribution Estimator (Nade) [25], the DocNade uses a neural autoregressive model to process multinomial word distributions and learn meaningful word representations from unlabelled texts [26]. The Replicated Softmax can be viewed as a generalization of the restricted Boltzmann machine (RBM), which deals with binary observed and hidden (latent) variables. The Replicated Softmax can handle multinomial observed variables, with shared connections among each multinomial observation and latent variables.

One disadvantage of the RBM is that the calculation of conditional probabilities is intractable and needs to be approximated by mean-field inference. Drawing upon the fact that a D-dimensional distribution can be denoted as a product of conditional distributions (the probability chain rule), the Nade assumes that the output in every step is a linear combination of the previous values and passes the inputs through a feed-forwarding neural network. The product of these previous conditional probabilities constitute a joint distribution over observations and can be readily maximized via the gradient of the negative log-likelihood [25].

By combining the Replicated Softmax and the Nade, the DocNade adopts a tree of binary logistic regressions to model conditional probabilities at each step [26]. More specifically, each root-to-leaf path in the probabilistic tree represents a word [27]. Moreover, each transition in the tree is controlled by a set of binary regressors and the occurrence of a specific word is determined by a product of transition probabilities of a particular tree path. Compared with the Replicated Softmax, the introduction of tree nodes has greatly reduced the computational complexity of the DocNade: its training complexity scales logarithmically, rather than linearly, with the vocabulary size [26]. Finally, it should be noted that the DocNade aims to provide a holistic view of semantic patterns in a document because it uses permutations of words in the whole document regardless of their order of appearance.

G. A Neural Variational Document Model

The Neural Variational Document Model (NVDM) provides a neural variational framework for topic modeling [28], [29]. Probability generative models including the LDA often rely on an analytical approximation (e.g., variational Bayes) for the distributions over latent variables. Yet, a high dimensional integration in Bayesian inference often becomes intractable when generative models are complex. Instead, the neural variational inference framework as a deep-learning method uses inference networks such as multilayer perceptrons (MLP) to model posterior probabilities of latent semantics [28]. In other words, the posterior probabilities are "learned" when the connection weights of perceptrons are updated via minimizing performance errors. As an unsupervised generative model, NVDM extracts a semantic latent variable for each document via an MLP encoder, which compresses text representations into hidden vectors, and uses a softmax decoder to generate the words. Similar to LDA, it deals with the bag-of-words representation.

The NVDM can be explained as follows [28]. For a latent variable h and a document-word vector x, we have the posterior distribution of the latent variable h as:

$$p(h|x) = \frac{p(h,x)}{p(x)}$$

where $p(x) = \int p(h,x)dh$. The idea of variational inference is to update the inference-network parameters ϕ so that an MLP encoder $q_\phi(h|x)$ is close to the posterior distribution $p_\theta(h|x)$, where θ parameterizes the generative distributions $p(h|x)$ [28]. The optimality can be achieved by minimizing their KL divergence, where the KL divergence characterises the difference between the cross entropy of the distribution q relative to the distribution p, and the entropy of p.

$$\begin{aligned}\mathsf{KL}(q_\phi(h|x)||p_\theta(h|x)) =&\mathbb{E}_{q_\phi(h|x)}[\log q_\phi(h|x)]\\ &- \mathbb{E}_{q_\phi(h|x)}[\log p_\theta(h|x)].\end{aligned}$$

It should be noted, however, the NVDM as an unsupervised method directly draws h from a prior $p(h)$ rather than the conditional distribution $p_\theta(h|x)$.

$$\begin{aligned}\mathsf{KL}(q_\phi(h|x)||p(h)) =&\mathbb{E}_{q_\phi(h|x)}[\log q_\phi(h|x)]\\ &- \mathbb{E}_{q_\phi(h|x)}[\log p(h)].\end{aligned}$$

The Evidence Lower Bound (ELBO) can be defined as follows.

$$L(\theta, \phi, x) = \log p(x) - \mathsf{KL}(q_\phi(h|x)||p(h)).$$

The decoder $p_\theta(x|h)$ is a softmax function shared across documents. Clearly, to maximize the lower bound, one needs to have maximized likelihood and minimized KL divergence. To utilize information provided by the encoder, a weight is often added to the KL divergence term [29].

