SENTENCE BASED TOPIC MODELING

A THESIS
SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Can Taylan SARI
January, 2014

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Özgür Ulusoy(Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Öznur Taştan

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Savaş Dayanık

Approved for the Graduate School of Engineering and Science:

Prof. Dr. Levent Onural
Director of the Graduate School

ABSTRACT

SENTENCE BASED TOPIC MODELING

Can Taylan SARI
M.S. in Computer Engineering
Supervisor: Prof. Dr. Özgür Ulusoy
January, 2014

Fast augmentation of large text collections in digital world makes inevitable to automatically extract short descriptions of those texts. Even if a lot of studies have been done on detecting *hidden topics* in text corpora, almost all models follow the *bag-of-words* assumption. This study presents a new unsupervised learning method that reveals topics in a text corpora and the topic distribution of each text in the corpora. The texts in the corpora are described by a generative graphical model, in which each sentence is generated by a single topic and the topics of consecutive sentences follow a hidden Markov chain. In contrast to bag-of-words paradigm, the model assumes each sentence as a unit block and builds on a memory of topics slowly changing in a meaningful way as the text flows. The results are evaluated both qualitatively by examining topic keywords from particular text collections and quantitatively by means of perplexity, a measure of generalization of the model.

Keywords: probabilistic graphical model, topic model, hidden Markov model, Markov chain Monte Carlo.

ÖZET

TÜMCE KÖKENLİ KONU MODELLEME

Can Taylan SARI
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Prof. Dr. Özgür Ulusoy
Ocak, 2014

Metin tipindeki veri kümesi sayısı her geçen gün akıl almaz bir şekilde artmaktadır. Bu durum, bu büyük metin veri kümelerinden el değmeden bilgisayarlar yardımıyla ve hızla kısa özetler çıkarmayı kaçınılmaz hale getirmektedir. Büyük metin veri kümelerinden, onların *bilinmeyen, saklı konularını* belirlemeye çalışan birçok çalışma olsa da, bunların hepsi *sözcük torbası* modelini kullanmışlardır. Bu çalışma, metin veri kümelerindeki bilinmeyen, saklı konuları ve bu konulara ait olasılık dağılımlarını ortaya çıkaran yeni bir gözetimsiz öğrenme metodu sunmaktadır. Bu çalışmaya göre veri kümesinde bulunan metinler, her tümcenin tek bir konudan türetildiği ve ardışık tümcelerin konularının bir gizli Markov zinciri oluşturduğu türetici bir çizgisel model tarafından açıklanmaktadır. Sözcük torbası modelinin tersine, önerdiğimiz model tümceyi metnin en küçük yapıtaşı olarak ele alır ve aynı tümce içerisindeki sözcüklerin birbirine anlamca sıkı sıkıya bağlı olduğunu, birbirini takip eden tümcelerin konularının ise yavaşça değiştiğini kabul eder. Önerilen modelin uygulama sonuçları hem konu dağılımlarının en olası kelimelerini ve tümcelere atanan konuları inceleyerek nitel, hem de modelin genelleştirme başarımını ölçerek nicel bir şekilde değerlendirilmektedir.

Anahtar sözcükler: olasılıksal çizgisel model, konu modeli, gizli Markov modeli, Markov zincirleri Monte Carlo.

Acknowledgment

I am deeply grateful to Assoc. Prof. Dr. Savaş Dayanık and Dr. Aynur Dayanık for their valuable guidance, scholarly inputs and consistent encouragement I received throughout the research work. I have been extremely lucky to have advisors and mentors who cared so much about my work, and who responded to my questions and queries so quickly and wisely. Besides, their guidance and encouragement open me a new door into my PhD education on a new field, Industrial Engineering. I have no doubt that we will do a better job on my further research.

I would like to express my gratitude to Prof. Dr. Özgür Ulusoy and Asst. Prof. Dr. Öznur Taştan for reading my thesis and giving me valuable comments.

I would like to thank my wife Aylin for her support, encouragement and quiet patience. She was always there cheering me up and stood by me through the good times and bad.

Finally I would also like to thank my parents Ali&Kader and brothers Süleyman&Mete Nuri for their faith and trust on me ever since I started primary school.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The amount of data in digital media is steadly increasing in parallel with the ever expanding internet and human needs due to cheaper manufacturing of storage devices and starvation for information which resides at the maximum level of all times. Text collections take the biggest share in this data mass in the forms of news portals, blogs, digital libraries. For example, *Wikipedia*, serves as a free digital reference manual to all Internet users. It is a collaborative digital encyclopedia and has 30 million articles in 287 languages[1]. Therefore, it is very difficult to locate the documents of primary interest by a manual or keyword search through the raw texts.

A scholarly article starts with an *abstract* and a number of *keywords*. An abstract is a summary of the entire article and gives brief information to help the reader decide whether it is of any interest. The keywords convey the main themes and the gist of an article. Instead of reading an entire article to find out whether it is related to the topic of current interest, reader can glance at the abstract. The reader can also make a search in the list of "keywords" of articles instead of the entire article. But, unfortunately, abstract and keywords are not included in all types of texts. Hereby, scientists propose *topic models* that extract short topical descriptions and gists of texts in the collections and annotate documents with

[1] http://en.wikipedia.org/wiki/Wikipedia

those *hidden topics*. Topic models help tremendously to organize, summarize and search the large text collections.

Topic models assume that each document is written about a mixture of some topics, and each topic is thought as a probability distribution over a fixed vocabulary. Each word of a document is generated from those topic distributions one by one. This process is referred as *generative process* of a topic model and we will discuss it in detail in Chapter 2. Table 1.1 displays four of the most frequent topics extracted from the articles of the *Science* journal.

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Table 1.1: 15 topmost words from four of most frequent topics, each on a separate column, from the articles of the *Science* journal

The most likely fifteen words for those four topics are listed in Table 1.1 and suggest that the topics are "genetics", "evolution", "disease" and "computers", respectively. Documents are thought to be formed by picking words from those topic distributions. For instance, a document related to "bioinformatics" is likely to be formed by the words picked from "genetics" and "computers" topics. A document on diseases may have been formed by the words picked from "evolution", "disease" and perhaps "genetics" topics.

The generative topic model assumes that each word of a document has a latent

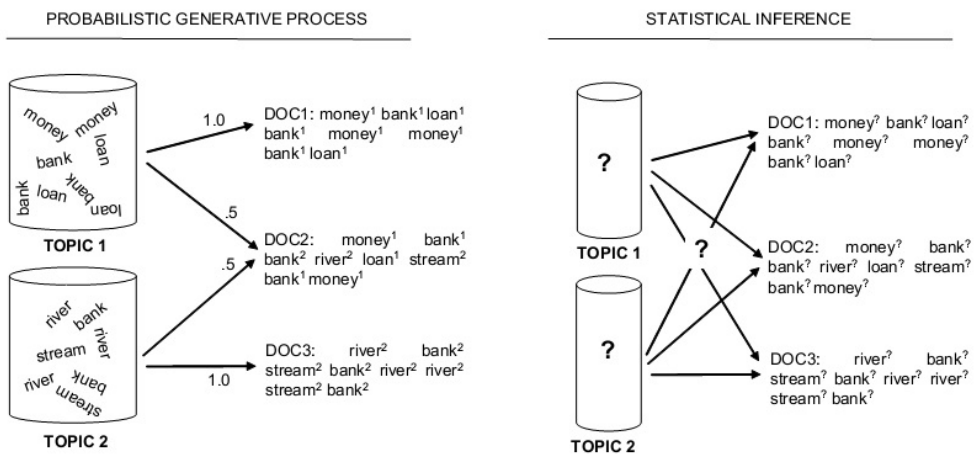

Figure 1.1: An illustration of probabilistic generative process and statistical inference of topic models (Stevyers and Griffiths [12])

topic variable and each word is sampled from the topic distribution identified by its latent variable. Statistical inference methods are used to predict the values of those latent variables to reveal the word usage and the document's topical content. Figure 1.1 is picked from Steyvers and Griffiths [12] article on topic models and illustrates the aims of both generative process and inference method. On the left, two topics are related to "banking" and "rivers", respectively. "DOC1" is totally generated from the words of "TOPIC 1" and "DOC3" is from "TOPIC 2". "DOC2" is, on the other hand, generated by the two topics with equal mixture probabilities. Note that word frequencies from two topics are completely the same for three documents. However, topics and topic assignments to the words are unobserved. Instead, topic models are proposed to extract topics and to assign topics to words and estimate the topic mixture probabilities of documents.

Figure 1.1 implicitly assumes that words generated from topic distributions are placed in a document in random order, and statistical inference method utilizes only the number of occurrences of each word in documents. Namely, each document is assumed to be a *bag-of-words*

The aim of this study is to develop a topic model, more aligned with the thought processes of a writer. This will hopefully result in better performing information retrieval methods in sequel. In realistic information retrieval tasks,

large text collections, incomprehensibly expanding day by day, are to be examined fast by the computer systems. We need more realistic mathematical models to obtain a more precise list of topics and statistical inference methods to fit those models to data fast. Those models will hopefully generate more informative descriptions of the texts in the collections and help users acquire relevant information and related texts smoothly and easily. Search engines, news portals, libraries can be counted among areas of usage.

According to our proposed model, the main idea of a document is often split into several supporting ideas, which are organized according to a topic and discussed in a chain of sentences. Each sentence is expected to be relatively more uniform and most of the time, devoted to a single idea. This leads us to think that every sentence is a bag of words associated with a single topic, and topics of consecutive sentences are related and change slowly. To meet the latter requirement, we assign to each sentence a hidden topic variable, and consecutive topic variables form a hidden Markov chain. Therefore, the proposed model can detect the topical correlations between words in the same sentence and closeby sentences. The proposed and competing models *Latent Dirichlet Allocation (LDA)* and *Hidden Topic Markov Model (HTMM)* are evaluated with four text collections, *Brown*, *Associated Press (AP)*, *Reuters* and *NSF* datasets both quantitatively by means of perplexity, a measure of generalization of the model and qualitatively by examining topic keywords from the text collections. The results show the proposed model has better generalization performance and more meaningful topic distributions/assignments on the text collections.

The thesis has five chapters. Chapter 2 reviews the existing topic models. Chapter 3 presents the Sentence Based Topic Model (SBTM) in detail by means of a generative probabilistic model as well as the Parameter inference by using Gibbs sampling, a special MCMC method. SBTM is evaluated and compared against the existing topic models in Chapters 4 and 5. The thesis concludes with a discussion about topic models and directions for future research.

Chapter 2

Literature Review

The majority of the topic models assume that documents are bags of words, the orders of which in the documents are unimportant. Meaningful words in the documents are collected in the *corpus vocabulary* and their counts are gathered in a *term-document* matrix. Each row and column of the matrix correspond to a word and a document in the corpus, respectively. Typically, a term-document matrix is a sparse matrix, because authors express their ideas by different, synonymous words. Thus, a person may not retrieve the most relevant documents to a query if s/he insists on an exact match between query and document terms. An effective information retrieval method must correlate the query and documents semantically instead of a plain keyword matching.

An early example of topic models is Latent Semantic Indexing (LSI) [3] [4] which represents documents by the column vectors of term-document matrix in a proper semantic space. Firstly, term-document matrix is represented as multiplication of three smaller matrices by *Singular Value Decomposition* (SVD). Let μ be the number of documents, σ the length of word dictionary and F the $\sigma \times \mu$ term-document matrix. The LSI organizes $F = U_0 T_0 V_0^T$ matrix as multiplications of U_0, T_0, V_0 matrices in dimensions of $\sigma \times \tau_0$, $\tau_0 \times \tau_0$, $\mu \times \tau_0$, respectively; see Figure 2.1. T_0 is a diagonal matrix and its diagonal holds singular values of F_0 matrix in decreasing order; U_0 and V_0 are orthogonal matrices; namely, $U_0^\top U_0 = I$ and $V_0^\top V_0 = I$. τ_0 is the number of singular values of F and is between the length

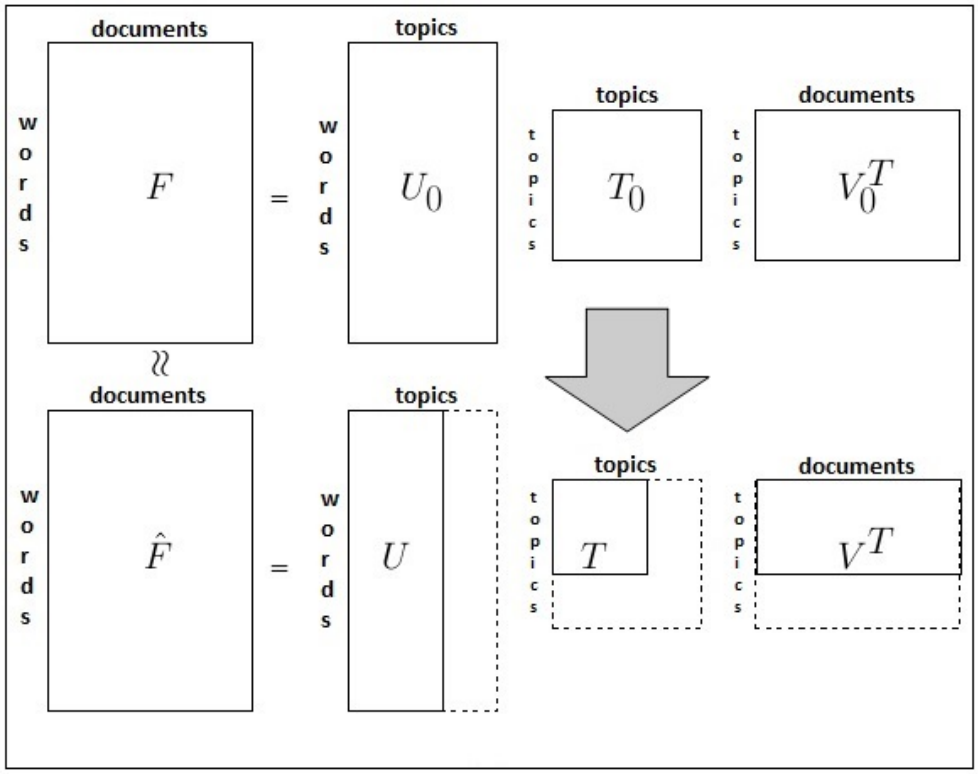

Figure 2.1: Singular Value Decomposition

of dictionary (σ) and number of documents (μ). The LSI obtains new T, U and V matrices by removing rows and columns from T_0 matrix corresponding to small singular values, also columns of U_0 and V_0 corresponding to those small singular values. Therefore,

$$F \approx \hat{F} = UTV^{\top} \tag{2.1}$$

approximation is obtained. The approximate $\hat{F}$ matrix is denser than F matrix. Thus, the LSI establishes a semantic relation between words and documents (even if document does not contain that word) and expresses this relation numerically. Each row of U corresponds to a word and each row of V corresponds to a document. Thus, words and documents can be expressed as τ-dimensional vectors in the same space, where τ is smaller than τ_0. Similarities between words, documents and words-documents can be measured by cosine values of angles between their representative vectors. Therefore, we can get an opportunity to solve the acquisition of similarities problem between words, documents and words-documents in a much smaller dimensional space.

Considering the rows (or columns) of $\hat{F}$ matrix corresponding to words (or documents) as μ-dimensional (or σ-dimensional) vectors in space, similarity between two words (or two documents) can be expressed with cosine of angles between their vectors. We must calculate the inner products of the rows (or columns) of $\hat{F}$ matrix to measure the similarities between words (or documents). Those inner products correspond to the elements of $\hat{F}\hat{F}^\top$ and $\hat{F}^\top\hat{F}$ matrices. Remembering $\hat{F} = UTV^\top$ equation and orthogonality of U, V matrices, we can calculate

$$\hat{F}\hat{F}^\top = (UTV^\top)(UTV^\top)^\top = (UT)(UT)^\top,$$
$$\hat{F}^\top\hat{F} = (UTV^\top)^\top(UTV^\top) = (TV^\top)^\top(TV^\top).$$

In the first equation that expresses the similarities between words, the rows of $\hat{F}$ and UT play same roles. Likewise, in the second equation that expresses the similarities between documents, the columns of $\hat{F}$ and VT play same roles. Therefore, we can express words and documents in the same τ-dimensional space by the rows of UT and VT matrices, respectively.

