COMPARISON OF METHODS FOR CHOOSING AN APPROPRIATE NUMBER OF TOPICS IN AN LDA MODEL

by

Patricia Jean Goedecke

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of Master of Science

Major: Mathematical Sciences

The University of Memphis

August 2017

Acknowledgements

I would like to thank Dr. Dale Bowman for encouraging me throughout my graduate studies at the University of Memphis, that I could succeed in this pursuit. I would like to thank Dr. Vasile Rus and Dr. Art Graesser for providing me with opportunities to apply my developing statistical skills, with internship and assistantship work in the Institute for Intelligent Systems. I would like to thank my classmates for keeping me going through our studies together, my friends and family for having my back, and my parents for always believing in me. I thank my colleagues at the University of Tennessee for encouraging me to continue my academic pursuits, and thesis committee members Dr. Su Chen and Dr Śaunak Sen for lending their expertise to this endeavor.

I would also like to thank Mr. Quentin Pleplé for his online bibliography of Topic Modeling (qpleple.com/bib), which helped make this subject approachable to me.

Abstract

Topic modeling is a technique for reducing dimensionality of large corpuses of text. Latent Dirichlet allocation (LDA), the most prevalent form of topic modeling, improved upon earlier methods by introducing Bayesian iterative updates, providing a sound theoretical basis for modeling by iteration. Yet a piece of the modeling puzzle remains unsolved; the number of topics to model, K, is an as yet unanswered question. This number of topics may also be called the dimensionality of the model. With this is an integrally related puzzle; how to determine when a model has been best fit. Presented here are a brief history of the development of topic modeling from its inception preceding LDA to the present; and a comparison of methods for determining what is a best-fit topic model, in pursuit of the most appropriate K.

Contents

List of Figures

Introduction

Topic modeling is a technique for reducing dimensionality of large corpuses of text. Blei et al. (2003) provided it with a method that has brought topic modeling to the forefront of machine learning techniques for text analysis. Bayesian latent Dirichlet allocation (LDA) modeling allows a topic model to continually improve with each iteration while effectively utilizing prior knowledge.

1 Methods leading to Latent Dirichlet Allocation

1.1 Latent Semantic Analysis and singular value decomposition

Topic modeling was arguably begun by Deerwester et al. (1990) in an article introducing Latent Semantic Analysis (LSA) as an indexing method. Indexing methods originated in information retrieval with text data, developing criteria for measuring similarity between words and documents. The authors sought to overcome a limitation of earlier indexing methods, namely that search terms might not precisely match the vocabulary used in a document. Their insight was that an underlying probabilistic relationship between terms and documents could be imputed and then described using statistical methods. Their approach uses singular value decomposition (SVD).

In the SVD approach of Deerwester et al. (1990), both terms and documents are represented as vectors in a space of (some chosen value) K dimensions; similarity is measured by the dot product or cosine between points in the space. Any term can be represented in this multidimensional space, with synonyms presumably occupying nearly the same space. A search term will now retrieve not only the identical term but terms with similar meaning, and also terms used in a similar context. A polysemic term may occupy two or more locations in this vector space, each accompanied by terms of the appropriate context. For example, “tee” as a shirt may collocate with terms such as “shirt,” “clothing,” and “casual wear,” while “tee” as a golf implementation may collocate near “putter” and “wood.” This collocating space is considered the topic.

These vectors comprise a large term by document matrix which is then decomposed into a set of orthogonal factors, from which the original matrix can be approximated by linear combination. The overall matrix is factored into two matrices, one representing the documents, and the other the topics. Any document now can be represented as a linear combination of topics. Landauer and Dumais (1997) expanded on the LSA concept with a psychological study of human understanding of meaning; they described LSA as “a new general theory of acquired similarity and knowledge representation,” and used it to simulate learning acheived by humans.

1.2 Probabilistic Latent Semantic Analysis

Hofmann (1999) introduced Probabilistic Latent Semantic Analysis (pLSA), a step toward latent Dirichlet allocation. Probabilistic Latent Semantic Analysis imagines an automated document generating process. For each word position i in document d, a topic z(d(i)) is first drawn probabilistically from a discrete set of topics, with probability θ(d). Given the topic z(d(i)), a word is then selected from the list associated with that topic. The words are probabilistically associated with each topic, with probability ϕ(z(d(i))). Each document is then represented as a linear combination of topics, and pLSA is likewise a linear topic model based on the factorization of the document-word matrix.

In pLSA, each document is a probabilistic mixture of topics; topics are modeled as occurring within documents according to a multinomial probability distribution. Blei et al. (2003) criticized pLSA for the following: 1) having ill-defined document generative semantics, and 2) having a number of parameters which increases linearly with the number of training documents, making the model susceptible to overfitting.

1.3 Non-Negative Matrix Factorization

In the same year as pLSA, Lee and Seung (1999) introduced Non-Negative Matrix Factorization (NMF), described as a neural network. NMF is contrasted with principle components analysis and vector quantization as representing parts-based rather than holistic representations. These constraints lead to a parts-based representation because they allow only additive, not subtractive, combinations. NMF is applied in representing facial images as well as semantic features of texts. As the authors explain, NMF uses an iterative algorithm, starting from non-negative initial conditions. When used with text data, a set of H hidden variables are iterated in tandem with a column of W word frequencies forming matrix V, constructing approximate factorizations.

Ding et al. (2008) showed NMF to be equivalent to pLSA, in that both optimize the same objective function (Pleple, 2013a):

$$F = \textstyle\sum_{i=1}^{n} \sum_{\mu=1}^{m} [V_{i\mu} log(WH)_{i\mu} - (WH)_{i\mu}]$$

1.4 Latent Dirichlet Allocation and its reception

The contribution of Blei et al. (2003), distinguishing Latent Dirichlet Allocation (LDA) from pLSA, was adding the Bayesian element of prior distributions on parameters. As with pLSA, Blei et al. imagine documents generated probabilistically from topics with multinomial probabilities, where a prior distribution is placed on the parameter of the multinomial distribution. Every iteration updating the model, though, now improves on the prior model in a theoretically consistent manner, based on Bayesian prior and posterior distributions.