H. Hierarchical Dirichlet Process

Different from LDA, the Hierarchical Dirichlet Process (HDP) algorithm adopts Dirichlet processes for topic modeling and allows the optimal number of topics to change as the size of a text corpus increases. Intuitively, one may expect a finer resolution of topics when a larger corpus is at stake. Although researchers do not need to specify the optimal number of topics as a hyper-parameter in HDP, an integer (often set as 150) is still needed to determine the right truncation of K. This integer represents an upper bound for the maximum number of topics. In HDP, each document is modeled by a probability distribution G concentrated on a countable set, where G is independently sampled from a Dirichlet process. Each word x is modeled as a random draw from a distribution $F(\phi)$ parameterized by a factor ϕ, which is randomly drawn from G. Due to the enhanced hierarchical structure of HDP, HDP is more complex and requires a higher computational cost than LDA for parameter estimation. In this research, the truncation number of topics for HDP is set as 1000.

IV. Data And Measures

A. Data

The text corpus was obtained from three mainstream newspapers in Canada: The Globe and Mail, The Toronto Star and National Post. All articles containing the word "Chinese" were retrieved and the reference period is from January 1st 1977 to June 30th 2019. There are 52,317, 43,529, and 23,634 articles retrieved from The Globe and Mail, The Toronto Star and National Post, respectively. Based on results from preliminary topic-modeling analysis using LSA and LDA, the research team performed multiple rounds of data compiling to remove stop words and meaningless words (e.g., journalists' names, physical address) prior to the analysis.

B. Measures

We use three (types of) measures to assess results from the eight topic-modeling methods pertaining to their choices of K: held-out likelihood (or reconstruction loss when applicable), coherence statistics, and graph-based dimensionality selection [30]–[33].

Fitting Error Measure: A 3-fold cross validation is used to calculate the held-out likelihood of fitted models [34]: the text corpus was divided into three parts, with one part as a testing set and the other two as training sets. For a topic-modeling method, we repeat the same estimation procedure for all three parts of the text corpus and then use the average of the held-out likelihood (or reconstruction loss) based on the three rounds of estimation as an indicator of model performance. The focus of the held-out-likelihood approach is the predictive power of a specific method (i.e., the fitness of data) instead of the coherence of the latent variables (topics) being investigated. The testing loss is used as a measure of fitness for the two neural models (DocNade and NVDM) [35]. A higher value of the held-out likelihood (or a lower value of the reconstruction/training loss) indicates better performance.

Coherence Statistics: We employ four measures of coherence in this study: C_v, C_{npmi}, C_{uci}, and U_{mass} [36]. The use of coherence measures follows the idea that a set of semantic expressions or terms is coherent if these expressions or terms agree with one another. For one specific topic, a coherence measure captures the degree of semantic similarities among words in this topic. The average of coherence statistics of each topic is used as a within-topic measure of coherence, which allows us to assess whether results from topic modeling represent actual semantic patterns or correspond to a methodological artifact. Despite their methodological connections, these coherence measures should be considered as independent to each other and we cannot directly compare values based on different coherence measures. For all four coherence measures, a higher value suggests that, on average, topics identified by a method are more coherent.

To facilitate our discussion on different coherence measures, we first recall the definition of the pointwise mutual information function [37]:

$$PMI(x, y) = \log \left(\frac{P(x, y) + \epsilon}{P(x)P(y)} \right),$$

where ϵ is a smoothing constant and is often set to 1. Next, we briefly describe these four coherence measures. C_{uci}, which was among the earliest statistics of topic coherence, uses a sliding window and pointwise mutual information to model the co-occurrence probability of every word pairs in a topic. Because C_{uci} needs to pair every single word with every other word in a topic, it can be argued that this measure provides an extrinsic rather than intrinsic evaluation of coherence [32]. We use a hypothetical topic of three words {a, b, c} to illustrate the calculation of C_{uci}. The co-occurrence probability of any pair of words in this topic is calculated based on sliding windows: if the text is "a has b", the documents obtained from a size-2 sliding window are "a has","has b". In this case, $P(a) = \frac{1}{2}$ (appeared once in the two documents obtained), $P(a, b) = 0$ (no co-occurrence of "a" and "b"), and we have C_{uci} as:

$$C_{uci} = \frac{1}{3} \left[PMI(a, b) + PMI(a, c) + PMI(b, c) \right].$$

C_{npmi} can be treated as an extension of C_{uci} given that the former uses normalized pointwise mutual information (NPMI) instead of pointwise mutual information [38]. NPMI is defined as:

$$NPMI(x, y) = \frac{\log \frac{P(x,y)+\epsilon}{P(x)P(y)}}{-\log(P(x, y) + \epsilon)},$$

where ϵ is a smoothing constant and the function is usually further weighted by raised to the power $\gamma > 0$.