Let q be a new document and F_q column vector have its word frequencies. The most similar documents to q, are the documents whose vectors in τ-dimensional space has the largest cosine similarity to F_q. Because each document is represented by columns of V matrix instead of $\hat{F}$ and $VT = \hat{F}^\top U$ is acquired from $\hat{F} = UTV^\top$ equation, q document can be represented by $F_q^\top U$ in τ-dimensional space.

Words can have several different meanings (like in "*odd* number" and "an *odd* man"). Unfortunately, the LSI cannot distinguish the meanings of those kind of words.

Another example of such an information retrieval method is Probabilistic Latent Semantic Indexing (PLSI) [5] [6] and is based on a generative probabilistic model, also known as *aspect model*; see its plate notation in Figure 2.2. Each word s in a document is associated with a latent variable, representing the *unobserved topic* of the word. According to PLSI, each document m is generated by a mixture of those latent topics according to the following steps:

1. Each topic t is a probability distribution over a set of vocabulary.

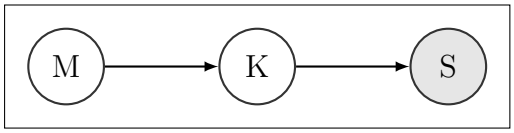

Figure 2.2: Plate notation of PLSI

2. Sample a document m from document probability distribution,$P_M(.)$.

3. Sample a topic k from topic-document conditional probability distribution, $P_{K|M}(.|m)$.

4. Sample a word s from term-topic conditional probability distribution, $P_{S|K}(.|k)$.

5. Add word s to document m and repeat steps 2-5.

Accordingly, probability of occurrence of word s in document m in association with topic k is $P_M(m)P_{K|M}(k|m)P_{S|K}(s|k)$. Note that the word selected when the topic is known, is statistically independent from the text to be added. Since we are not able to observe latent topic variable k, probability that word s occurred in document m is $\sum_k P_M(m)P_{K|M}(k|m)P_{S|K}(s|k) = P_M(m)\sum_k P_{K|M}(k|m)P_{S|K}(s|k)$. According to the model, the likelihood of $n(s,m)$ occurrences of word s in document m equals

$$\prod_m \prod_s \left[P_M(m) \sum_k P_{K|M}(k|m) P_{S|K}(s|k) \right]^{n(s,m)} . \tag{2.2}$$

With the help of observed $n(s,m)$ counts, we can estimate the unknown $P_M(m), P_{K|M}(k|m), P_{S|K}(s|k)$ distributions, by maximizing the log-likelihood function

$$\sum_m \left(\sum_s n(s,m) \right) \log P_M(m) + \sum_m \sum_s n(s,m) \log \sum_k P_{K|M}(k|m) P_{S|K}(s|k) \tag{2.3}$$

subject to

$$\sum_m P_M(m) = 1,$$
$$\sum_k P_{K|M}(k|m) = 1 \text{ (for each document } m),$$
$$\sum_s P_{S|K}(s|k) = 1 \text{ (for each topic } k).$$

Thereby, it is obvious to see that $P_M(m) \propto \sum_s n(s,m)$. Other conditional distributions cannot be estimated in closed form, but can be calculated by means of expectation-maximization [1] [2] iterations as follows:

$$\text{Expectation step:} \qquad P_{K|M,S}(k|m,s) \propto P_{S|K}(s|k)P_{K|M}(k|m), \tag{2.4}$$
$$\text{Maximization step:} \qquad P_{S|K}(s|k) \propto \sum_m n(s,m)P_{K|M,S}(k|m,s), \tag{2.5}$$
$$P_{K|M}(k|m) \propto \sum_s n(s,m)P_{K|M,S}(k|m,s). \tag{2.6}$$

We can divide data to training and validation sets in order to prevent the *overfitting*. We can estimate the conditional distributions from training data. After each expectation-maximization step, we can calculate the likelihood of the validation set, and terminate the estimation process as soon as that likelihood begins to decrease. The *perplexity* is a measure in language modeling to quantify the performance of the model on the validation set and is expressed as

$$\exp\left[-\frac{\sum_{s,m} n(s,m)\log P_{S,M}(s,m)}{\sum_{s,m} n(s,m)}\right] = e^{KL(\hat{P}_{s,m}\|P_{S,M})}e^{H(\hat{P}_{S,M})}. \tag{2.7}$$

Both summations in (2.7) are calculated over the words and documents in the validation set. In (2.7),

$$\hat{P}_{S,M}(s,m) \triangleq \frac{n(s,m)}{\sum_{s',m'} n(s',m')},$$
$$P_{S,M}(s,m) = P_M(m)\sum_k P_{S|K}(s|k)P_{K|M}(k|m)$$

are sample and population probability joint distributions of document (M) and word (S) random variables, respectively under PLSI model. $H(\hat{P}_{S,M})$ is the entropy of observed $\hat{P}_{S,M}$ distribution. $KL(\hat{P}_{S,M} \parallel P_{S,M})$ is Kullback-Leibler divergence between $\hat{P}_{S,M}$ and $P_{S,M}$ distributions. We can use the conditional distributions $P_{S|K}(s|k)$ estimated from the training set when we calculate $P_{S,M}(s,m)$

for each document m and word s in the validation set, but we have to estimate both $P_M(m)$ and $P_{K|M}(k|m)$ probabilities with expectation-maximization method. Hence, it is problematic to measure the generalization performance of the PLSI.

The same problem occurs when documents similar to a new document q are to be found. Intuitively, the conditional topic distribution $P_{K|M}(.|m)$ of a document m similar to document q must be "close" to the conditional topic distribution $P_{K|M}(.|q)$ of document q. For instance, we can claim that a document m resembles document q if the symmetric Kullback-Leibler divergence

$$\frac{1}{2}KL(P_{K|M}(.|m) \parallel P_{K|M}(.|q)) + \frac{1}{2}KL(P_{K|M}(.|q) \parallel P_{K|M}(.|m))$$

is below a proper threshold. If we never deal with document q, again, we can try to find the conditional probability distribution $P_{K|M}(.|q)$ with expectation-maximization method. The log-likelihood function becomes

$$\left(\sum_s n(s,q)\right) \log P_M(q) + \sum_s n(s,q) \log \sum_k P_{K|M}(k|q) P_{S|K}(s|k). \tag{2.8}$$

Thereby, we can directly use the conditional probability distributions $P_{S|K}(s|k)$ estimated from training set. Because q is the only new document, $P_M(q) = 1$ maximizes 2.8. At last, we can estimate $P_{K|M}(.|q)$ by repeating

$$\begin{aligned} &\text{Expectation step:} && P_{K|M,S}(k|q,s) \propto P_{S|K}(s|k) P_{K|M}(k|q), \\ &\text{Maximization step:} && P_{K|M}(k|q) \propto \sum_s n(s,q) P_{K|M,S}(k|q,s), \end{aligned}$$

until convergence of our estimations. Although the PLSI model gets fine mixture of topic distributions for documents, it has two vulnerabilities. The model generates topic mixtures only for the documents present in the training data and therefore it overfits topic distributions to unobserved documents. Moreover, the number of parameters of generating distributions for training documents grows linearly with the number of training documents.

The Latent Dirichlet Allocation *(LDA)* [7] [11] [12] [13] tries to overcome those difficulties with a new generative probabilistic model. The LDA is a graphical

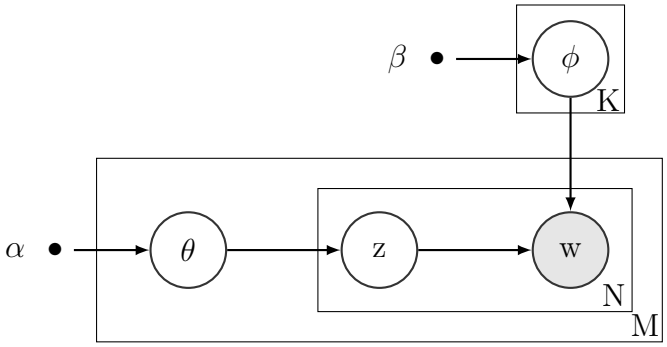

Figure 2.3: Plate notation of LDA

model like PLSI, and each document may have several topics; see its plate notation in Figure 2.3. Differently from PLSI, the words in a typical document generated from κ different topics have multinominal distribution, whose topic probabilities $\Theta = (\Theta_1, \ldots, \Theta_k)$ form a random variable with *Dirichlet* distribution with parameters $\alpha = (\alpha_1, \ldots, \alpha_\kappa)$ and probability density function

$$f_\Theta(\theta_1, \ldots, \theta_\kappa | \alpha) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_\kappa)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_\kappa)} \theta_1^{\alpha_1 - 1} \ldots \theta_\kappa^{\alpha_\kappa - 1}, \quad \sum_{k=1}^{\kappa} \theta_k = 1,\ \theta_k \geq 0,\ 1 \leq k \leq \kappa,$$

Each topic is a multinominal distribution over the dictionary with σ distinct words and topic-word probabilities, $\Phi = (\Phi_s)_{1 \leq s \leq \sigma}$, which is also a random variable having *Dirichlet* distribution with parameters $(\beta_s)_{1 \leq s \leq \sigma}$ and probability density function

$$f_\Phi(\phi_1, \ldots, \phi_\sigma | \beta) = \frac{\Gamma(\beta_1 + \ldots + \beta_\sigma)}{\Gamma(\beta_1) \ldots \Gamma(\beta_\sigma)} \phi_1^{\beta_1 - 1} \ldots \phi_\sigma^{\beta_\sigma - 1}, \quad \sum_{s=1}^{\sigma} \phi_s = 1,\ \phi_s \geq 0,\ 1 \leq s \leq \sigma$$

The hyperparameters of those Dirichlet distributions are usually set as symmetric parameters, $\alpha_1 = ... = \alpha_\kappa \equiv \alpha$ and $\beta_1 = ... = \beta_\sigma \equiv \beta$ to make parameter inference feasible and Dirichlet distributions are simply denoted by $\mathrm{Dir}(\alpha)$ and $\mathrm{Dir}(\beta)$, respectively. The generative process of LDA is as follows:

1. Draw multinominal topic-word distributions $(\phi_s^{(k)})_{1 \leq s \leq \sigma}$, $1 \leq k \leq \kappa$ from $\mathrm{Dir}(\beta)$ distribution on the $(\sigma - 1)$-simplex.

2. For each document, $1 \leq m \leq \mu$,

(a) Draw multinominal document-topic distribution, $(\theta_k^{(m)})_{1\leq k\leq \kappa}$, from $\text{Dir}(\alpha)$ distribution on the $(\kappa-1)$-simplex.

(b) To generate the words $S_{m,1}, S_{m,2}, ...$ in the m^{th} document,

i. draw topics $k_{m,1}, k_{m,2}, ...$ from the same document-topic multinominal distribution, $(\theta_k^{(m)})_{1\leq k\leq \kappa}$,

ii. draw words $s_{m,1}, s_{m,2}, ...$ from the dictionary according to the distributions

$$(\phi_s^{(k_{m,1})})_{1\leq s\leq \sigma}, (\phi_s^{(k_{m,2})})_{1\leq s\leq \sigma}, \cdots,$$

respectively.

The likelihood of the $n(s,m)$ occurrences of word s in document m equals

$$\int_{\phi^{(1)}...\phi^{(\kappa)}} \left(\prod_m \int_{\theta^{(m)}} \prod_s \left[\sum_k \phi_s^{(k)} \theta_k^{(m)} \right]^{n(s,m)} f_\Theta(\theta^{(m)}|\alpha) d\theta^{(m)} \right) \prod_l f_\Phi(\phi^{(l)}|\beta) d\phi^{(l)}.$$

The unknown hyperparameters of model α and β, number of topics κ, hidden topic-term distributions $(\phi_s^{(k)})_{1\leq s\leq \sigma}$, $1 \leq k \leq \kappa$, document-topic distributions $(\theta_k^{(m)})_{1\leq k\leq \kappa}$ can be obtained by maximizing the likelihood of observed words in the documents with Expectation-Maximization (EM) method. However, EM method converges to local maximas. Alternatively, unknown parameters and hidden variables can be estimated with variational Bayesian inference [18] or Markov Chain Monte Carlo (MCMC) [16] algorithms.

Let us explain the application of MCMC in LDA in more detail. There are several MCMC algorithms and one of the most popular MCMC algorithms is Gibbs sampling [22]. Gibbs sampling is a special case of Metropolis-Hastings algorithm and aims to form a Markov chain that has the target posterior distribution as its stationary distribution. After going through a number of iterations and *burn-in* period, it is possible to get samples from that stationary distribution by assuming it as true posterior distribution.

We want to obtain hidden topic-term distributions (Φ), document-topic distributions (Θ) and topic assignment z_i for each word i. The Gibbs sampling easily gives those conditional distributions. Because both topic-term (Φ) and

document-topic (Θ) variables can be calculated just using topic assignments z_i, the *collapsed* Gibbs sampler can be preferred after integrating out topic-term (Φ) and document-topic (Θ) variables.

The full conditional posterior latent topic-word distribution is

$$p(z_i|z_{-i}, \alpha, \beta, w) = \frac{p(z_i, z_{-i}, w|\alpha, \beta)}{p(z_{-i}, w|\alpha, \beta)},$$

where z_{-i} means all topic assignments except z_i. Thus,

$$p(z_i|z_{-i}, \alpha, \beta, w) \propto p(z_i, z_{-i}, w|\alpha, \beta) = p(z, w|\alpha, \beta). \tag{2.9}$$

The conditional distribution on the right side of (2.9) is

$$p(z, w|\alpha, \beta) = \int\int p(z, w, \theta, \phi|\alpha, \beta)d\theta d\phi$$

After expressing $p(z, w, \theta, \phi|\alpha, \beta)$ by means of the Bayesian network in Figure 2.3, we get:

$$\begin{aligned} p(z, w|\alpha, \beta) &= \int\int p(\phi|\beta)p(\theta|\alpha)p(z|\theta)p(w|\phi, z)d\theta d\phi, \\ &= \int p(z|\theta)p(\theta|\alpha)d\theta \int p(w|\phi, z)p(\phi|\beta)d\phi, \\ &= p(z|\alpha).p(w|z, \beta) \end{aligned}$$

is the product of which two integrals in each of which a multinominal distribution is integrated with respect to Dirichlet priors. Because Dirichlet and multinominal distributions are conjugate, we have

$$\begin{aligned} p(z|\alpha) = \int p(z|\theta)p(\theta|\alpha)d\theta &= \int \prod_i \theta_{m,z_i} \frac{1}{B(\alpha)} \prod_k \theta_{m,k}^{\alpha_k} d\theta_m \\ &= \frac{1}{B(\alpha)} \int \prod_k \theta_{m,k}^{n_{d,k}+\alpha_k} d\theta_m \\ &= \frac{B(n_{m,\cdot} + \alpha)}{B(\alpha)}, \end{aligned}$$

where $n_{m,k}$ is the number of words assigned to topic k in document m and $n_{m,\cdot} = (n_{m,1}, \cdots, n_{m,k})$ is the vector of the number of times that each word

in the vocabulary is used in document m. Likewise,

$$\begin{aligned} p(w|z,\beta) = \int p(w|\phi,z)p(\phi|\beta)d\phi &= \int \prod_m \prod_i \phi_{z_{m,i},w_{m,i}} \prod_k \frac{1}{B(\beta)} \prod_w \phi_{k,w}^{\beta_w} d\phi_k \\ &= \prod_k \frac{1}{B(\beta)} \int \prod_w \phi_{k,w}^{\beta_w + n_{k,w}} d\phi_k \\ &= \prod_k \frac{B(n_{k,\cdot} + \beta)}{B(\beta)}, \end{aligned}$$

where $n_{k,w}$ is the total number of times word w is assigned to topic k in the entire text collection, and $n_{k,\cdot} = (n_{k,1}, \cdots, n_{k,\sigma})$ is the vector of the number of assignments of words to topics in the entire text collection. Therefore, the joint distribution in (2.9) is

$$p(z,w|\alpha,\beta) = \prod_m \frac{B(n_{m,\cdot} + \alpha)}{B(\alpha)} \prod_k \frac{B(n_{k,\cdot} + \beta)}{B(\beta)}.$$