The LDA methodology of Blei et al. (2003) has been contested (Chang et al. 2009, Wallach et al. 2009, Stevens 2012) and perhaps improved upon. Mimno et al. (2011) provide three contributions to improve upon LDA modeling:

> (1) An analysis of the ways in which topics can be flawed; (2) an

automated evaluation metric for identifying such topics that does not rely on human annotators or reference collections outside the training data; (3) a novel statistical topic model based on this metric that significantly improves topic quality in a large-scale document collection.

Nevertheless, LDA is the current standard in topic modeling. Topic modeling is gaining recognition but is still under scrutiny as an effective data mining tool. Teh et al. (2004, 2012) propose a hierarchical Dirichlet process which they find to outperform the simple LDA model. Wei and Croft (2006) compare LDA with information retrieval using cluster-based models and find LDA compares favorably. AlSumait el al. (2009) examine the topic significance ranking of LDA models. Musat et al. (2011) propose improving topic evaluation using conceptual knowledge. Niraula et al. (2013) compare LDA with LSA, finding LSA to be more effective regarding semantic similarity. Banjade et al. (2015) seek to combine a variety of methods for measuring word-to-word similarity. Among topic modeling researchers, rather than replacing LDA, researchers are now seeking ways to use LDA more effectively. In addition, the method is being applied to different types of big data, such as biomedical data sets (Pritchard et al., 2000, Falush et al., 2003, Zhao et al., 2014, Zhao et al., 2015).

2 Current trends in LDA topic modeling

While LDA is now the accepted standard in topic modeling, and topic modeling is gaining popularity as a dimension reduction technique and clustering tool, the method may still be considered a work in progress. Two open areas of investigation are topic coherence and selecting the dimensionality of a model. Also of interest is the expansion of topic modeling into broader types of data mining.

2.1 Expanding topic modeling applications

Topic modeling developed from a pursuit for effective information retrieval, in an age of ever-increasing data. Even prior to Blei et al.'s (2003) contribution of latent Dirichlet allocation to the method, Lee and Seung (1999) had applied non-negative matrix factorization to image data as well as text. In their studies, Gerber et al. (2007) and Zhao et al. (2014) effectively apply topic modeling to human gene expression data. Mimno et al. (2011) anticipate that the methods of topic modeling will need to incorporate ever-larger corpora of data.

2.2 Topic coherence

A troubling concern regarding topic modeling as a dimension reduction technique for data is that topics modeled may make little sense to human interpreters, or may appear flawed. A number of researchers have sought to evaluate or improve on the coherence of topics modeled.

Newman et al. (2010) tested a number of automated techniques for evaluating topic coherence, including pointwise mutual information (PMI) and other lexical relatedness measures based on WordNet, Wikipedia and the Google search engine. They found several Wikipedia-based measures to show strong results, including one using PMI from Wikipedia source data. PMI measures the relatedness between any two words as the discrepancy between the relative frequency of their co-occurence $p(x, y)$ and the relative frequency predicted by their joint probabilities when independence is assumed $p(x)p(y)$.

Mimno et al. (2011) addressed the concern that topics derived from the automated LDA method frequently baffle human interpreters. They identified several specific types of topic flaws, presented a method for evaluating topic coherence automatedly, and provided a new statistical topic model which they call a Generalized Pólya Urn model. In their discussion, Mimno et al. (2011) express the concerns that going forward, improving semantic cohesion will be of primary concern to topic modeling researchers, along with scaling to ever larger data sets.

2.3 Selecting model dimensionality K

Another current issue in topic modeling, a question which remains unanswered, is how to mathematically determine the most appropriate K, the dimensionality of a model. This issue will be addressed in section 3.

3 Topic modeling using latent Dirichlet allocation

Topic modeling using latent Dirichlet allocation (LDA) rests on a number of mathematical propositions and assumptions. Among these are exchangeability, de Finetti's theorem and hierarchical Bayesian modeling.

3.1 Exchangeability and de Finetti's theorem

Exchangeability is a foundation of the latent Dirichlet allocation model. A sequence of random variables may be said to be exchangeable if any permutation of the sequence has the same joint probability; that is, for permutation $\pi_1, ..., \pi_n$ of the integers 1,...,n

$$\boldsymbol{p(i_1, ..., i_n) = p(z_{\pi(1)}, ..., z_{\pi(n)})}$$

In topic modeling, each document is treated as a "bag of words," i.e., the words are exchangeable within the document; word order is not taken into account. Each word is said to appear in a document according to a multinomial probability distribution, with a distribution parameter vector specific to a latent topic.

Exchangeable random variables in a sequence are identically distributed, but may not be independent. Independent, identically distributed variables have a number of known properties that are useful in modeling. Exchangeable random variables may be considered conditionally independent, conditioned on some latent variable. In topic modeling, words are said to appear in a document independently, according to a multinomial distribution conditioned on a topic. The joint probability of a set of words and topics therefore has the form:

$$p(\boldsymbol{w}, \boldsymbol{z}) = \int p(\theta) \left(\prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n) \right) d\theta$$

where θ is the random parameter of a multinomial over topics (Blei et al., 2003).

3.2 Latent Dirichlet allocation: Hypothetical document generation

Latent Dirichlet allocation assumes a probabilistic stepwise procedure of generating a document, given a set of V vocabulary words and K topics. The following procedure is adapted from Blei et al. (2003) and Bowman (in press).