C_v is a coherence measure proposed in more recent years to deal with indirect similarities between words [36], which means that some words should belong to the same topic yet they rarely occur together. Instead, researchers can learn indirect similarities through similar adjacent words. For example, if there are two sentences "McDonald makes chicken nuggets" and "KFC serves chicken nuggets", researchers will learn the indirect similarity between "McDonald" and "KFC" and put them together in the same topic. The mathematical details of C_v are somewhat complicated. Through the calculation of NPMI, a set of vectors are generated from the co-occurrence counts between every top word and every other top word. As a result, there is a corresponding vector for every top word in a topic. The indirect similarity is then calculated between the vector of every top word and the centroid of vectors of all other top words, where cosine distance is used as a measure of similarity.

Based on the principle that the occurrence of every top word should be informed by every preceding top word, the last coherence measure U_{mass} draws on the conditional probability of weaker words given the presence of their corresponding stronger words in a topic. Different from the other three measures, U_{mass} appears to be an intrinsic measure since the word list needs to be ranked and one word is only compared to its preceding and succeeding words [32], [33]. To avoid the logarithm of zero, U_{mass} uses a pairwise score function of the empirical conditional log-likelihood based on smoothing counts.

Dimensionality Selection: Graph-based dimensionality selection is also used to guide our choices of K. Methods like SVD (LSA) and PCA have a natural indicator of importance: the eigenvalue. Although scree plots have been used to select principal components, the traditional threshold of dimensionality selection, namely, the eigenvalue as 1.0, is not applicable to the high-dimensional data in this study. We thus use an automatic procedure to search for the elbow point in a scree plot via a simple profile likelihood method [31].

V. Results

We use three measures to assess the choices of K across the eight topic-modeling methods and results are presented from Figure 1 to Figure 17. For a specific topic-modeling method, we first assess whether different measures (likelihood/loss, coherence, dimensionality selection) tend to suggest similar choices of K. If multiple optimal solutions (e.g., a bimodal pattern) are suggested, we prefer a small optimal number of topics for the sake of simplicity in interpretation. In Figure 1, all the four coherence statistics favour fewer topics (see results for the SVD (LSA) method). But an opposite conclusion is suggested by both dimension selection and held-out likelihood/loss. The optimal number of topics appears to be fairly large according to results from dimensionality selection (669 topics, see Figure 3), while a larger number of topics is always preferred based on the reconstruction loss (see Figure 3).

Due to their methodological similarities, findings based on PCA are virtually the same as these based on LSA. All coherence statistics appear to suggest that fewer topics are preferred (see Figure 4). This conclusion is again different from these based on dimensionality selection and held-out likelihood. According to results from dimensionality section, the optimal number of topics should be 698 (see Figure 5). The likelihood measure also favours more topics (see Figure 6).

Because FA and other five topic-modeling methods do not explicitly consider eigenvalues, dimensionality selection is not applicable. For the FA method, these coherence statistics still prefer fewer topics (see Figure 7); the likelihood measure suggests that K should be around 100 (see Figure 8).

Different conclusions are suggested by the coherence statistics for the NMF method (see Figure 9). While the curves associated with C_{npmi} and C_v are flat, different results are suggested by the C_{uci} and U_{mass} measures: U_{mass} prefers a small K but C_{uci} suggests that K should be somewhere around 50. According to Figure 10, evidence from the held-out error suggests that a larger K is associated with better goodness-of-fit.

When we apply coherence statistics to assess results from the LDA method, we do not observe a clear pattern for the curves of C_{npmi} and C_v (see Figure 11). C_{uci} suggests that K should be between 50 and 75 but U_{mass} favours a smaller K. According to the held-out likelihood (see Figure 12), the optimal number of topics should be around 20.

We observe similar patterns for the two neural models (DocNade and NVDM). Coherence statistics presented in Figure 13 and Figure 15, especially the curve of C_{uci}, suggest that the optimal number of topics should be 50 (or above). Yet, the training loss declines with a larger number of topics (see Figure 14 and Figure 16). Coherence statistics tend to suggest a small optimal number of topics when the HDP method is employed (see Figure 17). The likelihood measure was not applicable to the assessment of results from HDP because, in theory, the number of topics is not a model parameter of HDP and the method has explored various choices of K. If we sort elements of the trained super-parameter α and apply the dimensionality-selection method to these ordered elements, the optimal solution to K is 2 (results omitted).

After we discuss results from one specific topic-modeling method based on different measures of optimality, our discussion above suggests that the same method do not necessarily produce similar optimal numbers of topics. The diverse results are particularly striking for classic data-reduction methods (SVD, PCA, FA, and NMF). In contrast, optimal numbers of topics reported by statistical generative models (LDA and HDP) tend to be similar according to different measures of optimality. Results from neural models (DocNade and NVDM) appear to suggest a tradeoff between topic coherence and goodness-of-fit. Next, by compiling results from eight topic-modeling methods (see Table I), we investigate whether the optimal numbers of topics specified by these different methods would agree with each other. As expected, when two approaches to topic modeling are methodologically related to each other, their choices of K are also similar regardless of the specific measure used. No matter whether we use coherence statistics, or likelihood/loss, or dimensionality selection to assess the results, SVD and PCA produce similar optimal numbers of topics. This conclusion also holds for results obtained from DocNade and NVDM.