The full conditional distribution of Gibbs sampler is then given by

$$\begin{aligned} p(z_i|z^{(-i)}, w, \alpha, \beta) =& \frac{p(w,z|\alpha,\beta)}{p(w,z^{(-i)}|\alpha,\beta)} = \frac{p(z|\alpha)}{p(z^{(-i)}|\alpha)} \cdot \frac{p(w|z,\beta)}{p(w^{(-i)})|z^{(-i)},\beta)p(w_i|\beta)} \\ \propto& \prod_m \frac{B(n_{m,\cdot} + \alpha)}{B(n_{m,\cdot}^{(-i)} + \alpha)} \prod_k \frac{B(n_{k,\cdot} + \beta)}{B(n_{k,\cdot}^{(-i)} + \beta)} \\ \propto& \prod_m \left(\prod_k \left(\frac{\Gamma(n_{m,k} + \alpha_k)}{\Gamma(n_{m,k}^{(-i)} + \alpha_k)} \right) \frac{\Gamma(\sum_{k=1}^K (n_{m,k}^{(-i)} + \alpha_k))}{\Gamma(\sum_{k=1}^K (n_{m,k} + \alpha_k))} \right) \\ & \times \prod_k \left(\prod_w \left(\frac{\Gamma(n_{k,w} + \beta_w)}{\Gamma(n_{k,w}^{(-i)} + \beta_w)} \right) \frac{\Gamma(\sum_{w=1}^W (n_{k,w}^{(-i)} + \beta_w))}{\Gamma(\sum_{w=1}^W (n_{k,w} + \beta_w))} \right) \\ \propto& \prod_m \prod_k (n_{m,k}^{(-i)} + \alpha_k) \prod_k \prod_w \frac{n_{k,w}^{(-i)} + \beta_w}{\sum_{w'} n_{k,w'}^{(-i)} + \beta_{w'}}, \end{aligned}$$

where $n_{m,k}^{(-i)}$ is the number of words in document m assigned to topic k, except the current topic i. After topic assignment variables z are drawn, topic-term (Φ) and document-topic distributions are recalculated by

$$\begin{aligned} \theta_{m,k} &= \frac{n_z(m,k) + \alpha}{\sum_{|l|} n_z(m,l) + \alpha}, \\ \phi_{k,w} &= \frac{n_z(k,w) + \beta}{\sum_{|l|} n_z(l,w) + \beta}, \end{aligned}$$

where $n_z(m, k)$ is the number of words in document m assigned to topic k and $n_z(k, w)$ is the number of words w assigned to topic k in the entire collection according to resample z.

The LDA is referred as a milestone model in the topic modeling using bag-of-words assumption. There are several implementations using different inference methods and number of versions with modified graphical models and different purposes.

The composite model [9] tries to organize words in a document into syntactic and semantic groups. The model assumes that syntactic structures have short-range dependencies: syntactic constraints apply within a sentence and do not persist across different sentences in a document. Nevertheless, semantic structures have long-range dependencies: an author organizes words, sentences even paragraphs along his/her thoughts. Thus, model offers a mixture of *syntactic classes* and *semantic topics* to detect those short and long-range dependencies of words with HMM and topic model, respectively, and obtains word distributions for each syntactic class and semantic topic.

Andrews and Vigliocco [19] propose a model where semantic dependencies among consecutive words follow a Hidden Markov Topics Model (HMTM). According to the model, the words are assigned to random topics, which form a hidden Markov chain.

Hidden Topic Markov Model (HTMM) [20] improves the HMTM model by constructing a Markov chain between sentences instead of words. The model assumes that topics of the words in a document follow a Markov chain. The consecutive words in same sentence are forced to have the same topic assignment. The words in the next sentence will have the same topic as the words of the previous sentence with a fixed probability otherwise. The topic of the next sentence is drawn from the document's topic distribution.

Some aspects of the HTMM is similar to those of our proposed model in the Chapter 3, and we compare the performances of both models in Chapter 4. Like HMTM model, HTMM introduces interdependence between the topics of

consecutive words in a document. But HTMM allows the topic transitions only between the last word of a sentence and first word of the next sentence. Thus, the model guarantees that all words in the same sentence have the same topic.

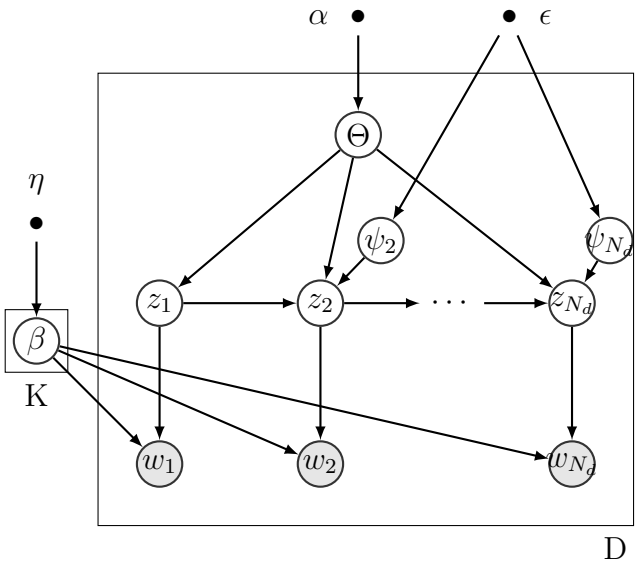

Figure 2.4: Plate notation of HTMM

The generative process of HTMM is as follows:

1. For $z = 1, \cdots, K$,
   Draw $\beta_z \sim \text{Dirichlet}(\eta)$.
2. For $d = 1, \cdots, D$,
   (a) draw $\theta \sim \text{Dirichlet}(\alpha)$,
   (b) set $\psi_1 = 1$,
   (c) for $n = 2, \cdots, N_d$,
      i. if (Begin-Sentence) draw $\psi_n \sim \text{Binom}(\epsilon)$,
         else $\psi_n = 0$.
   (d) for $n = 1, \cdots, N_d$,
      i. if $\psi_n = 0$, then $z_n = z_{n-1}$,
         else $z_n \sim \text{Multinominal}(\theta)$,

ii. draw $w_n \sim$ Multinominal(β_{z_n}).

Although authors state that model generates fine topic distributions and have lower perplexity than other competitive models, we notice some drawbacks in HTMM. Unfortunately, given Z_n, the distribution of Z_{n+1} still depends on whether the $(n+1)$st term is at the beginning of a sentence or not. Therefore, (Z_n) is not a Markov chain, strictly speaking.

The second serios drawback of HTMM that significantly limits the potential of latent variables to detect the smooth changes between topics is the following: if the sentence is going to have a different topic, the new topic is picked independently of the topic of the previous topic from the same distribution Θ_d all the time. However, even if the topics of consecutive sentences are different, they are expected to be locally related.

A third drawback of HTMM study is about the unrealistic and unfair comparisons of between HTMM and other methods. Authors divide each text two halves and one half is places in training set and other half in the test set. Then they run their model on training set and evaluate the model's generalization performance by calculating perplexity on the test set. Because each text appears in both training and testing, the HTMM is likely to overfit and the perplexity calculated during testing is going to be optimistic. In our reassessment of HTMM and other models in Chapter 4, a text is placed entirely in either training or testing set.

None of the existing models satisfactorily capture the thought processes behind creating a document. The main idea of a document is often split into several supporting ideas, which are organized around a specific topic and discussed in a chain of sentences. Each sentence is expected to be relatively uniform and most of the time, devoted to a single main idea. This leads us to think that every sentence is a bag of words associated with a single topic, and topics of consecutive sentences are related and change slowly. To meet the latter requirement, we assign to each sentence a hidden topic variable, and consecutive topic variables form a hidden Markov chain. This model will get rid of word sense ambiguity among synonyms and homonyms within consecutive words and sentences. If the

same word is used in different meanings in two different locations of the document, our proposed model in Chapter 3 will be more likely to distinguish their usages.

Chapter 3

Sentence-based topic modeling

We propose a topic model that, we believe, captures the semantic structure of the text better. Firstly, we ensure that all words in the *same sentence* are assigned to the *same topic*. What we mean by sentence is not a set of words terminated by the punctuations such as dot, two dots, question or exclamation mark. A few sentences or clauses describing distinct ideas may be connected each other by commas, semi-columns or conjunctions. A sentence can be defined as a phrase with a single subject and predicate. Based on this definition, we use an effective method to parse the sentences in a semantic manner [24]. We assume that all semantic words in the same sentence describe the same "topic". We assume that the order of the semantic words in the same sentence is unimportant, because the same ideas can be conveyed with many inverted sentences, albeit in different accents.

An author organizes his/her ideas into sections, each of which is further divided into paragraphs, each of which consists of related sentences and so on. The smallest semantically coherent/uniform building block of a text is a sentence, and each sentence is a collection of terms that describe one main idea/topic. The ideas evolve slowly across sentences: topics of closeby sentences are more closely related than the distant sentences. Therefore, the topical relation between sentences fades away as the distance between sentences increases.

Under those hypotheses, we assume that each sentence is a bag of words and has only one hidden topic variable. Topics of the consecutive sentences follow a hidden Markov chain. Thus, the words in each sentence are semantically very close, and hidden Markov chain allows the topics dynamically evolve across sentences.

The existing models (LSI, PLSI, LDA) mentioned in Chapter 2 neglect

- the order of the terms and sentences,
- topical correlations between terms in the same sentence,
- topical correlations between closeby sentences.

Therefore, we expect the "topics" extracted by our model will be different, more comprehensible and consistent, and present better the gists of documents.

Following the above description of the typical structure of a text, we assume that the documents are generated as shown in Figure 3.1. The latent topic variables $K_1, K_2, \ldots$ are the hidden topic variables for the consecutive sentences of a document and form a Markov chain on the state space $D = \{1, ..., \kappa\}$. The initial state distribution Θ of the Markov chain $K = (K_n)_{n\geq 1}$ is also a random variable and has Dirichlet distribution with parameter $\alpha_1 = ... = \alpha_\kappa = \alpha$ on the $(\kappa-1)$-simplex. Each row of one-step transition probability matrix Π is a discrete distribution over D and is also a random variable on the $(\kappa - 1)$-simplex. The rows $\Pi_1, \ldots, \Pi_\kappa$, have Dirichlet distributions, $\mathrm{Dir}(\gamma_1), \ldots, \mathrm{Dir}(\gamma_\kappa)$, respectively, for some $\gamma = (\gamma_1, \ldots, \gamma_\kappa)$. Each topic is represented by a discrete probability distribution $\Phi = (\Phi_s)_{1\leq s\leq \sigma}$ on the dictionary of σ terms, which is a random variable with Dirichlet distribution $\mathrm{Dir}(\beta)$ on the $(\sigma-1)$-simplex for some $\beta > 0$. The generative process of our model is as follows:

1. Draw independently multinominal topic-word distributions $(\phi_s^{(k)})_{1\leq s\leq \sigma}$, $1 \leq k \leq \kappa$ from $Dir(\beta)$ distribution on the $(\sigma - 1)$-simplex.
2. For each document $1 \leq m \leq \mu$,

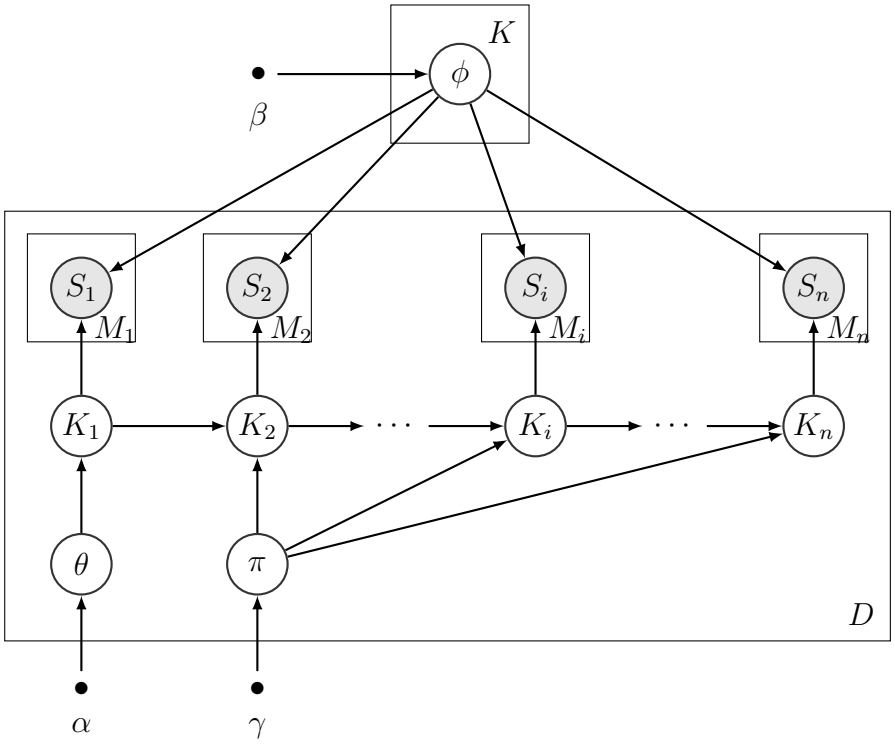

Figure 3.1: Plate notation of the proposed Sentence Based Topic Model

(a) Draw the initial state distribution $\theta^{(m)}, 1 \leq k \leq \kappa, 1 \leq m \leq \mu$ and independent rows $(P_{k,l}^{(m)})_{l \in D}, 1 \leq k \leq \mu, 1 \leq m \leq \mu$ of one-step transition probability matrix $P^{(m)} = [P_{k,l}^{(m)}]_{k,l \in D}$ from $Dir(\alpha)$ and $Dir(\gamma_1), ..., Dir(\gamma_\kappa)$ distributions on the $(\kappa - 1)$-simplex, respectively.

(b) Draw the topics $k_{m,1}, k_{m,2}, ...$ from a Markov chain with initial state distribution $(\theta_k^{(m)})_{1 \leq k \leq \kappa}$ and state-transition probability matrix $P^{(m)}$.

(c) Draw the words $s_{m,t,1}, s_{m,t,2}, ...$ in each $t = 1, 2, ...$ from the same distribution $(\phi_s^{(k_{m,t})})_{1 \leq s \leq \sigma}$, in the dictionary.

Let $n_i(s, m)$ be the number of occurrences of word s in the i^{th} sentence of document m. Then the likelihood of all known model parameters is

$$\prod_m \int_{\phi^{(1)},...,\phi^{(\kappa)}} \left(\int_{P_1^{(m)},...,P_\kappa^{(m)}} \int_{\theta_{(m)}} \left[\sum_{k_1,k_2,...} \left(\left(\prod_i \prod_s \left[\phi_s^{(k_i)} \right]^{n_i(s,m)} \right) \theta_{k_1}^{(m)} \prod_{j \geq 1} P_{k_j,k_{j+1}}^{(m)} \right) \right] \right.$$
$$\left. \times f_\Theta(\theta^{(m)}|\alpha) d\theta^{(m)} \prod_k f_{\Pi_k}(P_k^{(m)}|\gamma_k) dP_k^{(m)} \right) \prod_k f_\Phi(\phi^{(k)}|\beta) d\phi^{(k)}.$$

Unfortunately, the likelihood function of the model is intractable to compute exactly. The maximum likelihood estimator of unknown parameters, α, β and γ and the topic-word distributions $(\phi_s^{(k)})_{1\leq s\leq \sigma}$, $1 \leq k \leq \kappa$, initial topic distribution $(\theta_k^{(m)})_{1\leq k\leq \kappa}$ and topic transition probability matrix $P^{(m)} = [P_{k,l}^{(m)}]_{k,l\in D}$ of documents $1 \leq m \leq \mu$ can be found by approximate inference methods, expectation-maximization (EM), variational Bayesian (VB), or Markov Chain Monte Carlo (MCMC) methods.