1. The number of words (N) for the document is selected from a Poisson distribution with parameter $\boldsymbol{\xi}$.
2. Multinomial probability vector $\boldsymbol{\theta}$ of length K is selected from a Dirichlet distribution with parameter vector $\boldsymbol{\alpha}$.
3. The nth word (w_n) in a document is selected using the following process:
   (a) Select topic Z_n from the multinomial distribution with parameters (1, $\boldsymbol{\theta}$). In this model, there is a probability distribution over all vocabulary words for each topic.
   (b) Select vocabulary probability vector $\boldsymbol{\phi_{z_n}}$ of length V for topic Z_n from a Dirichlet distribution with parameter $\boldsymbol{\beta_{Z_n}}$. Vector $\boldsymbol{\phi_{z_n}}$ is dependent on the chosen topic, Z_n.
   (c) Select word w_n from probability distribution $p(w_n|Z_n,\boldsymbol{\phi_{Z_n}})$, a multinomial distribution with parameters (1, $\boldsymbol{\phi_{Z_n}}$); where $\boldsymbol{\phi_{Z_n}}$ is a vector of length V, the number of words in the vocabulary.

Let $w_{nj} = 1$ when the nth word in the document is the jth word in the vocabulary, and zero otherwise. When the nth word in the document is in topic

i, $z_{ni} = 1$, and zero otherwise. Then $\phi_{ij} \in \boldsymbol{\phi_{z_n}} = \text{Prob}(w_{nj} = 1|z_{ni} = 1)$. As only one word occurs at each location in a document,

$$\sum_{i=1}^{K} Z_{ni} = \sum_{j=1}^{V} w_{nj} = 1.$$

The LDA model can then be described as follows. For each document,the vector $\boldsymbol{Z_n} = (Z_{n1}, ..., Z_{nK})$ indicates the topic of the nth word. The model assumes a multinomial distribution for $\boldsymbol{Z_n}$,

$$\boldsymbol{Z_n} \sim Mult(1, \theta_1, ..., \theta_K) \text{ for } n = 1, ..., \text{N}.$$

Then the conditional probability distribution of the Z_n vector, given the $\boldsymbol{\theta}$ vector and K, is a product of theta values, where $\sum_{i=1}^{K} Z_{ni} = 1$ and $\sum_{i=1}^{K} \theta_i = 1$,

$$\text{f}(\boldsymbol{z_n}|\theta, K) = \prod_{i=1}^{K} \theta_i^{z_{ni}}$$

The $\boldsymbol{Z_n}$ vector is a vector of a single 1, and (K-1) zeroes. $Z_{ni} = 1$ when the nth word is from topic i; θ_i is interpreted as the probability topic i is in the document.

Let $\boldsymbol{Z} = (\boldsymbol{Z_1, ..., Z_N})$. Then due to the assumption of conditional independence of topics given $\boldsymbol{\theta}$, the likelihood of $\boldsymbol{Z}$ given $\boldsymbol{\theta}$ is a product of thetas over topics and over the number of words in the document, N,

$$\boldsymbol{L(Z|\theta}, K) = \prod_{n=1}^{N} \prod_{i=1}^{K} \theta_i^{z_{ni}}$$

The vector representing the nth word in a document, $\boldsymbol{w_n} = (w_{n1}, ..., w_{nV})$, is also a vector of zeroes with a single one.

$$w_{nj} = \begin{cases} 1 & \text{if } n\text{th word in document } d \text{ is the } j\text{th word in the vocabulary} \\ 0 & \text{otherwise} \end{cases}$$

The conditional distribution of $\boldsymbol{w_n}|\boldsymbol{Z_n}$ is multivariate, $\boldsymbol{w_n} \sim mult(1, \phi_{1zn}, ..., \phi_{Vzn})$. From this we have

$$\text{f}(\boldsymbol{w_n}|\boldsymbol{Z_n}, \boldsymbol{\phi_n}, K) = \prod_{i=1}^{K} \prod_{j=1}^{V} (\phi_{ji}^{z_{ni}})^{w_{nj}}$$

Then the likelihood of $\boldsymbol{w}|\boldsymbol{Z},\boldsymbol{\phi}$,$K$, where $\boldsymbol{w} = \boldsymbol{(w_1, ..., w_N)}$, is found as

$$\boldsymbol{L(w|Z,\phi,}K) = \prod_{n=1}^{N} \prod_{i=1}^{K} \prod_{j=1}^{V} (\phi_{ji})^{z_{ni}w_{nj}}$$

Here, ϕ_{ji} may be interpreted as the probability given topic allocation i, that word j is in the document, where j = 1,...,V and i = 1,...,K.

A Dirichlet($\boldsymbol{\alpha}$) prior distribution is assumed on $\boldsymbol{\theta}$; a Dirichlet($\boldsymbol{\beta_j}$) prior distribution is assumed for $\boldsymbol{\phi}$, with j = 1,...,K; i.e., there is a unique β for each topic. These priors are proportional to the following:

$$\pi(\boldsymbol{\theta}|\alpha_0, K = k) \propto \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}$$
$$\pi(\boldsymbol{\phi}\ |\beta_{0j}, K = k) \propto \prod_{i=1}^{K} \prod_{j=1}^{V} \phi_{ji}^{\beta_{ji}-1}$$

Also described by Blei et al. (2003), the dimensionality (K) of the Dirichlet distribution, which is the number of topics modeled, is assumed to be known and fixed. In practical application, dimensionality (K) is determined arbitrarily or post hoc. Estimation of parameters may be accomplished in a theoretically consistent manner using Bayesian methods. These include Gibb's sampling - a Markov Chain Monte Carlo (MCMC) algorithm - and a variational expectation-maximization (EM) algorithm (Blei et al., 2003).