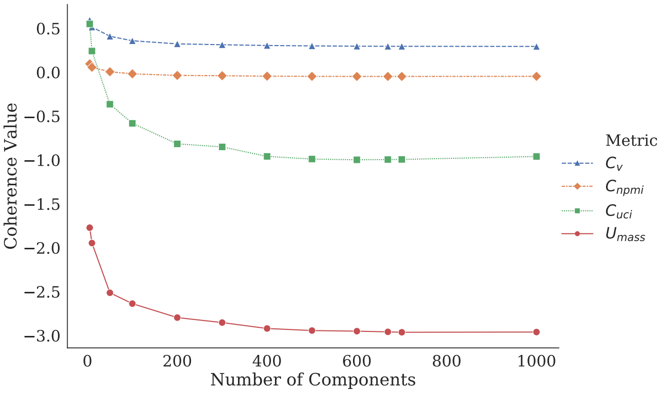

Fig. 1. The SVD (LSA) method: Coherence.

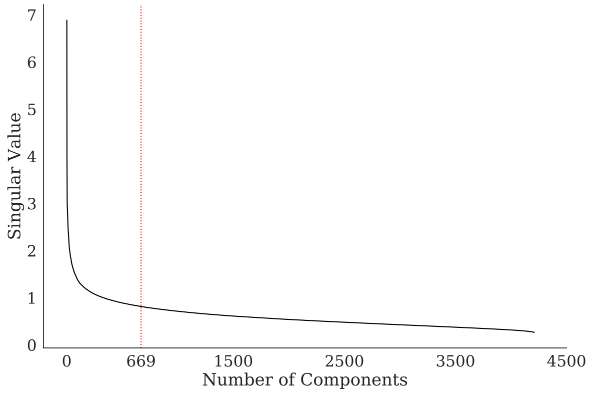

Fig. 2. The SVD (LSA) method: Dimensionality selection.

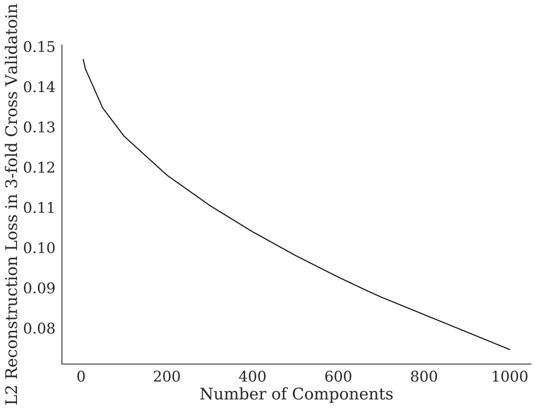

Fig. 3. The SVD (LSA) method: Held-out error.

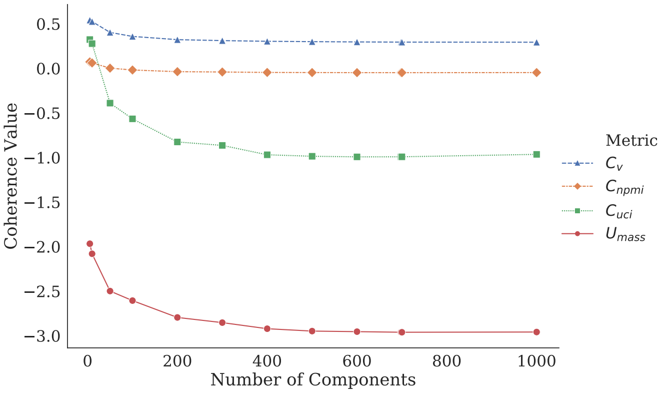

Fig. 4. The PCA method: Coherence.

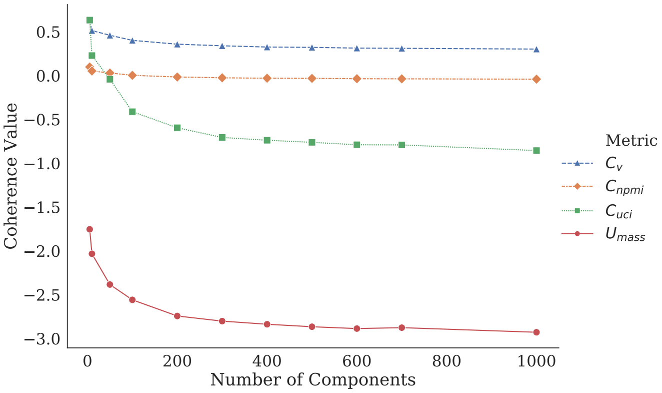

Fig. 7. The FA method: Coherence.