For fitting LDA model, Blei [7] proposed mean-field variational expectation maximization, which is a variation of EM algorithm, in which the topic distribution of each document is obtained from the estimated model variables from the last expectation step. This algorithm generates a variational distribution from the same family of true posterior distribution and tries to minimizes the *Kullback-Leibler divergence* between them. However, variational EM method can be stuck to a local optimum and it is more effective on higher dimensional problems. Minka and Lafferty [15] try to overcome those obstacles and increase the accuracy by means of higher-order variational algorithms but at the expense of high computational cost. Therefore, in spite of less scalability, we decided to implement a special MCMC method, known as collapsed Gibbs sampler [21] [23] in order to infer the parameters of our own model. It is easy to implement for LDA-type graphical models, converges more rapidly and does not get stuck to a local optimum.

As an MCMC method, Gibbs sampling [22] method generates random samples from the joint distribution of the variables where a direct inference seems intractable. Suppose one needs a sufficiently large sample to approximate accurately a multivariate distribution $p(x_1, \ldots, x_n)$. Gibbs sampling generates samples iteratively from a conditional distribution of all variables x_{-i} except x_i for $i = 1, \ldots, n$. In other words, each variable x_i is sampled from the conditional joint distribution $p(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ at each iteration. The sequence of samples generated by this process forms a Markov chain, whose stationary distribution is $p(x_1, \ldots, x_n)$ and this sample set will be used to infer the desired functions of the random variables $x_1, \cdots, x_n$.

In SBTM, full joint distribution of the variables, Φ, Θ, Π, K, S, given the hyperparameters α, β and γ is

$$\begin{aligned}P(\Phi,\Theta,\Pi,K,S|\alpha,\beta,\gamma) =&(\prod_{k=1}^{\kappa} p(\phi_k|\beta))(\prod_{m=1}^{M} p(\theta_m|\alpha))(\prod_{m=1}^{M}\prod_{k=1}^{\kappa} p(\pi_{m,k}|\gamma))\\ &\times \prod_{m=1}^{M}\Bigg(p(K_{m,1}|\theta_m)\prod_{n=1}^{N_{m,1}} p(S_{m,1,n}|\phi,K_{m,1})\\ &\times \prod_{t=2}^{T_m} p(K_{m,t}|\pi,K_{m,t-1})\prod_{n=1}^{N_{m,t}} p(S_{m,t,n}|\phi,K_{m,t})\Bigg),\end{aligned}$$

where $\kappa, M, T_m, N_{m,t}$ denote number of topics, number of documents, number of sentences in document m and number of words in sentence t of document m, respectively. Other random variables are as described in Figure 3.1. After multinominal and Dirichlet distributions are plugged in the conditional densities as described in the model, full joint distribution simplifies to

$$\begin{aligned}P(\Phi,\Theta,\Pi,K,S|\alpha,\beta,\gamma) =&\prod_{k=1}^{\kappa}\frac{1}{\Delta(\beta)}\prod_{n=1}^{N}\phi_{k,n}^{\beta_n+f_{k,n}-1}\\ &\times\prod_{m=1}^{M}\frac{1}{\Delta(\alpha)}\prod_{k=1}^{\kappa}\theta_{m,k}^{\alpha_k+e_{m,k}-1}\\ &\times\prod_{m=1}^{M}\prod_{k=1}^{\kappa}\frac{1}{\Delta(\gamma)}\prod_{l=1}^{\kappa}\pi_{m,k,l}^{\gamma_l+g_{m,k,l}-1},\end{aligned}$$

where

$$f_{k,s} = \sum_{m=1}^{M}\sum_{t=1}^{T_m}\sum_{n=1}^{N_{m,t}} 1_{\{K_{m,t}=k,S_{m,t,n}=s\}} \text{ (total count of word } s \text{ assigned to topic } k),$$

$$g_{m,k,l} = \sum_{t=2}^{T_m} 1_{\{K_{m,t-1}=k,K_{m,t}=l\}} \text{ (total count of topic transitions from topic } k \text{ to topic } l),$$

$$e_{m,k} = 1_{\{K_{m,1}=k\}} \text{ (equals one if the first sentence of document } m \text{ is assigned to topic } k).$$

As we can infer from full joint distribution, we want to infer topic-word distribution Φ, initial topic distribution Θ, and topic-transition probability matrix Π, as well as, topic assignments K for each sentence; hence, each word in the document. A Gibbs sampler is in the form of conditional distributions for each

variable given all other variables ($p(x_i|x_{-i})$). When we have topic assignments, K, for each sentence, hence for each word in the sentence, topic-word, initial and topic-transition distributions can be reestimated. Therefore, we implement a collapsed Gibbs sampler by integrating out Φ, Θ and Π variables, which leads to acquire simpler derivations, faster convergence and lower computation cost.

If we integrate out Φ, Θ and Π variables, we obtain

$$\begin{aligned}
P(K,S|\alpha,\beta,\gamma) = \int & p(\phi,\theta,\pi,K,S|\alpha,\beta,\gamma) \\
& \times \prod_{k,n} d\phi_{k,n} \prod_{m,k} d\theta_{m,k} \prod_{m,k,l} d\pi_{m,k,l} \\
& \times \prod_{k=1}^{K} \frac{1}{\Delta(\beta)} \int \prod_{n=1}^{N} \phi_{k,n}^{\beta_n + f_{k,n} - 1} d\phi_{k,n} \\
& \times \prod_{m=1}^{M} \frac{1}{\Delta(\alpha)} \int \prod_{k=1}^{K} \theta_{m,k}^{\alpha_k + e_{m,k} - 1} d\theta_{m,k} \\
& \times \prod_{m=1}^{M} \prod_{k=1}^{K} \frac{1}{\Delta(\gamma)} \int \prod_{l=1}^{K} \pi_{m,k,l}^{\gamma_l + g_{m,k,l} - 1} d\pi_{m,k,l},
\end{aligned}$$

which simplifies to

$$P(K,S|\alpha,\beta,\gamma) = \left(\prod_{k=1}^{K} \frac{\Delta(\beta + f_k)}{\Delta(\beta)} \right) \left(\prod_{m=1}^{M} \frac{\Delta(\alpha + e_m)}{\Delta(\alpha)} \right) \left(\prod_{m=1}^{M} \prod_{k=1}^{K} \frac{\Delta(\gamma + g_{m,k})}{\Delta(\gamma)} \right).$$

After integrating out Φ, Θ and Π variables and making required derivations and simplifications as shown in Appendix A, full conditional distribution of topic assignments $K_{m,1}$ and $K_{m,t}$, of the first and t^{th} sentences document m given other variables and hyperparameters for the collapsed Gibbs sampler are

$$\begin{aligned}
P(K_{m,1} = k | \underset{\sim}{K}^{-(m,1)}) &= \underset{\sim}{k}^{-(m,1)}, \underset{\sim}{S} = \underset{\sim}{s}, \alpha, \beta, \gamma\} \\
&\propto \Delta(\alpha + e_m^{-(m,1),k}) \times \prod_{l=1}^{K} \left(\Delta(\beta + f_l^{-(m,1),k}) \Delta(\gamma + g_{m,l}^{-(m,1),k}) \right),
\end{aligned}$$

$$\begin{aligned}
P(K_{m,t} = k | \underset{\sim}{K}^{-(m,t)}) &= \underset{\sim}{k}^{-(m,t)}, \underset{\sim}{S} = \underset{\sim}{s}, \alpha, \beta, \gamma\} \\
&\propto \prod_{l=1}^{K} \left(\Delta(\beta + f_l^{-(m,t),k}) \Delta(\gamma + g_{m,l}^{-(m,t),k}) \right),
\end{aligned}$$

respectively, for every $k = 1, \ldots, K$, $t = 2, \ldots, T_m$, $m = 1, \ldots, M$ where the count variables $f_l^{-(m,t),k}$, $e_{m,l}^{-(m,1),k}$ and $g_{m,l}^{-(m,t),k}$ are involved and described in detail in Appendix A.

As we calculate the number of transitions $g_{m,k,l}$ from topic k to l in document m, we need to divide the full conditional distribution of topic assignments $K_{m,t}$ into three variables where first (last) sentence does not have any previous (next) sentence. After making required derivations and simplifications as shown in Appendix B, we get full conditional derivations for the first, intermediate and last sentences in (3.1)-(3.6).

Therefore, for every $k \neq k_{m,1}$,

$$\begin{aligned} &P\{K_{m,1} = k | \underset{\sim}{K}^{-(m,1)} = \underset{\sim}{k}^{-(m,1)}, \underset{\sim}{S} = \underset{\sim}{s}, \alpha, \beta, \gamma\} \\ &\quad \propto \frac{\alpha_k}{\alpha_{k_{m,1}}} \frac{P_{N_{m,1}}^{\sum_{s=1}^{N}(\beta_s + f_{k_{m,1},s}) - 1}}{P_{N_{m,1}}^{\sum_{s=1}^{N}(\beta_s + f_{k,s}) - 1 + N_{m,1}}} \prod_{s \in \underset{\sim}{s}_{m,1}} \frac{P_{c_{m,1,s}}^{\beta_s + f_{k,s} - 1 + c_{m,1,s}}}{P_{c_{m,1,s}}^{\beta_s + f_{k_{m,1},s} - 1}} \qquad (3.1) \\ &\quad \times \frac{\sum_{l=1}^{K}(\gamma_l + g_{m,k_{m,1},l}) - 1}{\sum_{l=1}^{K}(\gamma_l + g_{m,k,l})} \frac{\gamma_{k_{m,2}} + g_{m,k,k_{m,2}}}{\gamma_{k_{m,2}} + g_{m,k_{m,1},k_{m,2}} - 1}, \end{aligned}$$

and

$$P\{K_{m,1} = k_{m,1} | \underset{\sim}{K}^{-(m,1)} = \underset{\sim}{k}^{-(m,1)}, \underset{\sim}{S} = \underset{\sim}{s}, \alpha, \beta, \gamma\} \propto 1. \qquad (3.2)$$

For every $t = 2, \ldots, T_m - 1$ and $k = k_{m,t}$

$$\begin{aligned} &P\{K_{m,t} = k | \underset{\sim}{K}^{-(m,t)} = \underset{\sim}{k}^{-(m,t)}, \underset{\sim}{S} = \underset{\sim}{s}, \alpha, \beta, \gamma\} \\ &\quad \propto \frac{P_{N_{m,t}}^{\sum_{s=1}^{N}(\beta_s + f_{k_{m,t},s}) - 1}}{P_{N_{m,t}}^{\sum_{s=1}^{N}(\beta_s + f_{k,s}) - 1 + N_{m,t}}} \prod_{s \in \underset{\sim}{s}_{m,t}} \frac{P_{c_{m,t,s}}^{\beta_s + f_{k,s} - 1 + c_{m,t,s}}}{P_{c_{m,t,s}}^{\beta_s + f_{k_{m,t},s} - 1}} \\ &\quad \times \frac{\sum_{l=1}^{K}(\gamma_l + g_{m,k_{m,t},l}) - 1}{\sum_{l=1}^{K}(\gamma_l + g_{m,k,l})} \frac{\gamma_{k_{m,t+1}} + g_{m,k,k_{m,t+1}}}{\gamma_{k_{m,t+1}} + g_{m,k_{m,t},k_{m,t+1}} - 1} \frac{\gamma_k + g_{m,k_{m,t-1},k}}{\gamma_{k_{m,t}} + g_{m,k_{m,t-1},k_{m,t}} - 1}, \end{aligned} \qquad (3.3)$$

and

$$P\{K_{m,t} = k_{m,t} | \underset{\sim}{K}^{-(m,t)} = \underset{\sim}{k}^{-(m,t)}, \underset{\sim}{S} = \underset{\sim}{s}, \alpha, \beta, \gamma\} \propto 1. \qquad (3.4)$$

For the last sentence of document $t = T_m$ and $k \neq k_{m,T_m}$

$$
\begin{aligned}
&P\{K_{m,T_m} = k | \underset{\sim}{K}^{-(m,T_M)} = \underset{\sim}{k}^{-(m,T_m)}, \underset{\sim}{S} = \underset{\sim}{s}, \alpha, \beta, \gamma\} \\
&\quad \propto \frac{P_{N_{m,T_m}}^{\sum_{s=1}^{N}(\beta_s + f_{k_{m,T_m},s})-1}}{P_{N_{m,T_m}}^{\sum_{s=1}^{N}(\beta_s + f_{k,s})-1+N_{m,T_m}}} \prod_{s \in \underset{\sim}{s}_{m,T_m}} \frac{P_{c_{m,T_m,s}}^{\beta_s + f_{k,s}-1+c_{m,T_m,s}}}{P_{c_{m,T_m,s}}^{\beta_s + f_{k_{m,T_m},s}-1}} \qquad (3.5) \\
&\quad \times \frac{\gamma_k + g_{m,k_{m,T_m-1},k}}{\gamma_{k_{m,T_m}} + g_{m,k_{m,T_m-1},k_{m,T_m}} - 1},
\end{aligned}
$$

and

$$
P\{K_{m,T_m} = k_{m,T_m} | \underset{\sim}{K}^{-(m,T_M)} = \underset{\sim}{k}^{-(m,T_m)}, \underset{\sim}{S} = \underset{\sim}{s}, \alpha, \beta, \gamma\} \propto 1, \qquad (3.6)
$$

where $\underset{\sim}{k} = (k_{m,t})$, $\underset{\sim}{s} = (s_{m,t,n})$ denote the current values of $\underset{\sim}{K} = (K_{m,t})$, $\underset{\sim}{S} = (S_{m,t,n})$, and $(\underset{\sim}{K}^{-(m,t)})$ and $(\underset{\sim}{k}^{-(m,t)})$ denote $\underset{\sim}{K}$ without $K_{m,t}$ and $\underset{\sim}{k}$ without $k_{m,t}$, respectively. $c_{m,t,s}$ holds the number of times that word s appears in the t^{th} sentence of the m^{th} document.

The inference with the collapsed Gibbs sampler is as follows. It initially assigns topics to each sentence at random and sets up topic count variables. At each iteration, (3.1)-(3.6) are computed for the sentences of all documents. New topics are assigned to the sentences, topic count variables are updated, and model parameters, Φ, Θ and Π are predicted by using the updated topic count variables. This iterative process is repeated until the distributions converge.

Chapter 4

Evaluation

In this section, we describe how SBTM is tested on several datasets and compared with other topic models. We start with describing datasets we used in our experiments. Then preprocessing of those datasets is described step by step and results are reported both qualitatively, by in terms of the topics found in the text collections and quantitatively, in terms of perplexity, which measures the generalization performance of the model.

4.1 Datasets

We apply SBTM, LDA, HTMM to four different text corpora in order to investigate the effects of the number of documents, unique words, sentences in a typical document, and the variety of topics in the corpora on the model performance (soundness and relevance of topics found by the model).

The smallest of all four corpora is the *Brown University Standard Corpus of Present-Day American English* (also known as *Brown Corpus*). It contains approximately one million words in 500 texts published in 1961 and is regarded as a fine selection of the contemporary American English. The Brown Corpus is well studied in the field of linguistics, and texts in the corpus range across 15

text categories and more subcategories:

- A. PRESS: Reportage *(44 texts)*
  - Political
  - Sports
  - Society
  - Spot News
  - Financial
  - Cultural
- B. PRESS: Editorial *(27 texts)*
  - Institutional Daily
  - Personal
  - Letters to the Editor
- C. PRESS: Reviews *(17 texts)*
  - *theatre*
  - *books*
  - *music*
  - *dance*
- D. RELIGION *(17 texts)*
  - Books
  - Periodicals
  - Tracts
- E. SKILL AND HOBBIES *(36 texts)*
  - Books
  - Periodicals

- F. POPULAR LORE *(48 texts)*
  - Books
  - Periodicals
- G. BELLES-LETTRES - Biography, Memoirs, etc. *(75 texts)*
  - Books
  - Periodicals
- H. MISCELLANEOUS: US Government and House Organs *(30 texts)*
  - Government Documents
  - Foundation Reports
  - Industry Reports
  - College Catalog
  - Industry House organ
- J. LEARNED *(80 texts)*
  - Natural Sciences
  - Medicine
  - Mathematics
  - Social and Behavioural Sciences
  - Political Science, Law, Education
  - Humanities
  - Technology and Engineering
- K. FICTION: General *(29 texts)*
  - Novels
  - Short Stories
- L. FICTION: Mystery and Detective Fiction *(24 texts)*

  - Novels
  - Short Stories
- M. FICTION: Science *(6 texts)*
  - Novels
  - Short Stories
- N. FICTION: Adventure and Western *(29 texts)*
  - Novels
  - Short Stories
- P. FICTION: Romance and Love Story *(29 texts)*
  - Novels
  - Short Stories
- R. HUMOR *(9 texts)*
  - Novels
  - Essays, etc.