3.3 Selecting model dimensionality

As noted above, a question in topic modeling which remains unanswered is how to mathematically determine the most appropriate K, the dimensionality of a model. Thus far, dimensionality has often been determined in an ad hoc or post hoc manner. A common technique to find the most appropriate K is to fit an LDA model over a broad set of K values, comparing results at each level. We see this for example in Heinrich et al. (2005), comparing correlation levels at different *scopes*, where scope refers to levels of K as well as to numbers of documents, term types, and document types. Some studies that have addressed the issue are described below. Application of several of these methods will be demonstrated in section 4. As we will see, the definition of a best-fitting topic model remains to be determined.

3.3.1 Perplexity (P)

Blei et al. (2003) suggested a method for measuring the fit of a topic model. They used *perplexity*, which had already been in common usage with language modeling methods such as Latent Semantic Analysis (LSA). Blei et al. (2003) provide the following definitions:

1. A *word* is a discrete unit of data, defined as an item from a vocabulary indexed numerically.

2. A *document* is a collection of N words, $\boldsymbol{w} = (w_1, ..., w_N)$. As described above, w_n represents the nth word in a document.

3. A *corpus* is a collection of M documents, denoted by $D = \boldsymbol{(w_1, ..., w_M)}$, with each of these representing a document-length vector of words. Typically, a model is trained on a *training* portion of the data and tested on a *test* set to check the model fit.

With these denotations in mind, *perplexity* is then defined as

$$perplexity(D_{test}) = \exp\left(-\frac{\sum_{d=1}^{M_t} log\ p(\boldsymbol{w_d})}{\sum_{d=1}^{M_t} N_d}\right)$$

where w_d is the set of words in the dth document; D_{test} is the test set of documents not used to train the model; M_t is the number of documents in the test set; and $\mathrm{P}(\boldsymbol{w_d})$ is the probability of observing $\boldsymbol{w_d}$ using the parameters estimated from the training document set.

The meaning of perplexity is in its relationship to likelihood. As Blei et al. define it, perplexity is algebraicly equivalent to the inverse of the geometric mean per-word likelihood. This means that the greater the likelihood of the model, on an aggregated per-word level, the lower the perplexity score will be. Lower perplexity scores thus indicate better performance of the topic model.

Perplexity is an output of most LDA computational packages - whether Mallet (McCallum, 2002) or another tool - as it is an intrinsic measure of the model. No extrinsic modeling needs to be done to get these measures. A weakness of the perplexity optimization method, particularly for text data, is that it may not correlate well with human judgment regarding topics (Chang et al., 2009). Topic models selected by perplexity scores tend to rate lower in human coherence than other models. Perplexity score selections are also described as lacking in stability and efficiency (Zhao et al., 2015).

A challenge that remains, with perplexity and most other measures of optimization, is in interpreting the meaning of the measures. Perplexity falls quickly with increasing K but eventually starts increasing, perhaps due to overfitting. Smaller values of K have particular value in some cases. In text data, ten or twenty topics can generally be understood meaningfully by humans. At larger values, the intuitive meaning of topics becomes lost. Not all topic modeling is of text data, however, and intuitive interpretation is not always the goal.

3.3.2 Change in Perplexity (dP)

Zhao et al. (2015) argue that the change in perplexity is a stable and efficient means for finding an appropriate number of topics, from a machine learning standpoint. As perplexity score (P) changes from one level of K (K_i) to the next (K_{i+1}), change in perplexity is measured as:

$$dP = \frac{P_{i+1} - P_i}{K_{i+1} - K_i}$$

Zhao et al. (2015) use the absolute value of change in perplexity. Their proposed method is to select the K with the first change in direction of the dP measure, as the appropriate model dimensionality. This method also corresponds with finding a second difference of P at zero, as demonstrated in the next section. Zhao et al. (2015) measure the stability and effectiveness of the rate of perplexity change using the following methods. To evaluate stability, dP is compared with the perplexity-based method. In repeated sampling, change in perplexity provided more stable results than did perplexity. To evaluate efficiency, Zhao et al. (2015) used cluster analysis on model output. Again, the dP method outperformed the perplexity-based method for selecting the best number of topics to use.

3.3.3 Goodness Of Fit (GOF)

Goodness Of Fit is a method which compares the assignment of topics in documents with a uniform distribution which would occur if they were randomly assigned. Bowman (2016) applies this method and describe its theoretical underpinnings. The goodness of fit method applies a hypothesis testing approach. Here, θ_i^d represents that probability that topic i is included in document d. Z_i^d is the number of words in document d that are from topic i. The null hypothesis is that for $i, l = 1, ..., K i \neq l$ and $d = 1, ..., D$,

$$H_0 : \theta_i^{(d)} = \theta_l^{(d)} = 1/K$$

The application of this method involves measuring what proportion of documents, for any level of K, significantly differ from the uniform distribution. A strength of this method is that it is a purely mathematical approach, requiring little in the way of interpretive skill. A weakness is that, as it tends to favor larger values of K, it may not lead to topics which have intuitive interpretability. As noted above, smaller numbers of topics modeled tend to lead toward greater interpretability by humans.

3.3.4 Pointwise Mutual Information (PMI)

Pointwise Mutual Information (PMI) is a statistical measure of the discrepancy between observed coincidence of random variables X and Y given their joint probability, with the coincidence which would be expected assuming independence. Mathematically,

$$\text{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

Meaningfully, PMI is used to gauge the degree of topic coherence in a model. The dominant words within each topic, those with highest probability of occurring within the topic, are rated pairwise for PMI with one another. The scoring is standardized to adjust for the dimensionality of a model. Higher PMI scores are interpreted as an indication of greater topic coherence.

The Pointwise Mutual Information approach to optimizing K requires a sophisticated analytical tool extrinsic to the topic model. As applied by Niraula, Banjade, Ştefănescu, and Rus (2013), this involved forming

> ... all possible pairs with the top 10 or 20 words in each topic... [The corpus] contained 1,284,156,826 tokens and 5,693,208 word types (i.e. unique words) counted after removing digits and punctuation and changing to lower case. After removing the stop words, the number of tokens was 672,542,579.