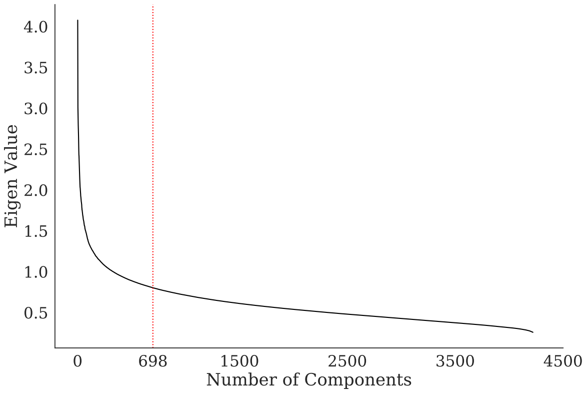

Fig. 5. The PCA method: Dimensionality selection.

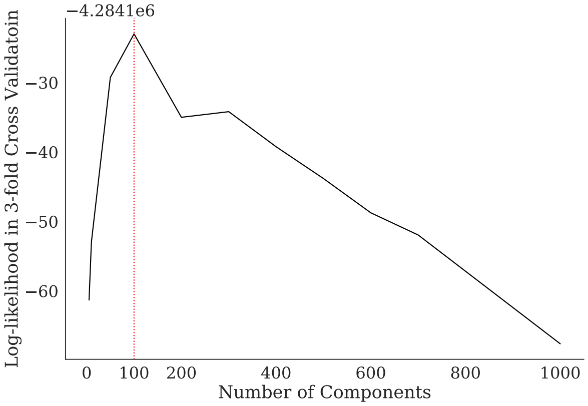

Fig. 8. The FA method: held-out likelihood.

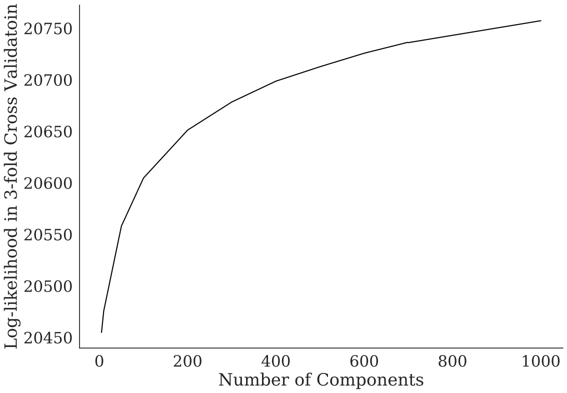

Fig. 6. The PCA method: held-out likelihood.

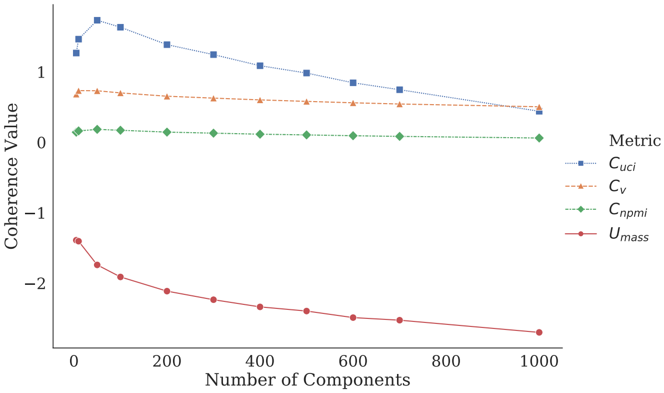

Fig. 9. The NMF method: Coherence.

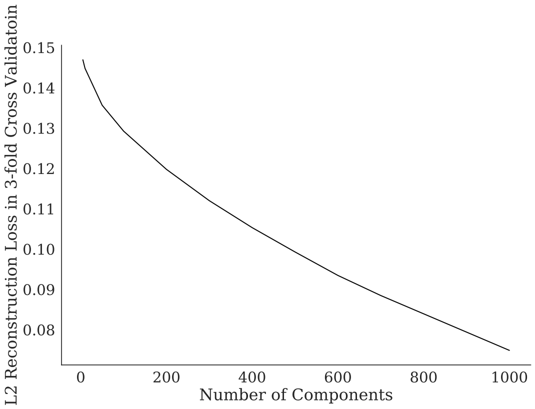

Fig. 10. The NMF method: Held-out error.