The second corpus we used is extracted from the *Associated Press* news and contains a subset of 2250 documents of TREC AP corpus. It can be downloaded from David M. Blei's webpage `http://www.cs.princeton.edu/~blei/lda-c/index.html` and also another subset of that corpus is used in LDA [7].

Since information retrieval algorithms perform better on large datasets, we also want to measure our model's performance on larger text corpora. The third dataset is a subset from *Reuters* news collection of 12900 documents. Reuters corpus is collected by the Carnegie Group Inc. and Reuters, Ltd. in the course of developing the CONSTRUE text categorization system and documents in the corpus appeared on the Reuters newswire in 1987. It is the most widely used test collection for text categorization methods.

The fourth corpus is NSF dataset and is the largest of all four test corpora. It contains the abstracts of the National Science Foundation Research proposals awarded between 1990 and 2003 and we worked with a subset of 24010 abstracts.

4.2 Text Preprocessing

All four datasets have different formats. For example, in the Brown corpus, each word annotated with its type by its part of speech (verb, pronoun, adverb). Those type of information about words are not used by our model. NSF corpus contains proposal abstracts with some irrelevant information for our model, such as paper publish date, file id, paper award number. In order to remove information and have only the raw text of the documents, we implemented tiny parser programs specific to each corpus. Our text preprocessor needs each text/document in a single file and all documents in a corpus in a folder system named with the title of corpus.

A typical corpus goes through the following six preprocessing steps before the topic model is applied:

Sentence parsing : As our model aims to detect topical relations across sentences, raw text must be broken into sentences. As it was described in Chapter 3, what we mean by sentence is not a set of words terminated by the punctuations such as dot, two dots, question or exclamation mark. A few sentences or clauses describing distinct ideas may be connected each other by commas, semi-columns or conjunctions. A sentence can be defined as a phrase with a single subject and predicate. Based on this definition, we use an effective and "highly accurate" sentence/paragraph breaker as described in its own web page `http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector`. It has been developed by Scott Piao [24] and it employs heuristic rules for identifying boundaries of sentences and paragraphs. An evaluation of a sample text collection shows that it achieved a precision of 0.997352 with a recall of 0.995093.

Conversion to lower case : Along with the sentence parser, we benefit from an *R* package `tm` which provides a text mining framework. It presents methods for data import, corpus handling, preprocessing, meta data management and creation of term-document matrices. We apply data transformation functions of the package, "to lower case" is the first one among them. It converts all words in a corpus to lower case to prevent word ambiguity among same words in different cases.

Remove numbers : The "remove numbers" function of `tm` package is used to remove any numbers in the text collection.

Stop words : "Stop words" are the terms in a language that lack any semantic content such as adverbs, prepositions, conjunctions. A list of stop words in English is included in `tm` package and the words included in that list are removed from the text collections.

Strip whitespace : Extra whitespaces are trimmed from the datasets.

Remove punctuations : All punctuations are removed from the datasets by a tiny script that we developed.

At the end of preprocessing steps, each sentence of a text document appears on a new line and each text document is stored in a separate file. A sample document from AP dataset is shown in Figure 4.1 (line numbers are added for convenience).

4.3 Evaluation of SBTM and comparisons with LDA and HTMM

We evaluate SBTM's performance on four datasets and compare with the most popular topic model LDA, which follows bag-of-words assumption, and with Hidden Topic Markov Model (HTMM) [20], which is a recently proposed model also taking text structure into account as described in Chapter 2.

```
1 embassy provide help victims floods mudslides killed people left
homeless rio de janeiro announced friday
2 consulgeneral louis schwartz check amount gov
3 wellington moreira franco saturday consulates press office reported
4 money total amount embassys emergency fund
5 funds provided government embassies world
6 addition embassy fund united offered provide brazil emergency
supplies
7 money expected petropolis mountain resort miles north rio de janeiro
city
8 civil defense officials friday days rains caused flooding landslides
killed people petropolis left people homeless
9 consular officials visited petropolis check families citizens living
found seriously affected press office
10 countries offered brazil aid supplies include britain france italy
nicaragua
```

Figure 4.1: A sample text document after preprocessing

4.3.1 Generalization performance of models

In the first part of experiments, we compare the competing models with their generalization performance. To evaluate language models, perplexity is the most common performance measure. The aim of topic models is to achieve the highest likelihood on held-out test set. Perplexity is defined as the reciprocal geometric mean of the likelihood of a held-out test corpus; hence, lower perplexity means better generalization performance. The formula of perplexity is

$$P(W|M) = \left\{ \exp - \frac{\sum_{m=1}^{M} \log p(w_m|M)}{\sum_{m=1}^{M} N_m} \right\}, \tag{4.1}$$

where M is all documents in the test corpus, the denominator equals total word count of the test corpus and $\log p(w_m|M)$ is per-word log likelihood.

The joint likelihood of all words in SBTM is

$$L(S|\Theta, \Phi, \Pi) = \prod_{m=1}^{M} \left(\sum_{k_1=1}^{K} \cdots \sum_{k_{T_m}=1}^{K} \theta_m^{k_1} \pi_{m,k_1,k_2} \cdots \pi_{m,k_{T_m-1},k_{T_m}} \times \prod_{t=1}^{T_m} \prod_{s \in S_{m,t}} (\phi_{k_t}^{(s)})^{c(m,t,s)} \right), \tag{4.2}$$

where T_m is number of sentences in document m, $S_{m,t}$ is the set of words in

sentence t of document m, and the perplexity is given by

$$P(S|M) = \exp\left(\frac{-\sum_{m=1}^{M} \log L_m(\underset{\sim}{S}_m|\Theta, \Phi, \Pi)}{|\underset{\sim}{S}|}\right). \tag{4.3}$$

The number of summations in the likelihood formula increases exponentially with the number of documents. Therefore, the likelihood is intractable to be calculated. Instead, we decide to simulate large numbers of samples of θ and π variables in the likelihood function and topic assignments to sentences according to those samples and, approximately calculate the perplexity by

$$P(S|M) = \sum_{m=1}^{M} \log \frac{1}{I} \sum_{i=1}^{I} \prod_{t=1}^{T_m} \prod_{s \in S_{m,t}} (\phi_{k_t,i}^{(s)})^{(i)c(m,t,s)}, \tag{4.4}$$

where I is the number of samples. Figure 4.2 shows the perplexity values for AP dataset as the number of samples changes 1000 to 100000. The perplexity decreases with the sample size. We set the number of samples to a number around which the perplexity levels off.

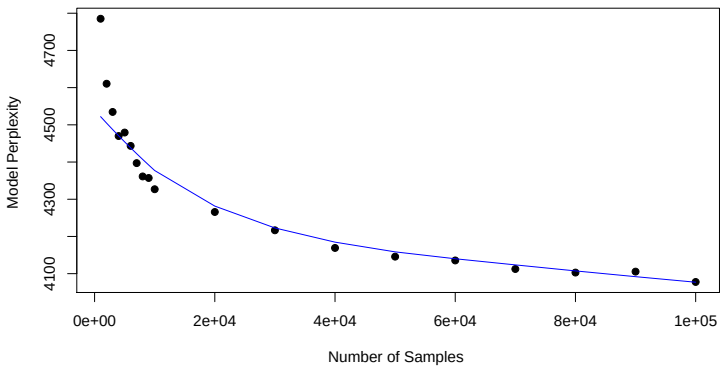

Figure 4.2: Perplexity vs number of samples of perplexity for SBTM

Before comparing the perplexity values of each model, we need to decide on number of iterations of the models. As mentioned in the previous chapters, SBTM and LDA make parameter estimation by Gibbs sampling and HTMM with EM and the forward-backward algorithms. Hence, we try to decide on number of iterations in the training phase with the help of perplexity computing.

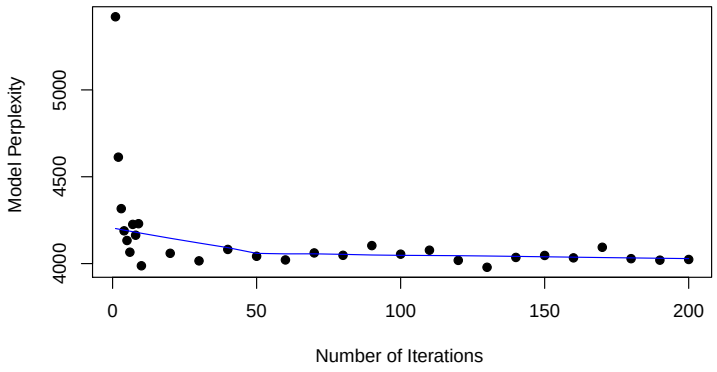

Figure 4.3: Perplexity vs number of iterations for SBTM

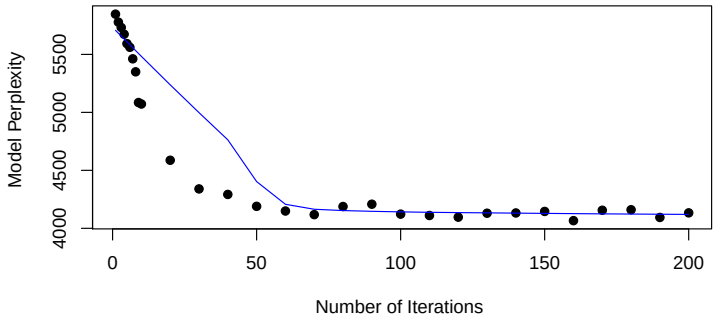

Figure 4.4: Perplexity vs number of iterations for LDA

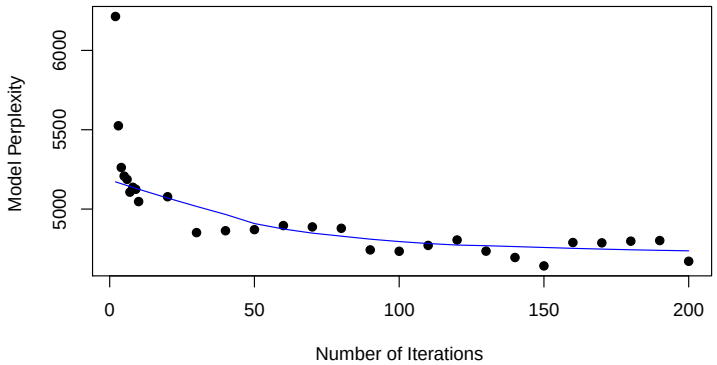

Figure 4.5: Perplexity vs number of iterations for HTMM

We computed perplexity for each model with number of iterations from 1 to 200. Figures 4.3, 4.4 and 4.5 show that all models quickly converge to optimum; in other words, minimum perplexity values are quickly reached. Therefore, we decided to run the competing models for 100 iterations.

After determining the number of iterations for Gibbs sampling, the last issue we need to go through is number of samples obtained from Gibbs sampling. In MCMC sampling methods, one constructs a Markov chain and gets a single sample at the end of each iteration process. Then, s/he can use that single sample or continue sampling from the same Markov chain and get the average of those samples as an estimate of the mean of the distribution of interest. Alternatively, parallel Markov chains can be simulated to get more samples.

But in topic modeling, parallel MCMCs are useless. Stevyers [12] states that, there is not a constant order of topics; at each sample, topics may appear in different orders. Consequently, it is impossible to average the samples from different Markov chains to calculate performance measures. Therefore we decided to run SBTM on a single MCMC realization and average several samples obtained from the same run. Figure 4.6 shows that optimal number of topics found after one and 100 samples do not significantly differ. The topic assignments and topic distributions obtained with 1- and 100-samples also look quite similar. Therefore, we take sample size one.

We applied LDA, HTMM and SBTM to all four datasets as the number of topics changes between 2 and 20. To account for the sampling variance of perplexity and investigate the model accuracy better for each number of topics, we applied K-fold cross-validation for each model. We partitioned AP, Reuters and NSF datasets into 10 folds and Brown dataset to 50 folds at random. Each time one fold is retained as the test set and the remaining folds are used as training set.

Figures 4.7, 4.8, 4.9 and 4.10 display the perplexity values versus the number of topics obtained by applications of LDA, HTMM and SBTM, respectively, for Brown, AP, Reuters and NSF datasets. The vertical bars at each data point extend one standard deviation up and down from the average perplexity values

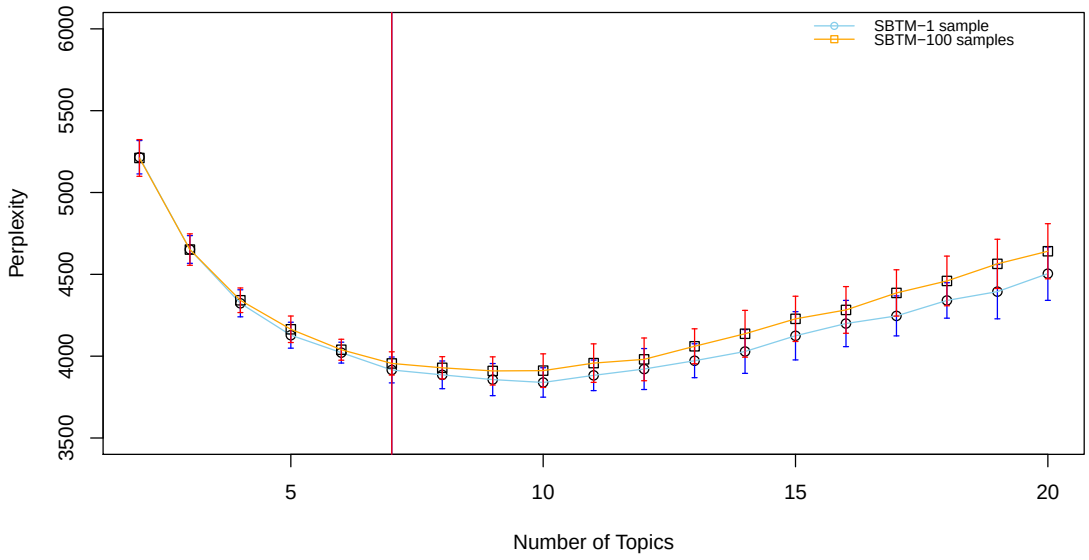

Figure 4.6: Comparison of perplexity results obtained from 1-sample and 100-samples for AP corpus

obtained by cross-validation for each number of topics. The vertical lines mark the number of topics for each model obtained with *1-SE rule*.

Figure 4.7 shows that LDA and HTMM have better generalization performance than SBTM after number of topics exceeds 7. SBTM has larger variances than other models for all number of topics and all models have their highest standard deviation and perplexity values (around 6000-9000) among all four datasets. Because Brown is a rather small text collection, we have only 10 documents for each fold. The topical contents of each fold cannot be uniform among all test folds and that increase the variance. Another problem of having small text collection is that the number of words in the test set which also appear in the training phase is rather small. That lowers the generalization performance and causes lower perplexity values for all models. Besides, SBTM method needs sufficient number of sentence transitions to reliably capture topical relations in a text collection, and Brown corpus falls short of examples to learn from.

Figure 4.8 shows the effect of large dataset on both perplexity and its variance.