The major strength of the PMI approach for finding K is that for text documents, the resulting model correlates well with human judgment. A challenge of this approach is that it requires the development of an extensive PMI comparison database, for text or any chosen topic medium.

4 Topic Modeling Example

The data set used to compare K optimization methods here originated with Dr. James Pennebaker of the University of Texas at Austin. Dr. Pennebaker shared a data set that included two reading assignments for students in an introductory undergraduate psychology course, and web-interfaced chats students had discussing these readings. This seed data of 2387 words was expanded by finding related documents in a Wikipedia search. The resulting corpus was 808 documents averaging 97 words per document. These documents were combined and the corpus repeatedly sampled, dividing it by 50 percent into training and testing sets. The vocabulary of the entire corpus was 78,266 unique terms; the testing and training sets each had 364,387 tokens and were broken into 37 documents of nearly 10,000 words each.

Latent Dirichlet allocation topic models were applied to the data set using Mallet (McCallum, 2002), at K levels from 10 to 100 by tens. To the resulting models were applied the K optimization methods of perplexity (P), change of perplexity (dP), Goodness Of Fit (GOF) and pointwise mutual information (PMI). The resulting measures are graphed, in some cases with linear or other trend lines, and possible interpretations of the findings are discussed. As noted above, interpreting the output of optimization measures is not a straightforward matter.

4.1 Perplexity (P), Change in Perplexity (dP), and Second Difference of Perplexity ($d2P$)

When the graph of perplexity as a function of K is examined, a few things become clear. Figure 1 shows that while the trend is generally declining as expected, this is not a smooth curve. This analysis finds a local minimum at K = 20, and another at K = 70. Either of these might be useful, depending on the purpose for which modeling is done.

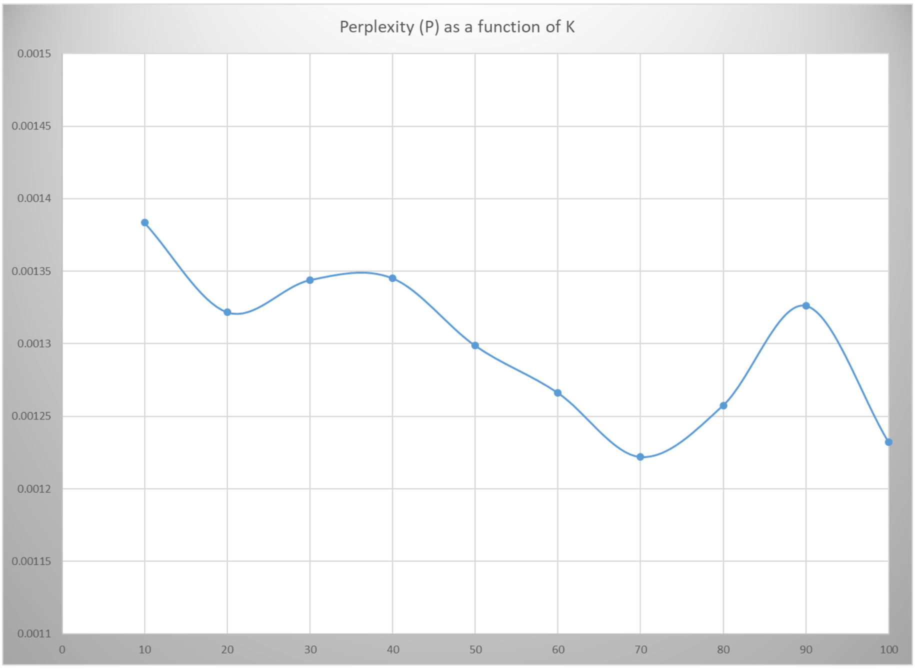

Figure 1: Perplexity as a Function of K

With perplexity, the lower the scale, the better the model is considered to be. Perplexity tneds to favor models with more topics, which may not be useful in cases of text analyses which are looking for intuitive interpretability of topics. Partly for this reason, analysts look to the change in perplexity as another measure, and sometimes to the second difference. Recall that Zhao et al. (2015) found change in perplexity to be a stable and efficient method for optimizing K.

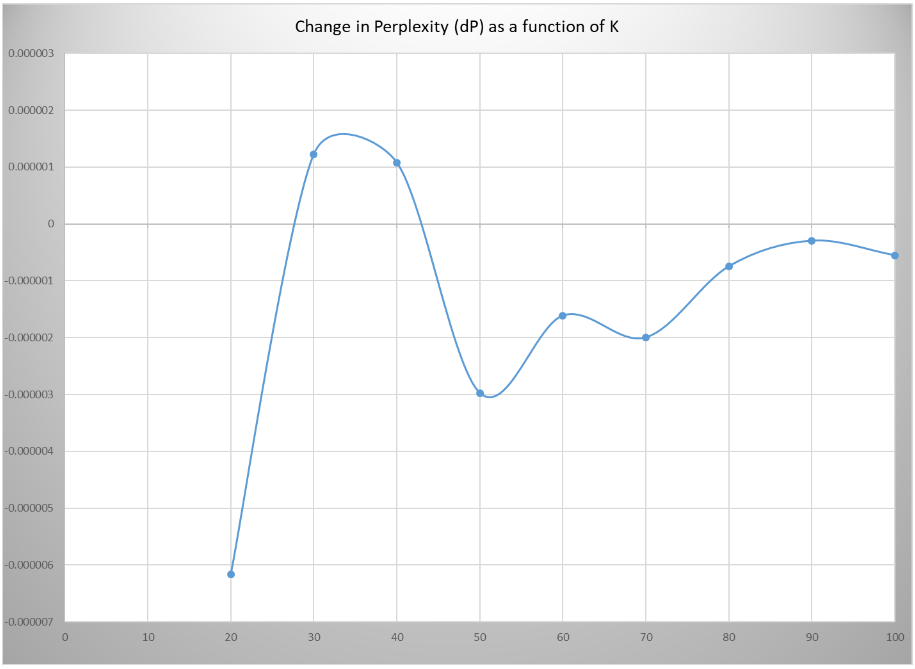

Figure 2: Change in Perplexity as a Function of K

Figure 2 shows that in the case of this data, there is a large increase in dP at the start of this curve, between 20 and 30 K. It then drops off, and the turning point Zhao et al. (2015) describe occurs at K = 50. This turning point is also revealed in Figure 3, a graph of $d2P$, as the first location where the second difference crosses the zero line.