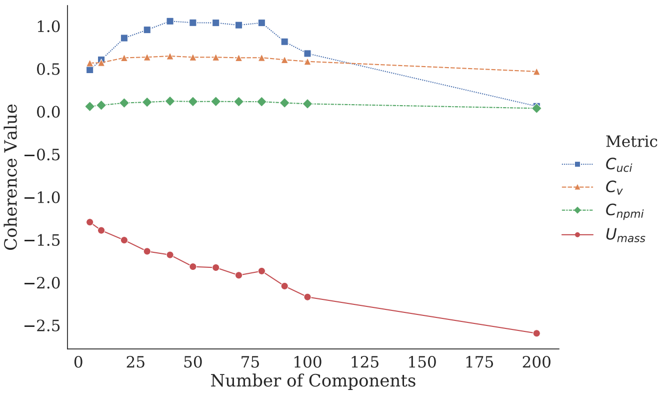

Fig. 11. The LDA method: Coherence.

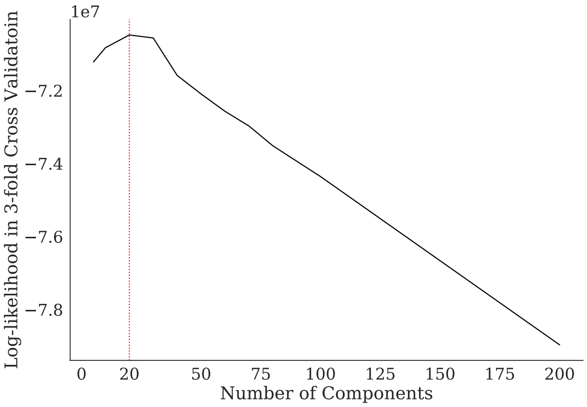

Fig. 12. The LDA method: Held-out likelihood.

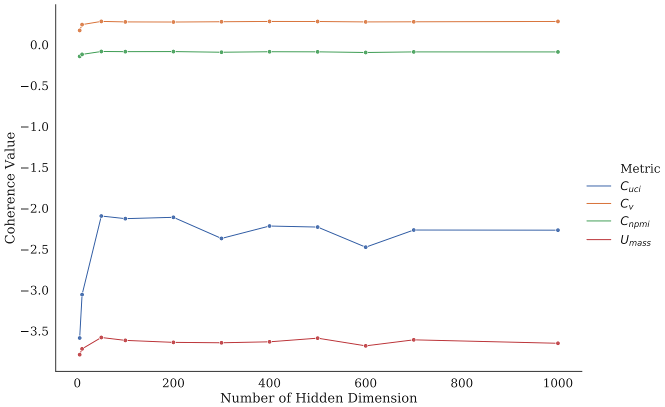

Fig. 13. The DocNade method: Coherence.

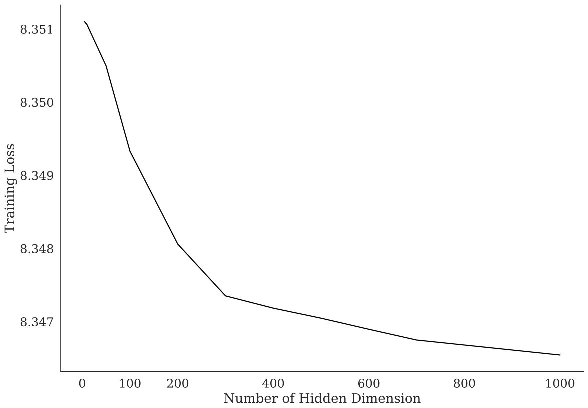

Fig. 14. The DocNade method: Training loss.

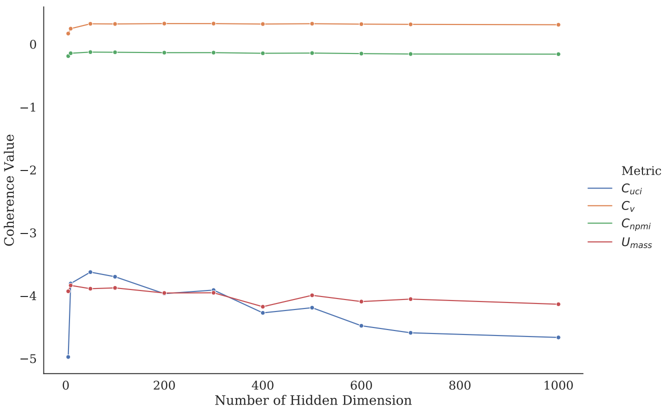

Fig. 15. The NVDM method: Coherence.