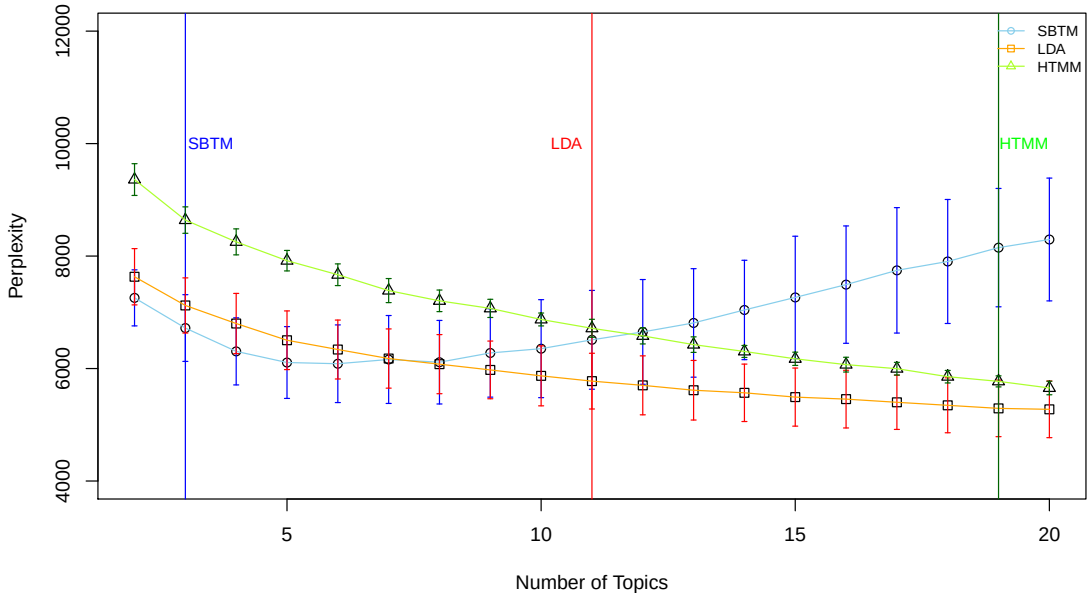

Figure 4.7: Perplexity vs number of topics for Brown corpus

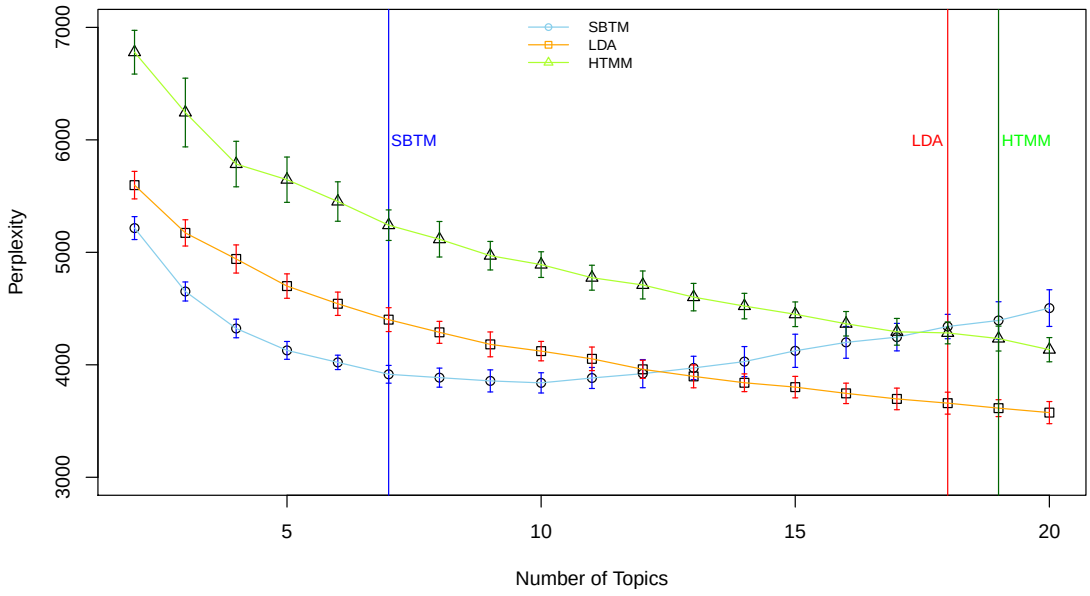

Figure 4.8: Perplexity vs number of topics for AP corpus

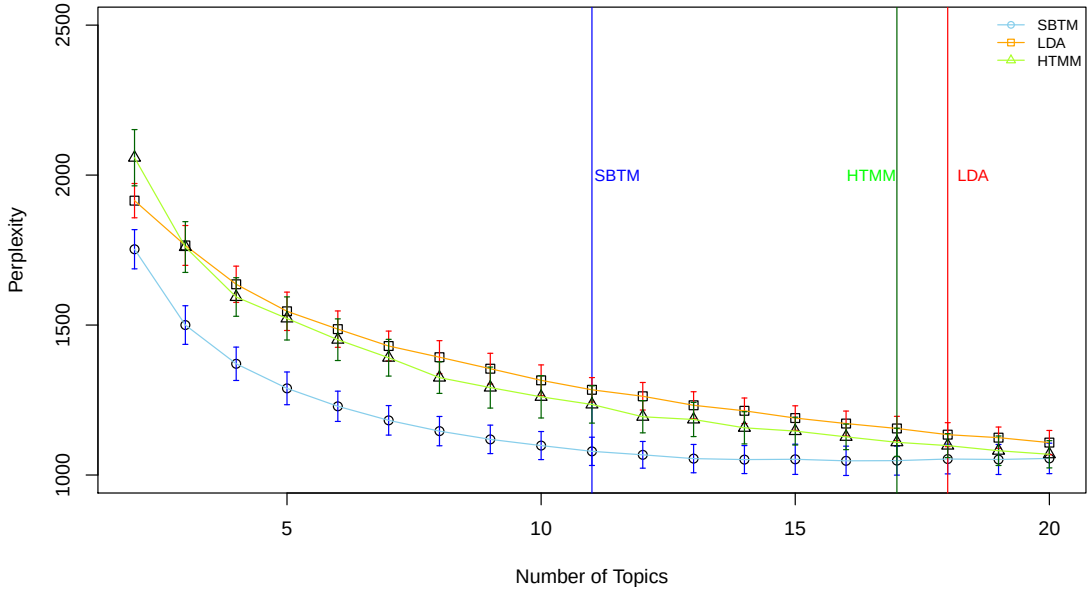

Figure 4.9: Perplexity vs number of topics for Reuters corpus

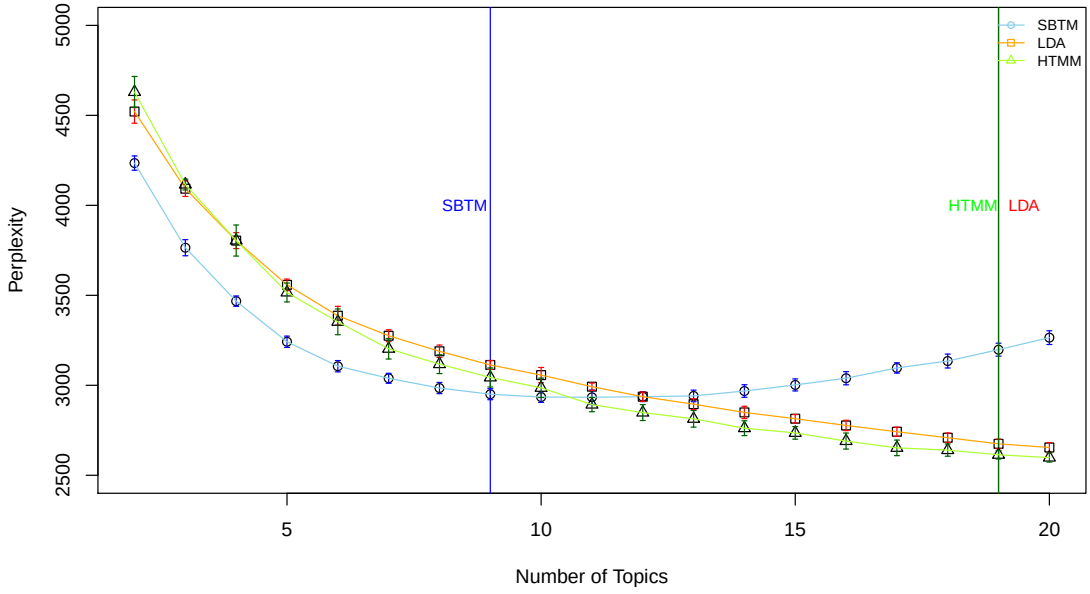

Figure 4.10: Perplexity vs number of topics for NSF corpus

AP corpus has documents approximately four times more than Brown has. SBTM has significantly lower perplexity than HTMM for even all number of topics and LDA until 13 topics. The variances of perplexity values under SBTM are much lower for AP corpus than Brown corpus.

SBTM outperforms LDA and HTMM for all number of topics on Reuters corpus as shown in Figure 4.9. Also SBTM has best generalization performance on Reuters corpus among all datasets where it has perplexity values around 1000. The optimal numbers of topics for SBTM, HTMM, and LDA are 11, 17, and 18, respectively.

The size of the large text collection NSF is reflected by the small variances of perplexity values as indicated by the tiny vertical bars in Figure 4.10. SBTM has lower perplexity than HTMM and LDA until 10 topics and SBTM perplexity starts to increase where other models continue to decrease. The optimal numbers of topics for SBTM, HTMM, and LDA are 9, 19, and 19, respectively.

Even if the perplexity is the most common quantitative measure of generalization performance of the language models, it is not used alone to decide on the best model and number of topics. The topic distributions, the aptness of topic assignments to words/sentences and mixtures of topic proportions for documents are just as important as the perplexity in the final model choice. So, in the following section, we focus on those qualitative issues.

4.3.2 Aptness of topic distributions and assignments

In this section, we compare SBTM with LDA and HTMM by some qualitative measures. We present the topic distributions and topic assignments to words/sentences of two datasets, AP and NSF, by SBTM, LDA, and HTMM. In each table, from Table 4.1 to 4.6, the most likely 20 words of each 10 topics extracted from AP by SBTM, LDA and HTMM are shown, respectively. First we will examine those topics for each model.

We recall that AP dataset contains news articles published in AP newswire.

Therefore, we expect that each model is supposed to form topic distributions about different news topics mentioned in those news articles. Indeed, Table 4.1 indicates that SBTM detects those news topics. The most likely 20 words of "Topic 1" are related with "law" like "court", "attorney", "judge". Another topic, "Topic 2" is obviously related to the "Cold War" by means of terms "soviet", "united", "bush". Also we can point "Topic 6" as last where "bush", "dukakis", "campaign" terms are the most common terms when one talk about "presidential elections in the US". Among those fine topic distributions, we can indicate "Topic 4" as a futile one, in which words do not represent any semantic topic.

LDA is seemed to be very successful on AP dataset like SBTM. "Topic 5" is related with "law" as "Topic 1" of SBTM model. The words "court", "attorney", "judge" have higher probabilities among the vocabulary. LDA and SBTM are in a large agreement on the choice and composition of topic distributions. LDA's "Topic 3" and SBTM's "Topic 9" consist of words related to "finance", like "million", "billion", "trade". Some of the LDA topics however have some defects. "Topic 1" seems to be a futile one since it does not have words about a single topic. "Topic 2" is a mixed one where words about "middle east" and "health" topics are appeared together. Also, "Topic 7" seems to be about "presidential elections in the US" but some words about "USSR" or "Cold War" like "soviet", "gorbachev" also appeared at the top of the list.

HTMM is not as successful as LDA or SBTM. Although some topics have a few words around a single topic like "million", "stock", "market" in "Topic 4" and "prices", "billion", "market" in "Topic 9", we cannot say that HTMM extracted clear, consistent and meaningful topic distributions. That result may have arisen from the drawbacks of HTMM model mentioned in the literature review.

After examining topic distributions extracted from AP dataset, we want to measure whether models assign those topics to words/sentences correctly. We pick a sample document from AP dataset which is about a "financial disclosure by the Democratic presidential candidate Micheal Dukakis". We run the models on that document to get topic assignments for each word/sentence.

Figure 4.11 shows topics assigned to the sentences of an AP document by

SBTM. SBTM assigns "US elections" topic to the introductory sentences of the document in terms of having more information about "US elections" like "presidential candidate Micheal Dukakis", "Federal Election Commission", "Massachusetts governor". Afterwards, topics of consecutive sentences start to evolve from "US elections" to "financial" issues. Document refers to "trust funds", "stocks", companies like "Pepsico", "Kentucky Fried Chicken", "IBM" and several amounts of money. SBTM assigns "financial" topic to almost all remaining sentences. Also it assigns "social welfare" topic to a sentence that mentions "a Harvard University program that cleans up and preserves public grounds".

Correspondingly, Figure 4.12 shows topics assigned to the words of the same AP document by LDA. LDA makes sensible topic assignments to a number of words. "Dukakis", "election", "campaign" terms are assigned to topic related with "US elections". The company brands like "IBM Corp.", "American Telephone and Telegraph" and economy terms like "investment", "trust fund", "holdings", "financial" are assigned to "finance". Meanwhile, as it can be remembered from the description of the LDA model, the model does not consider the semantic structure of document and labels each word independently of the nearby words. As a result of that, considerable number of words are assigned to topics in which those words are included with their connotations. The terms of "South" and "Africa" are assigned to topic related with "crime", "police" and "violence" because "South" and "Africa" terms are usually referred in those type of documents. Also, model cannot distinguish a word's local meaning in a specific the document's topic from its mos frequently used meaning in the entire corpus. The "disclosure" term can be regarded as a "law" term in general, but "financial" adjective and the context of document in which it appears should obviously assign it a "financial" meaning. However, all occurrences of "disclosure" term are assigned to "law" topic. Lastly, LDA model instinctively splits proper nouns into its words like "Micheal", "Dukakis" and assigns them to separate topics.

As we mentioned before, HTMM could not extract consistent and meaningful topic distributions from the AP dataset. Even topic distributions and assignments are unstable and change drastically from one traning to another.

Topic 0		Topic 1		Topic 2		Topic 3		Topic 4	
people	0,0063	court	0,0090	party	0,0097	soviet	0,0090	people	0,0032
percent	0,0052	attorney	0,0053	government	0,0092	united	0,0089	world	0,0031
children	0,0051	judge	0,0051	soviet	0,0073	president	0,0069	time	0,0029
school	0,0032	charges	0,0049	president	0,0063	government	0,0053	york	0,0028
health	0,0032	trial	0,0048	political	0,0059	bush	0,0049	news	0,0024
dont	0,0030	police	0,0046	people	0,0050	military	0,0045	im	0,0024
aids	0,0029	federal	0,0040	national	0,0042	war	0,0042	dont	0,0024
officials	0,0027	prison	0,0040	communist	0,0038	east	0,0040	john	0,0023
program	0,0036	former	0,0037	gorbachev	0,0037	officials	0,0038	family	0,0023
report	0,0026	drug	0,0036	leader	0,0036	talks	0,0038	american	0,0022
care	0,0025	office	0,0032	opposition	0,0034	trade	0,0037	school	0,0021
hospital	0,0025	jury	0,0030	union	0,0033	west	0,0037	home	0,0021
medical	0,0024	death	0,0029	minister	0,0032	nations	0,0035	president	0,0020
drug	0,0024	convicted	0,0029	south	0,0029	foreign	0,0034	yearold	0,0019
university	0,0024	district	0,0029	elections	0,0028	iraq	0,0034	wife	0,0019
women	0,0024	told	0,0028	former	0,0028	union	0,0034	million	0,0018
study	0,0024	law	0,0028	rights	0,0027	world	0,0033	book	0,0018
students	0,0023	department	0,0028	news	0,0026	minister	0,0033	war	0,0018
time	0,0023	investigation	0,0027	told	0,0025	meeting	0,0031	won	0,0018
public	0,0022	defense	0,0027	power	0,0025	american	0,0031	city	0,0017
Social Service		Judiciary System		Politics		Foreign Politics		Life at large	