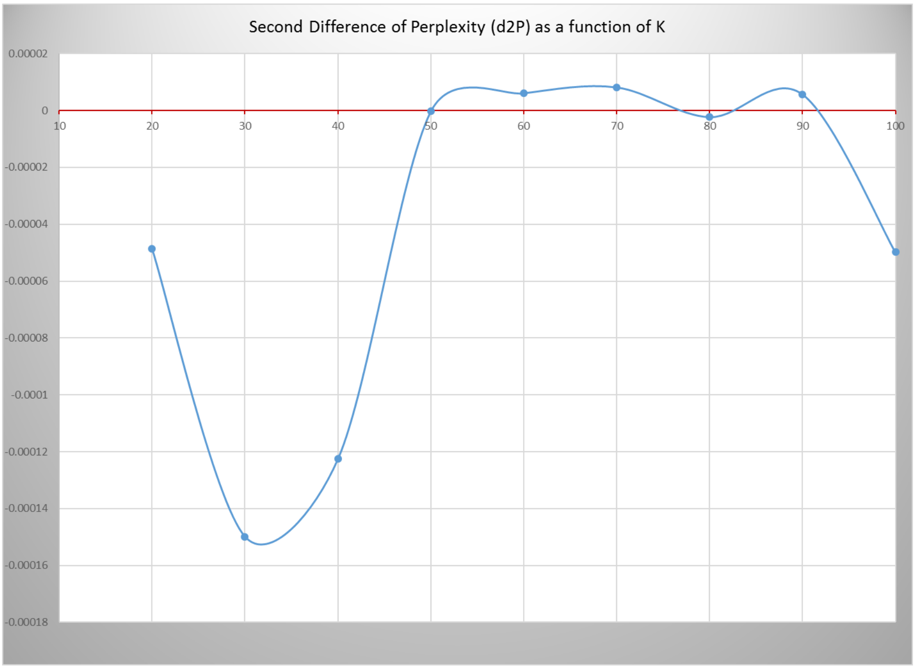

Figure 3: Second Difference of Perplexity as a Function of K

Where perplexity itself is used as a measure, K values of 20 or 70 would appear to be appropriate choices. The change of perplexity scale favors a K of 50. These demonstrate immediately that there is not agreement among differing methods, as to the appropriate value of K.

4.2 Goodness Of Fit (GOF)

The goodness of fit metric for determining K is the proportion or percentage of documents differing significantly from the uniform distribution, with regard to assignment of topics in documents. Finding this measure is a relatively straightforward mathematical procedure, given a reasonably sized corpus, and its interpretation is straightforward as well. When topics have not been assigned randomly within documents, intuitively the topic modeling will be a better fit to the corpus.

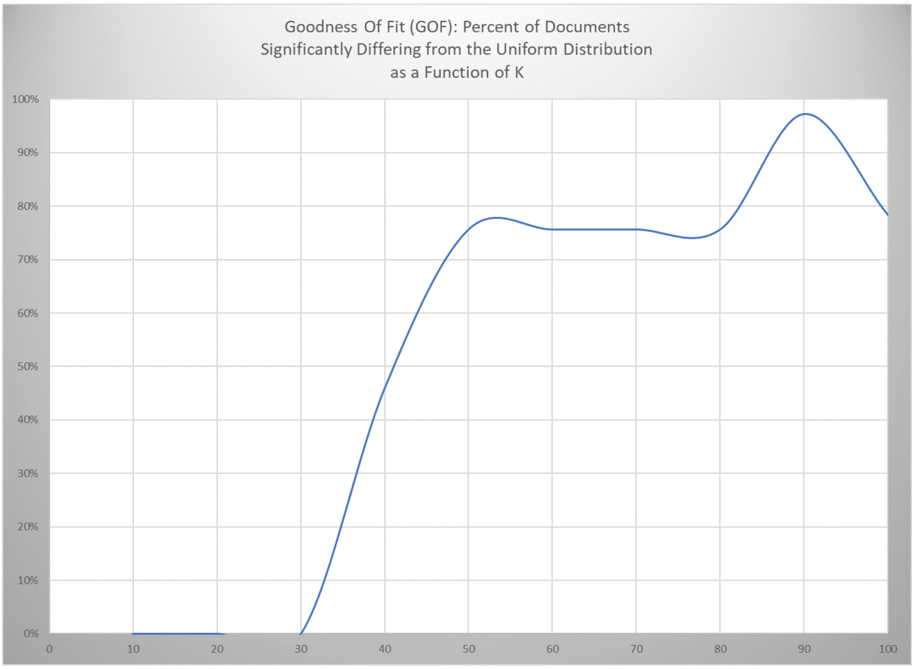

Figure 4: Goodness Of Fit (GOF): Percent of Documents Significantly Differing from the Uniform Distribution, as a Function of K

Figure 4 shows the graph of the Goodness Of Fit analysis outcome, with its highest peak at K = 90. As discussed earlier, some purposes of topic modeling benefit from utilizing lower values of K. An alternate optimum in this case would be indicated at K = 50, where there is an early peak, and the start of a plateau. The only two indicators which agree of the four examined thus far are this intermediary GOF measure at K = 50, with the change of perplexity (dP) measure at K = 50.