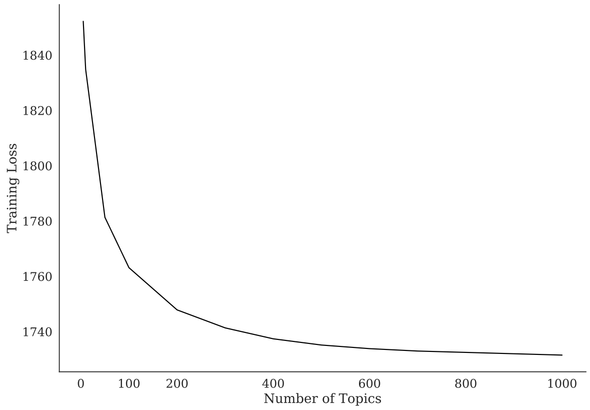

Fig. 16. The NVDM method: Training loss.

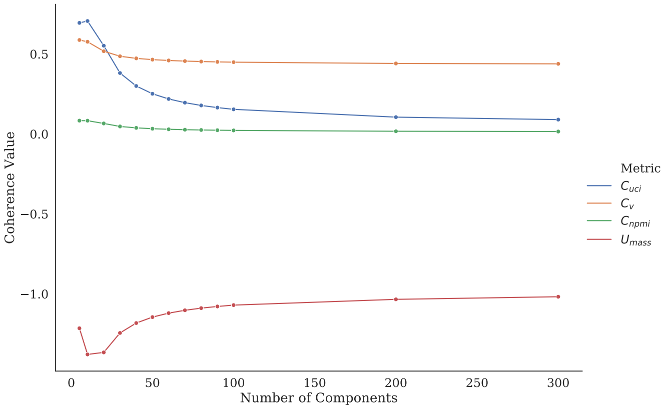

Fig. 17. The HDP method: Coherence.

TABLE I
A summary of optimal numbers of topics, by methods and measures

	SVD	PCA	FA	NMF
C_{uci}	small	small	small	50±
C_v	small	small	small	10-50*
C_{npmi}	small	small	small	50±*
U_{mass}	small	small	small	small
Held-out likelihood (or loss)	large	large	100	large
Dimensionality Selection	669	698	NA	NA
	LDA	DocNade	NVDM	HDP
C_{uci}	50-75	50-200	50±	10±
C_v	25-75*	50+*	50+*	small
C_{npmi}	25-75*	50+*	50+*	small*
U_{mass}	small	50-500*	10-300*	small
Held-out likelihood (or loss)	20	large	large	NA

Note: *Optimal choices are not very clear.

VI. Conclusion

Based on about 120,000 Canadian newspaper articles, this study uses three measures of optimality (coherence statistics, held-out likelihood/loss, and dimensionality selection) to assess the performance of eight approaches to topic modeling. For their choices of optimal numbers of topics, results from different approaches to topic modeling often do not agree with one another, even if the same measure of optimality is used to assess the choice of K. Yet, a variety of methodologically related approaches (e.g., SVD and PCA, DocNade and NVDM, LDA and HDP) do suggest similar choices of K, especially when the same measure of optimality is used. Statistical generative models including LDA and HDP report similar optimal numbers of topics under different measures of optimality and may be preferred over others. Finally, it should be noted that these findings are based on one text corpus of Canadian newspaper articles and may vary with different data sources.

To put our findings in perspective, we argue that these eight methods contribute to the methodological repertoire of topic modeling due to their shared purpose, rather than their methodological similarities. By reviewing their methodological details, we show that these eight methods employ a wide range of modeling philosophies to leverage semantic information and attributes. Optimality in the topic-modeling setting can also be defined in different ways. Depending on whether the fundamental goal is to have coherent topics or to achieve better goodness-of-fit, the choice of K could be drastically different. In practice, researchers' choices of K need to balance diverse optimality criteria, and should be informed by knowledge from domain experts. Based on the premise that "all models are wrong, but some are useful" [39], researchers need to identify a useful lens though which the rich information embedded in texts can be exploited, analysed, and interpreted [1]. A topic-modeling method is useful as long as it enhances our understanding of society.

Acknowledgement

The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. PolyU 15334616].

References

[1] P. DiMaggio, M. Nag, and D. Blei, "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding," Poetics, vol. 41, no. 6, pp. 570–606, 2013.

[2] M. E. Martin and N. Schuurman, "Area-based topic modeling and visualization of social media for qualitative GIS," Annals of the American Association of Geographers, vol. 107, no. 5, pp. 1028–1039, 2017.

[3] D. A. McFarland, D. Ramage, J. Chuang, J. Heer, C. D. Manning, and D. Jurafsky, "Differentiating language usage through topic models," Poetics, vol. 41, no. 6, pp. 607–625, 2013.