Topic 5		Topic 6		Topic 7		Topic 8		Topic 9	
miles	0,0043	bush	0,0108	police	0,0146	percent	0,0239	million	0,0081
people	0,0042	dukakis	0,0083	people	0,0084	million	0,0114	company	0,0074
fire	0,0036	president	0,0069	killed	0,0058	market	0,0091	billion	0,0053
water	0,0034	house	0,0064	army	0,0037	prices	0,0086	percent	0,0048
air	0,0031	campaign	0,0060	city	0,0035	billion	0,0083	corp	0,0043
space	0,0029	senate	0,0049	officials	0,0034	stock	0,0067	president	0,0041
officials	0,0028	republican	0,0049	miles	0,0033	dollar	0,0063	federal	0,0040
city	0,0028	percent	0,0047	military	0,0033	rose	0,0058	inc	0,0037
north	0,0025	democratic	0,0045	reported	0,0030	cents	0,0057	workers	0,0036
coast	0,0025	committee	0,0042	government	0,0030	trading	0,0050	co	0,0033
southern	0,0024	bill	0,0042	fire	0,0029	rates	0,0049	business	0,0029
time	0,0023	sen	0,0040	shot	0,0027	sales	0,0047	court	0,0029
national	0,0022	people	0,0038	monday	0,0026	index	0,0045	government	0,0028
northern	0,0022	budget	0,0037	night	0,0026	lower	0,0045	pay	0,0027
rain	0,0021	jackson	0,0034	air	0,0026	late	0,0042	air	0,0026
hours	0,0021	presidential	0,0034	spokesman	0,0026	exchange	0,0041	employees	0,0026
south	0,0021	congress	0,0033	soldiers	0,0026	price	0,0041	contract	0,0025
reported	0,0021	vote	0,0032	injured	0,0026	yen	0,0041	service	0,0025
central	0,0020	told	0,0031	five	0,0025	fell	0,0041	bank	0,0025
day	0,0019	rep	0,0031	troops	0,0024	rate	0,0040	plan	0,0025
Civil Service		Presidential Elections		Military Actions		Finance		Government Expenditures	

Table 4.1: Topics extracted from AP by SBTM

Financial disclosure forms indicate a $1 million trust fund in which Democratic presidential candidate Michael Dukakis is co-beneficiary sold stock worth up to $65,000 last year from companies with ties to South Africa. The forms, filed with the Federal Election Commission, are consistent with but more detailed than state financial disclosure forms filed by the Massachusetts governor last month. Dukakis gained no income in 1987 from the family trust established by his late father, whose equal beneficiaries are Dukakis and Bates College, the elder Dukakis' alma mater. The trust drew attention last year when similar forms filed with the FEC disclosed that until 1986 it included stock in companies that do business in South Africa. Dukakis has no role in investment decisions of the trust but in 1986, when questioned about its holdings during a gubernatorial re-election campaign, told the bank which controls the account to divest any stock in companies with South African dealings. Stock in 10 such companies was immediately sold. The trust gained between $15,001 and $50,000 last year by selling stock in Unisys Corp., which maintains operations in South Africa. Dukakis campaign spokesman Steven Akey said the trust had owned stock in a company acquired by Unisys last year and sold it because of Unisys' ties to South Africa. The trust also reported capital gains of between $5,001 and $15,000 from the sale of Pepsico Inc. stock. The company maintains indirect ties to South Africa through its Kentucky Fried Chicken subsidiary. Three companies in which the trust has holdings have indirect ties to racially segregated South Africa, largely through distribution agreements and spare parts sales contracts, according to the Washington-based Investor Responsibility Research Center. They are IBM Corp., General Electric Corp. and American Telephone & Telegraph. Akey said the trust's investment policy is to not acquire stock in any company on the Washington group's list of companies with direct dealings in South Africa. Dukakis is also a potential co-beneficiary of another $1 million trust left by his father. But that trust is not considered part of his net worth because Dukakis' 84-year-old mother controls it and could decide to alter arrangements which currently call for that fund to be split by Dukakis and Bates. The FEC forms, which cover the period from Jan. 1, 1987, to March 31, 1988, show Dukakis' principal source of income during that period was $106,310 from his salary as governor. The forms were filed Friday. He and his wife, Kitty, also earned between $1,001 and $2,500 in interest from a savings account that holds between $15,001 and $50,000; between $101 and $1,000 in interest from a smaller bank account in Mrs. Dukakis' name; and between $5,000 and $15,000 in interest from the governor's holdings in the state retirement plan. Mrs. Dukakis also earned $20,000 last year in salary from a Harvard University program that cleans up and preserves public grounds. An investment account in the name of the couple's three children earned between $2,600 and $5,100 last year, according to the FEC forms. The couple also owns a home in suburban Brookline that is worth more than $100,000. Dukakis and his wife, who do not use credit cards, listed no liabilities. The largest investment in the Panos Dukakis trust is a tax-exempt money market fund worth between $100,001 and $250,000. Five of the trust's investments are worth between $50,001 and $100,000: holdings in Maine municipal bonds, a tax-exempt bank trust fund, General Electric, AT&T and Atlantic Richfield Corp.

	Topic 0 - Social Service
	Topic 6 - Presidential Elections
	Topic 9 - Government Expenditures

Figure 4.11: Topics assigned to sentences of an AP document by SBTM

Topic 0		Topic 1		Topic 2		Topic 3		Topic 4	
police	0,0177	people	0,0093	united	0,0090	percent	0,0439	air	0,0118
government	0,0120	school	0,0077	court	0,0074	million	0,0188	th	0,0054
people	0,0111	home	0,0072	american	0,0068	billion	0,0114	time	0,0053
south	0,0087	family	0,0071	iraq	0,0065	report	0,0091	space	0,0045
military	0,0080	children	0,0070	human	0,0058	oil	0,0090	flight	0,0044
killed	0,0073	life	0,0065	health	0,0057	trade	0,0084	plane	0,0043
officials	0,0061	dont	0,0065	medical	0,0052	prices	0,0082	news	0,0041
army	0,0058	yearold	0,0064	decision	0,0050	months	0,0076	aircraft	0,0036
troops	0,0054	university	0,0050	rights	0,0050	sales	0,0076	navy	0,0036
official	0,0053	wife	0,0050	iraqi	0,0048	cents	0,0063	monday	0,0035
forces	0,0045	hospital	0,0045	kuwait	0,0048	increase	0,0060	accident	0,0035
war	0,0043	im	0,0044	system	0,0044	rate	0,0058	airport	0,0035
reported	0,0040	students	0,0043	international	0,0044	month	0,0051	wednesday	0,0035
security	0,0039	black	0,0043	gulf	0,0043	economic	0,0049	spokesman	0,0034
attack	0,0038	thats	0,0038	saudi	0,0040	economy	0,0049	american	0,0034
soldiers	0,0038	women	0,0037	americans	0,0039	workers	0,0048	week	0,0033
israel	0,0038	time	0,0036	claims	0,0036	price	0,0047	force	0,0032
condition	0,0037	ms	0,0036	public	0,0036	reported	0,0046	airlines	0,0032
africa	0,0037	care	0,0035	aids	0,0036	lower	0,0045	base	0,0031
capital	0,0036	hes	0,0033	drug	0,0036	expected	0,0043	hours	0,0030
Military Actions		Life at large		UN/Middle East/Foreign Politics		Finance		Air Force	

Topic 5		Topic 6		Topic 7		Topic 8		Topic 9	
court	0,0098	house	0,0111	president	0,0160	city	0,0094	company	0,0145
former	0,0083	federal	0,0088	soviet	0,0159	officials	0,0081	million	0,0133
attorney	0,0083	committee	0,0082	bush	0,0152	people	0,0074	market	0,0110
told	0,0079	bill	0,0080	party	0,0096	water	0,0060	stock	0,0109
judge	0,0078	program	0,0080	political	0,0078	california	0,0059	york	0,0102
office	0,0074	senate	0,0076	union	0,0074	san	0,0058	inc	0,0081
charges	0,0071	congress	0,0071	dukakis	0,0072	fire	0,0058	late	0,0080
trial	0,0068	defense	0,0068	campaign	0,0062	national	0,0053	bank	0,0080
prison	0,0066	money	0,0065	democratic	0,0060	plant	0,0050	dollar	0,0076
district	0,0053	budget	0,0063	told	0,0058	miles	0,0049	board	0,0071
federal	0,0052	plan	0,0055	meeting	0,0057	workers	0,0045	trading	0,0069
drug	0,0051	tax	0,0054	east	0,0053	southern	0,0044	corp	0,0068
investigation	0,0050	administration	0,0052	talks	0,0052	texas	0,0043	business	0,0066
justice	0,0046	government	0,0048	united	0,0051	friday	0,0039	co	0,0065
law	0,0044	rep	0,0048	gorbachev	0,0050	center	0,0039	exchange	0,0063
found	0,0044	department	0,0046	leader	0,0047	service	0,0038	share	0,0058
death	0,0042	sen	0,0045	minister	0,0044	damage	0,0037	financial	0,0055
charged	0,0042	public	0,0042	support	0,0043	near	0,0037	billion	0,0054
john	0,0039	private	0,0042	presidential	0,0043	coast	0,0034	thursday	0,0053
convicted	0,0037	national	0,0041	government	0,0041	county	0,0034	index	0,0051
Judiciary System		Parliament/Politics		Cold War/Presidential Elections		Civil Service		Finance	

Table 4.2: Topics extracted from AP by LDA

Financial disclosure forms indicate a $1 million trust fund in which Democratic presidential candidate Michael Dukakis is co-beneficiary sold stock worth up to $65,000 last year from companies with ties to South Africa. The forms, filed with the Federal Election Commission, are consistent with but more detailed than state financial disclosure forms filed by the Massachusetts governor last month. Dukakis gained no income in 1987 from the family trust established by his late father, whose equal beneficiaries are Dukakis and Bates College, the elder Dukakis' alma mater. The trust drew attention last year when similar forms filed with the FEC disclosed that until 1986 it included stock in companies that do business in South Africa. Dukakis has no role in investment decisions of the trust but in 1986, when questioned about its holdings during a gubernatorial re-election campaign, told the bank which controls the account to divest any stock in companies with South African dealings. Stock in 10 such companies was immediately sold. The trust gained between $15,001 and $50,000 last year by selling stock in Unisys Corp., which maintains operations in South Africa. Dukakis campaign spokesman Steven Akey said the trust had owned stock in a company acquired by Unisys last year and sold it because of Unisys' ties to South Africa. The trust also reported capital gains of between $5,001 and $15,000 from the sale of Pepsico Inc. stock. The company maintains indirect ties to South Africa through its Kentucky Fried Chicken subsidiary. Three companies in which the trust has holdings have indirect ties to racially segregated South Africa, largely through distribution agreements and spare parts sales contracts, according to the Washington-based Investor Responsibility Research Center. They are IBM Corp., General Electric Corp. and American Telephone & Telegraph. Akey said the trust's investment policy is to not acquire stock in any company on the Washington group's list of companies with direct dealings in South Africa. Dukakis is also a potential co-beneficiary of another $1 million trust left by his father. But that trust is not considered part of his net worth because Dukakis' 84-year-old mother controls it and could decide to alter arrangements which currently call for that fund to be split by Dukakis and Bates. The FEC forms, which cover the period from Jan. 1, 1987, to March 31, 1988, show Dukakis' principal source of income during that period was $106,310 from his salary as governor. The forms were filed Friday. He and his wife, Kitty, also earned between $1,001 and $2,500 in interest from a savings account that holds between $15,001 and $50,000; between $101 and $1,000 in interest from a smaller bank account in Mrs. Dukakis' name; and between $5,000 and $15,000 in interest from the governor's holdings in the state retirement plan. Mrs. Dukakis also earned $20,000 last year in salary from a Harvard University program that cleans up and preserves public grounds. An investment account in the name of the couple's three children earned between $2,600 and $5,100 last year, according to the FEC forms. The couple also owns a home in suburban Brookline that is worth more than $100,000. Dukakis and his wife, who do not use credit cards, listed no liabilities. The largest investment in the Panos Dukakis trust is a tax-exempt money market fund worth between $100,001 and $250,000. Five of the trust's investments are worth between $50,001 and $100,000: holdings in Maine municipal bonds, a tax-exempt bank trust fund, General Electric, AT&T and Atlantic Richfield Corp.

	Topic 0 - Military Actions
	Topic 1 - Life at large
	Topic 2 - UN/Middle East/Foreign Policy
	Topic 3 - Finance
	Topic 5 - Judiciary System
	Topic 6 - Parliament/Politics
	Topic 7 - Cold War/Presidential Elections
	Topic 8 - Civil Service
	Topic 9 - Finance

Figure 4.12: Topics assigned to words of an AP document by LDA

Topic 0		Topic 1		Topic 2		Topic 3		Topic 4	
people	0,0054	people	0,0059	million	0,0067	police	0,0056	million	0,0081
president	0,0041	bush	0,0038	billion	0,0065	cents	0,0048	percent	0,0071
court	0,0039	million	0,0037	bush	0,0047	cent	0,0031	stock	0,0057
government	0,0031	president	0,0035	percent	0,0042	people	0,0031	market	0,0055
time	0,0028	police	0,0035	house	0,0042	west	0,0028	yen	0,0043
political	0,0025	government	0,0034	president	0,0037	city	0,0028	dollar	0,0042
officials	0,0024	percent	0,0029	budget	0,0037	york	0,0026	index	0,0040
former	0,0023	time	0,0029	federal	0,0032	united	0,0026	people	0,0038
soviet	0,0023	news	0,0026	officials	0,0032	thursday	0,0026	prices	0,0033
day	0,0021	students	0,0025	company	0,0031	lower	0,0025	rose	0,0033
told	0,0021	school	0,0025	tax	0,0030	time	0,0025	average	0,0032
party	0,0020	dont	0,0023	defense	0,0028	wednesday	0,0023	trading	0,0032
department	0,0020	day	0,0021	dukakis	0,0027	told	0,0023	exchange	0,0032
public	0,0020	court	0,0021	committee	0,0027	government	0,0023	soviet	0,0030
john	0,0019	fire	0,0021	air	0,0026	monday	0,0022	week	0,0030
dont	0,0019	thursday	0,0021	senate	0,0026	bid	0,0022	police	0,0029
company	0,0019	told	0,0020	told	0,0025	late	0,0022	government	0,0029
found	0,0018	dukakis	0,0020	campaign	0,0025	north	0,0022	billion	0,0028
tuesday	0,0017	days	0,0019	congress	0,0024	million	0,0022	york	0,0027
killed	0,0017	public	0,0019	united	0,0023	national	0,0021	shares	0,0027

Topic 5		Topic 6		Topic 7		Topic 8		Topic 9	
president	0,0052	soviet	0,0076	government	0,0047	million	0,0058	percent	0,0164
police	0,0051	people	0,0048	president	0,0046	people	0,0044	prices	0,0046
people	0,0046	officials	0,0044	people	0,0043	dukakis	0,0039	billion	0,0044
government	0,0040	president	0,0043	united	0,0042	percent	0,0034	market	0,0041
bush	0,0040	news	0,0034	south	0,0035	national	0,0032	million	0,0039
united	0,0034	government	0,0033	bush	0,0030	federal	0,0027	soviet	0,0036
time	0,0033	told	0,0026	american	0,0025	jackson	0,0026	government	0,0035
party	0,0031	time	0,0026	officials	0,0025	government	0,0025	united	0,0033
trade	0,0030	union	0,0026	told	0,0024	york	0,0025	sales	0,0030
drug	0,0030	gorbachev	0,0026	thursday	0,0023	former	0,0024	oil	0,0030
house	0,0027	military	0,0025	house	0,0022	news	0,0022	east	0,0028
military	0,0023	meeting	0,0025	africa	0,0022	court	0,0022	economic	0,0027
dont	0,0022	troops	0,0024	national	0,0021	law	0,0022	president	0,0027
committee	0,0022	north	0,0023	friday	0,0021	week	0,0022	west	0,0026
spokesman	0,0021	wednesday	0,0023	congress	0,0021	political	0,0022	rate	0,0025
percent	0,0021	south	0,0023	trade	0,0021	dont	0,0021	workers	0,0024
foreign	0,0020	spokesman	0,0021	bill	0,0020	city	0,0021	american	0,0024
city	0,0020	war	0,0021	nations	0,0020	president	0,0020	report	0,0023
officials	0,0020	west	0,0020	war	0,0020	bush	0,0019	union	0,0022
bill	0,0019	tuesday	0,0020	military	0,0019	police	0,0018	economy	0,0022

Table 4.3: Topics extracted from AP by HTMM

NSF dataset contains the abstracts of research proposals awarded by the National Science Foundation. Table 4.4 shows the topic distributions discovered by SBTM. SBTM detect the topics those are referred in the NSF dataset. "Topic 2" refers to "molecular biology" with the words like "protein", "cell", "molecular", "gene". Another topic "Topic 4" is about "mathematics" where it contains the words "equations", "mathematical", "numerical", "dimensional". Distinctly from the AP dataset, all topics are related with a single scientific topic among academic fields.