4.3 Pointwise Mutual Information (PMI), PMI/K and dPMI/K

Pointwise Mutual Information, as modeled here using the Niraula et al. (2013) method, is shown in Figure 5. The graph of PMI as a function of K almost perfectly fits a power curve, PMI $= 31.249K^{-0.257}$, with an R^2 value of 0.9643. The PMI model fitting method differs from the others in that it favors lower values of K; the graph begins at its optimum and declines from there. While this method corroborates human judgment in favoring lower values of K, the smoothness of this does little to reassure a researcher regarding the correct selection of K.

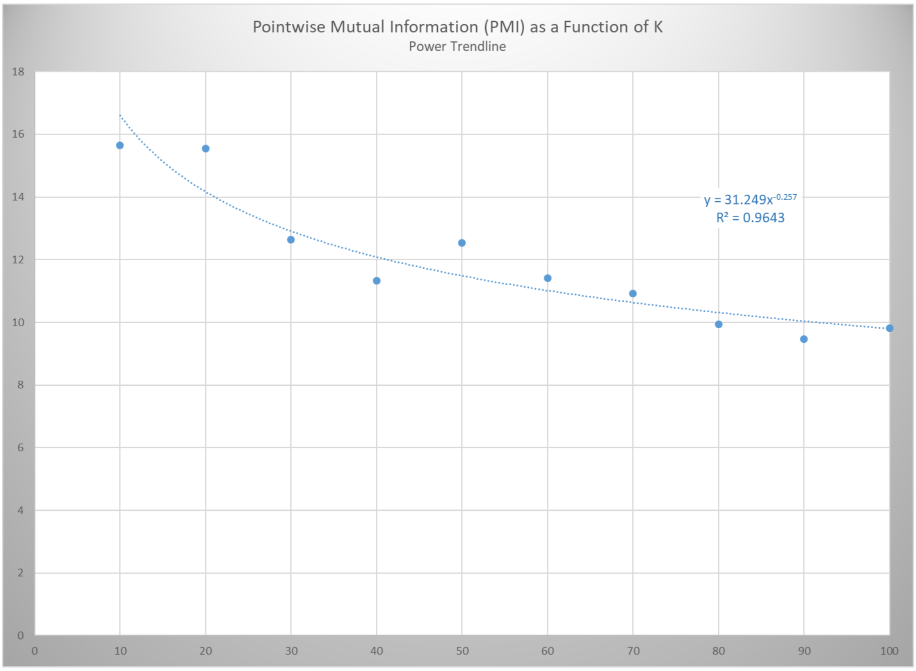

Figure 5: Pointwise Mutual Information (PMI) as a Function of K

Because the incline of this curve is slight, one can also measure the proportional relationship between PMI and K, to find a level of K which provides the most PMI-efficient use of topics. This is shown in Figure 6. Taking a strict proportion of the PMI per K is a start; interestingly, it also provides a smooth power-function curve, with anR^2 value of .9968. When the difference of this ratio is viewed in Figure 7, the change in PMI per K (dPMI/K) provides the only real inflection point this method has to offer, at K = 20. We see in this graph that above 20 the change in PMI per K quickly attains zero.

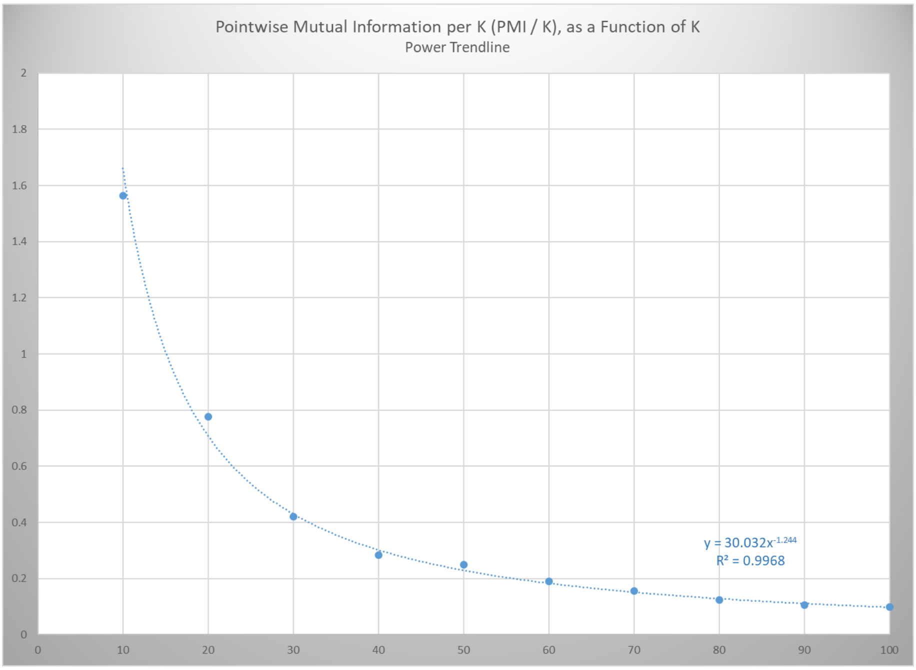

Figure 6: Pointwise Mutual Information per K, (PMI/K)