[4] M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand, "Structural topic models for open-ended survey responses," American Journal of Political Science, vol. 58, no. 4, pp. 1064–1082, 2014.

[5] Q. Fu, Y. Zhuang, J. Gu, Y. Zhu, H. Qin, and X. Guo, "Search for k: Assessing five topic-modeling approaches to 120,000 canadian articles," in 2019 IEEE International Conference on Big Data (Big Data), pp. 3640–3647, IEEE, 2019.

[6] D. Lazer and J. Radford, "Data ex machina: introduction to big data," Annual Review of Sociology, vol. 43, pp. 19–39, 2017.

[7] P. DiMaggio, "Adapting computational text analysis to social science (and vice versa)," Big Data & Society, vol. 2, no. 2, p. 2053951715602908, 2015.

[8] E. Loper and S. Bird, "NLTK: The natural language toolkit," in Association for Computational Linguistics, 2004.

[9] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," Text Mining: Applications and Theory, vol. 1, pp. 1–20, 2010.

[10] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," Discourse Processes, vol. 25, no. 2-3, pp. 259–284, 1998.

[11] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," Journal of Documentation, vol. 60, no. 5, pp. 503–520, 2004.

[12] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information processing & management, vol. 24, no. 5, pp. 513–523, 1988.

[13] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American Society for Information Science, vol. 41, no. 6, pp. 391–407, 1990.

[14] J. Gao and J. Zhang, "Clustered SVD strategies in latent semantic indexing," Information Processing & Management, vol. 41, no. 5, pp. 1051–1063, 2005.

[15] G. Strang, Introduction to Linear Algebra, vol. 3. Wellesley Cambridge Press, 1993.

[16] I. Jolliffe, Principal Component Analysis. Springer, 2011.

[17] Z. Ghahramani, G. E. Hinton, et al., "The EM algorithm for mixtures of factor analyzers," tech. rep., Technical Report CRG-TR-96-1, University of Toronto, 1996.

[18] D. B. Rubin and D. T. Thayer, "EM algorithms for ML factor analysis," Psychometrika, vol. 47, no. 1, pp. 69–76, 1982.

[19] N. Péladeau and E. Davoodi, "Comparison of latent Dirichlet modeling and factor analysis for topic extraction: A lesson of history," in Proceedings of the 51st Hawaii International Conference on System Sciences, 2018.

[20] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, no. 6755, p. 788, 1999.

[21] E. Gaussier and C. Goutte, "Relation between PLSA and NMF and implications," in Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 601–602, ACM, 2005.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, vol. 3, no. Jan, pp. 993–1022, 2003.

[23] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent Dirichlet allocation," in Advances in Neural Information Processing Systems, pp. 856–864, 2010.

[24] R. R. Salakhutdinov and G. E. Hinton, "Replicated softmax: an undirected topic model," in Advances in neural information processing systems, pp. 1607–1614, 2009.

[25] H. Larochelle and I. Murray, "The neural autoregressive distribution estimator," in Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pp. 29–37, 2011.

[26] H. Larochelle and S. Lauly, "A neural autoregressive topic model," in Advances in Neural Information Processing Systems, pp. 2708–2716, 2012.

[27] F. Morin and Y. Bengio, "Hierarchical probabilistic neural network language model.," in Aistats, vol. 5, pp. 246–252, Citeseer, 2005.

[28] Y. Miao, L. Yu, and P. Blunsom, "Neural variational inference for text processing," in International conference on machine learning, pp. 1727–1736, 2016.

[29] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," arXiv preprint arXiv:1511.06349, 2015.

[30] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in Advances in Neural Information Processing Systems, pp. 288–296, 2009.

[31] M. Zhu and A. Ghodsi, "Automatic dimensionality selection from the scree plot via the use of profile likelihood," Computational Statistics & Data Analysis, vol. 51, no. 2, pp. 918–930, 2006.

[32] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108, Association for Computational Linguistics, 2010.

[33] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 262–272, Association for Computational Linguistics, 2011.

[34] S. Arlot, A. Celisse, et al., "A survey of cross-validation procedures for model selection," Statistics Surveys, vol. 4, pp. 40–79, 2010.

[35] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," arXiv preprint arXiv:1702.05659, 2017.

[36] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in Proceedings of the 8th ACM International Conference on Web Search and Data Mining, pp. 399–408, ACM, 2015.

[37] A. Islam and D. Inkpen, "Second order co-occurrence PMI for determining the semantic similarity of words," in Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), pp. 1033–1038, 2006.

[38] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013), pp. 13–22, 2013.

[39] G. E. Box, "All models are wrong, but some are useful," Robustness in Statistics, vol. 202, p. 549, 1979.