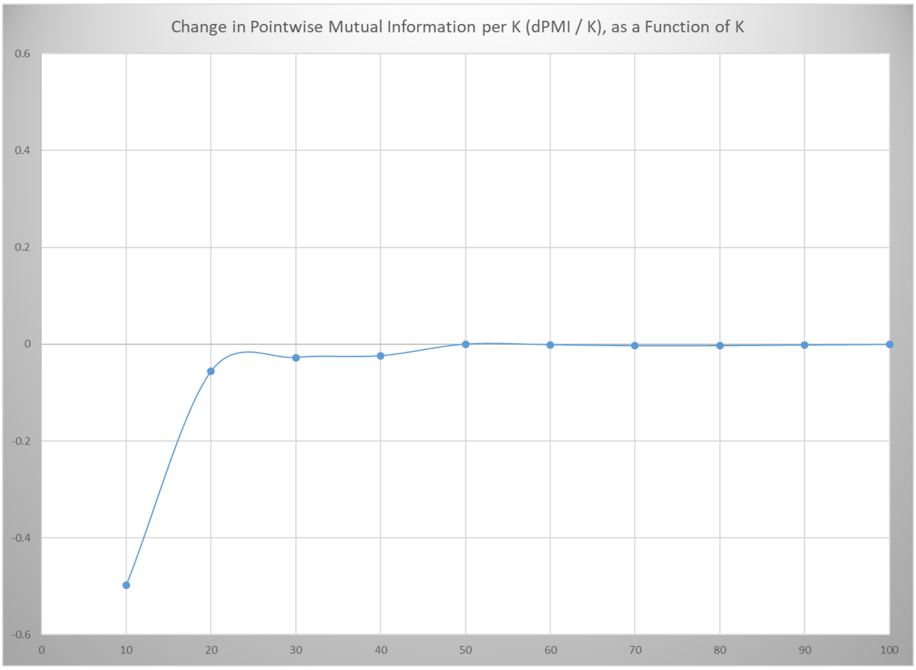

Figure 7: Change in Pointwise Mutual Information per K (dPMI/K)

Discussion and Conclusion

Due to the method of developing the corpus modeled here, the assumption of independence of documents has not been met. In order to include dependencies among documents, modification should occur in the modeling process. Blei and Frazier (2011) for example propose a distance dependent Chinese restaurant process that allows for dependencies between the elements.

In discussing this analysis with the thesis committee, an alternative to perplexity was proposed for measuring the fit of models. As LDA is a Bayesian modeling method, a Bayes Factor is suggested as a possible comparison factor between models. Described by Goodman (1999), a Bayes Factor is the ratio of the likelihood of two competing hypotheses.

After reviewing seven depictions of metrics of K intended to provide optimization, we see that there is not a general agreement among these methods. They do not point toward one K as the most appropriate level of dimensionality to model. In fact, they do not agree upon the measure of a best fit. The inconsistency among our results requires us to continue in the mode of critical thinkers. In each case, we will have to ask ourselves, for what purpose are we developing our topic models? Do we want mathematical purity, a correspondence with human intuition, or a metric intrinsic to the modeling process? The purposes of modeling will have to continue to guide our pursuit of determining K.

Bibliography

AlSumait, L., Barbará, D., Gentle, J., and Domeniconi, C. (2009, September). Topic significance ranking of LDA generative models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 67-82. Springer Berlin Heidelberg.

Banjade, R., Maharjan, N., Niraula, N. B., Rus, V., and Gautam, D. (2015, April). Lemon and tea are not similar: Measuring word-to-word similarity by combining different methods. *International Conference on Intelligent Text Processing and Computational Linguistics*, 335-346. Springer International Publishing, Chicago.

Blei, D. M., & Frazier, P. I. (2011). Distance dependent Chinese restaurant processes.*Journal of Machine Learning Research*, 12(Aug), 2461-2488.

Blei, D. M., and Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 127-134. ACM.

Bowman D. (in press). Variational EM Algorithm for Estimating the Number of Topics in an LDA Model.

Bowman D. (2016). Selecting the number of topics in a latent Dirichlet allocation model. Statistical Learning and Data Science Section, Alexandria, VA, American Statistical Association, 1849-1857.

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., and Blei, D. M. (2009, December). Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems (NIPS)*, 31, 1-9.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), 391.

Ding, C., Li, T. and Peng, W. (2008). On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. Computational Statistics and Data Analysis, 52(8), 3913-3927.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics, 164(4), 1567-1587.

Gerber, G. K., Dowell, R. D., Jaakkola, T. S., and Gifford, D. K. (2007). Automated discovery of functional generality of human gene expression programs. *PLOS Computational Biology*, 3(8), e148.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 2: The Bayes factor. Annals of internal medicine, 130(12), 1005-1013.

Heinrich, G., Kindermann, J., Lauth, C., Paaβ, G., and Sanchez-Monzon, J. (2005). Investigating word correlation at different scopes—a latent concept approach. In *Workshop Lexical Ontology Learning* at Int. Conf. Mach. Learning.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50-57. ACM.

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem:

The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211.

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788-791.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262-272. Association for Computational Linguistics.

Musat, C., Velcin, J., Trausan-Matu, S., and Rizoiu, M. A. (2011). Improving topic evaluation using conceptual knowledge. In *22nd International Joint Conference on Artificial Intelligence (IJCAI)* 3, 1866-1871.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100-108. Association for Computational Linguistics.

Niraula, N., Banjade, R., Ştefănescu, D., and Rus, V. (2013). Experiments with semantic similarity measures based on LDA and LSA. In *Statistical Language and Speech Processing*, 188-199. Springer Berlin Heidelberg.

Pleple, Q. (2013). Topic modeling bibliography. Retrieved from http://qpleple.com/bib/

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945-959.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D. (2012, July). Exploring topic coherence over many models and many topics. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 952-961. Association for Computational Linguistics.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2004). Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. In *Neural Information Processing Systems (NIPS)*, 1385-1392.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2012). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association (JASA)*, 1566-1581.

Wei, X., and Croft, W. B. (2006, August). LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 178-185. ACM.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 1105-1112. ACM.

Zhao, W., Zou, W. and Chen, J.J. (2014). Topic modeling for cluster analysis of large biological and medical datasets. *BMC Bioinformatics*, 15, Suppl 11:511.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, 16(13), S8.