The topics discovered by LDA model are illustrated in Table 4.5. Again, LDA shows a good performance. The "molecular biology" topic is included in "Topic 7", "mathematics" in "Topic 9", "evolution" in "Topic 2" and so on. Also we can say that, all topic distributions of two models overlap with almost all words among the most likely 20 ones. Therefore, we can count LDA just as successful as SBTM on discovering the topic distributions.

Table 4.6 shows the poor performance of HTMM as shown in the AP dataset. Again, HTMM cannot extract clear, apparent and semantic topic distributions. Therefore, it prevents to estimate topic assignments of the sentences of a sample document.

We pick an abstract about "chemistry" from NSF dataset and estimate the topics of words/sentences of that document by the competitive models. Again, we color each word/sentence according to the topic it is assigned. For example, we color sentences assigned to "Topic 1" to "red" for SBTM and words assigned to "Topic 8" to "purple" for LDA.

SBTM assigns three different of topics to sentences of document. The sentences about the "chemical compound" of the "materials" and the "analytical" measures are assigned to "chemistry" topic. In a number of sentences, the "research", "educational plan" and courses of "Professor Collins" are referred and SBTM annotates those sentences with "student affairs and academic issues" topic. Also a sentence that refers to "student interest" on the "chemical analysis" is assigned to "social activity" topic.

Correspondingly, Figure 4.14 shows topics assigned to the words of the same NSF abstract by LDA. It is obvious that most of the word are assigned to plausible topics. The "chemistry" topic is used repeated with the proper words like "electroanalytical", "organic", "materials". But meanwhile, LDA suffers again from word disambiguation problem. The topical assignment "Cyclic voltammetry (CV)" can be counted among the most striking examples. LDA assigns "chemistry" topic to both "cyclic" and "voltammetry" words, successfully. But, it also assigns "social activity" topic where it misperceives "CV" as "curriculum vitae". Other problems like assigning "modified" directly to "molecular biology" without considering the context or splitting proper nouns into words are occurred again as encountered in AP dataset.

Topic 0		Topic 1		Topic 2		Topic 3		Topic 4	
research	0,0112	research	0,0079	protein	0,0105	research	0,0082	theory	0,0120
systems	0,0096	chemistry	0,0058	proteins	0,0095	data	0,0079	methods	0,0078
data	0,0092	molecular	0,0057	cell	0,0095	study	0,0048	systems	0,0072
system	0,0078	chemical	0,0056	cells	0,0076	provide	0,0042	research	0,0069
design	0,0077	studies	0,0055	gene	0,0070	project	0,0041	models	0,0069
control	0,0066	reactions	0,0051	genes	0,0067	region	0,0040	equations	0,0069
project	0,0061	molecules	0,0051	dna	0,0055	time	0,0039	study	0,0061
based	0,0054	materials	0,0050	function	0,0052	sites	0,0039	project	0,0057
network	0,0053	properties	0,0050	specific	0,0049	ice	0,0038	analysis	0,0054
time	0,0050	metal	0,0049	studies	0,0048	archaeological	0,0036	mathematical	0,0053
information	0,0050	structure	0,0049	molecular	0,0047	site	0,0035	model	0,0047
applications	0,0046	study	0,0047	expression	0,0046	analysis	0,0032	time	0,0044
performance	0,0045	systems	0,0047	plant	0,0039	climate	0,0028	data	0,0039
analysis	0,0043	organic	0,0040	binding	0,0037	dr	0,0027	dimensional	0,0038
models	0,0042	phase	0,0039	role	0,0037	studies	0,0026	space	0,0037
development	0,0042	understanding	0,0039	research	0,0037	understanding	0,0026	applications	0,0037
algorithms	0,0040	project	0,0037	system	0,0036	field	0,0025	numerical	0,0037
develop	0,0039	surface	0,0037	project	0,0036	change	0,0024	nonlinear	0,0035
techniques	0,0038	using	0,0037	understanding	0,0035	changes	0,0023	techniques	0,0034
methods	0,0038	processes	0,0036	rna	0,0035	record	0,0023	algorithms	0,0033
Research Methods and Means		Chemistry		Biology		Research Site Visits		Mathematical Models	

Topic 5		Topic 6		Topic 7		Topic 8		Topic 9	
species	0,0088	research	0,0218	research	0,0341	research	0,0132	materials	0,0091
research	0,0075	students	0,0207	university	0,0224	project	0,0083	research	0,0085
study	0,0055	science	0,0156	program	0,0107	data	0,0066	properties	0,0064
data	0,0049	project	0,0115	award	0,0094	social	0,0062	optical	0,0049
studies	0,0043	program	0,0105	project	0,0083	study	0,0059	systems	0,0049
understanding	0,0043	education	0,0082	chemistry	0,0079	information	0,0044	magnetic	0,0048
project	0,0039	engineering	0,0074	support	0,0075	economic	0,0044	project	0,0043
processes	0,0038	graduate	0,0070	dr	0,0075	political	0,0038	using	0,0043
water	0,0037	undergraduate	0,0065	science	0,0061	understanding	0,0035	phase	0,0040
provide	0,0032	school	0,0056	students	0,0059	policy	0,0035	devices	0,0040
model	0,0031	development	0,0055	department	0,0055	development	0,0034	study	0,0037
plant	0,0031	university	0,0053	national	0,0048	model	0,0034	electron	0,0034
effects	0,0030	faculty	0,0052	laboratory	0,0047	analysis	0,0032	temperature	0,0034
populations	0,0030	mathematics	0,0050	center	0,0046	studies	0,0030	applications	0,0033
population	0,0028	technology	0,0049	engineering	0,0046	theory	0,0029	surface	0,0032
changes	0,0028	teachers	0,0048	nsf	0,0041	models	0,0029	field	0,0032
genetic	0,0028	training	0,0045	materials	0,0040	public	0,0027	system	0,0032
ocean	0,0026	student	0,0044	graduate	0,0039	children	0,0027	structures	0,0031
time	0,0026	activities	0,0043	institute	0,0037	science	0,0027	low	0,0031
dynamics	0,0025	learning	0,0043	training	0,0034	time	0,0026	structure	0,0031
Life Sciences		Teaching		Human Resources for Research		Social Impact of Science		Research Materials	

Table 4.4: Topics extracted from NSF by SBTM

Electroanalytical Applications of Organically Modified Sol-Gel Materials.

This Faculty Early Career Development (CAREER) project, supported in the Analytical and Surface Chemistry Program, aims to explore and characterize the formation, properties and applications of organically modified sol-gels. The unique properties of these hybrid inorganic-organic materials remain to be exploited fully. The mechanism by which solutes become entrapped in these materials and the control of this process will be studied. Cyclic voltammetry (CV) with ultramicroelectrodes will be used to follow local physical and chemical changes that occur during hydrolysis and condensation of alkoxysilanes on the surfaces of these materials. CV and electrogenerated chemiluminescence will be used to characterize the mobility and accessibility of small charged redox probes entrapped in the sol-gel matrix under processing conditions. Practical results from this CAREER research project will focus on the development of permselective coatings for electroanalytical investigations and the fabrication of electrochemiluminscent sensors. Professor Collinson will combine these research thrusts with an educational plan that includes the development of a course in scientific ethics for both undergraduate and graduate students. A major revision of the laboratory component of an undergraduate course in chemical analysis is also planned. Student interest is to be cultivated by also employing more creative and realistic samples for analysis in this course with particular attention to forensic and environmental applications. The development of a fundamental understanding of the characteristics of the sol-gel matrix as an environment in which to do chemistry is an important objective of this CAREER research proposal. In the long term being able to tailor these matrices to a particular application using organically modified silicates will have strategic impact. The xerogels that are produced upon drying of sol-gels have interesting properties in their own right which could lead to useful applications in electronic, magnetic, optical materials and derived products. Professor Collinson's aim to introduce formally contemporary issues in scientific ethics and scientific conduct is timely, and this course will enable the students at Kansas State University to more clearly understand this topic.

	Topic 1 - Chemistry
	Topic 6 - Teaching
	Topic 8 - Social Impact of Science

Figure 4.13: Topics assigned to sentences of an NSF document by SBTM

Topic 0		Topic 1		Topic 2		Topic 3		Topic 4	
research	0,0132	species	0,0146	water	0,0092	experiments	0,0115	materials	0,0198
project	0,0112	data	0,0077	processes	0,0080	using	0,0109	properties	0,0123
data	0,0089	study	0,0074	data	0,0078	results	0,0072	chemical	0,0087
social	0,0087	provide	0,0062	climate	0,0059	experimental	0,0072	chemistry	0,0084
study	0,0078	research	0,0060	study	0,0058	data	0,0070	phase	0,0080
information	0,0069	population	0,0060	global	0,0054	field	0,0065	electron	0,0074
economic	0,0066	sites	0,0057	soil	0,0048	time	0,0060	surface	0,0073
policy	0,0051	human	0,0054	ice	0,0048	proposed	0,0056	magnetic	0,0067
political	0,0049	populations	0,0053	studies	0,0047	energy	0,0054	metal	0,0063
model	0,0048	evolution	0,0050	marine	0,0047	system	0,0053	structure	0,0060
understanding	0,0043	patterns	0,0049	ocean	0,0046	response	0,0053	studies	0,0059
knowledge	0,0042	site	0,0049	changes	0,0045	project	0,0051	molecular	0,0059
public	0,0040	dr	0,0044	model	0,0045	structures	0,0049	optical	0,0058
children	0,0038	natural	0,0044	time	0,0044	developed	0,0045	temperature	0,0057
people	0,0036	region	0,0044	carbon	0,0043	low	0,0043	systems	0,0054
human	0,0035	understanding	0,0040	sea	0,0043	provide	0,0042	electronic	0,0051
studies	0,0034	variation	0,0038	understanding	0,0043	study	0,0042	reactions	0,0050
issues	0,0034	diversity	0,0038	rates	0,0041	resolution	0,0042	molecules	0,0049
decision	0,0034	evolutionary	0,0037	project	0,0040	studies	0,0042	organic	0,0049
effects	0,0033	archaeological	0,0036	surface	0,0039	light	0,0041	understanding	0,0048
Social Impact of Research		A Mixed Topic		Marine Life		Experiments		Research Materials	

Topic 5		Topic 6		Topic 7		Topic 8		Topic 9	
systems	0,0179	research	0,1032	protein	0,0130	students	0,0290	theory	0,0169
design	0,0168	university	0,0263	cell	0,0116	science	0,0286	models	0,0148
control	0,0146	program	0,0165	proteins	0,0111	project	0,0152	methods	0,0142
system	0,0143	laboratory	0,0138	cells	0,0091	education	0,0129	systems	0,0097
based	0,0115	award	0,0124	dna	0,0084	program	0,0122	model	0,0088
data	0,0101	support	0,0098	molecular	0,0078	engineering	0,0107	mathematical	0,0086
performance	0,0098	center	0,0079	function	0,0075	technology	0,0091	study	0,0086
network	0,0089	training	0,0078	gene	0,0075	school	0,0083	analysis	0,0076
develop	0,0075	graduate	0,0078	specific	0,0073	learning	0,0078	equations	0,0062
applications	0,0074	department	0,0077	genes	0,0073	development	0,0076	project	0,0061
information	0,0074	dr	0,0075	studies	0,0063	student	0,0072	computational	0,0058
research	0,0068	scientists	0,0074	plant	0,0054	undergraduate	0,0071	nonlinear	0,0055
time	0,0063	development	0,0072	role	0,0049	activities	0,0071	investigator	0,0054
computer	0,0062	biology	0,0070	expression	0,0049	course	0,0067	dimensional	0,0053
development	0,0062	projects	0,0068	mechanisms	0,0048	mathematics	0,0067	numerical	0,0053
software	0,0061	researchers	0,0065	binding	0,0045	community	0,0064	space	0,0053
networks	0,0058	equipment	0,0065	structure	0,0041	programs	0,0063	dynamics	0,0050
developed	0,0054	chemistry	0,0064	regulation	0,0041	educational	0,0063	time	0,0049
proposed	0,0053	environmental	0,0059	determine	0,0041	faculty	0,0062	applications	0,0048
techniques	0,0053	engineering	0,0057	genetic	0,0040	career	0,0060	complex	0,0048
Research Methods		Human Resources for Research		Biology		Teaching		Mathematical Models	

Table 4.5: Topics extracted from NSF by LDA

Electroanalytical Applications of Organically Modified Sol-Gel Materials.

This Faculty Early Career Development (CAREER) project, supported in the Analytical and Surface Chemistry Program, aims to explore and characterize the formation, properties and applications of organically modified sol-gels. The unique properties of these hybrid inorganic-organic materials remain to be exploited fully. The mechanism by which solutes become entrapped in these materials and the control of this process will be studied. Cyclic voltammetry (CV) with ultramicroelectrodes will be used to follow local physical and chemical changes that occur during hydrolysis and condensation of alkoxysilanes on the surfaces of these materials. CV and electrogenerated chemiluminescence will be used to characterize the mobility and accessibility of small charged redox probes entrapped in the sol-gel matrix under processing conditions. Practical results from this CAREER research project will focus on the development of permselective coatings for electroanalytical investigations and the fabrication of electrochemiluminscent sensors. Professor Collinson will combine these research thrusts with an educational plan that includes the development of a course in scientific ethics for both undergraduate and graduate students. A major revision of the laboratory component of an undergraduate course in chemical analysis is also planned. Student interest is to be cultivated by also employing more creative and realistic samples for analysis in this course with particular attention to forensic and environmental applications. The development of a fundamental understanding of the characteristics of the sol-gel matrix as an environment in which to do chemistry is an important objective of this CAREER research proposal. In the long term being able to tailor these matrices to a particular application using organically modified silicates will have strategic impact. The xerogels that are produced upon drying of sol-gels have interesting properties in their own right which could lead to useful applications in electronic, magnetic, optical materials and derived products. Professor Collinson's aim to introduce formally contemporary issues in scientific ethics and scientific conduct is timely, and this course will enable the students at Kansas State University to more clearly understand this topic.

	Topic 0 - Social Impact of Research
	Topic 1 - A Mixed Topic
	Topic 2 - Marine Life
	Topic 3 - Experiments
	Topic 4 - Research Materials
	Topic 5 - Research Methods
	Topic 6 - Human Resources for Research
	Topic 7 - Biology
	Topic 8 - Teaching
	Topic 9 - Mathematical Models

Figure 4.14: Topics assigned to words of an NSF document by LDA

Topic 0		Topic 1		Topic 2		Topic 3		Topic 4	
continue	0,0078	research	0,0140	research	0,0221	research	0,0253	research	0,0177
research	0,0072	semiconductor	0,0108	project	0,0110	students	0,0214	university	0,0106
iterates	0,0058	essential	0,0063	data	0,0094	science	0,0144	data	0,0085
understanding	0,0042	properties	0,0054	science	0,0087	program	0,0115	linguistic	0,0067
modeled	0,0041	project	0,0050	iterates	0,0064	project	0,0112	project	0,0067
data	0,0040	university	0,0045	development	0,0056	education	0,0080	support	0,0055
electronic	0,0038	systems	0,0043	events	0,0056	engineering	0,0080	drainage	0,0046
project	0,0038	program	0,0043	scientific	0,0051	current	0,0080	east	0,0046
significant	0,0038	experimentally	0,0040	information	0,0047	undergraduate	0,0073	proof	0,0044
changes	0,0035	development	0,0040	university	0,0045	university	0,0072	perform	0,0044
furman	0,0034	insulator	0,0040	analysis	0,0038	hold	0,0058	program	0,0042
surrounding	0,0032	applications	0,0039	disciplines	0,0036	training	0,0053	international	0,0041
atmospheric	0,0029	european	0,0038	atmospheric	0,0034	teachers	0,0052	furman	0,0040
assays	0,0028	modeled	0,0037	foremost	0,0034	development	0,0050	names	0,0034
monte	0,0028	stressed	0,0037	publications	0,0034	ties	0,0048	science	0,0032
droplets	0,0028	body	0,0036	technology	0,0034	intensively	0,0046	iterates	0,0030
time	0,0028	contribute	0,0035	modeled	0,0033	commitment	0,0046	analysis	0,0029
differences	0,0028	using	0,0035	understanding	0,0033	technology	0,0045	physical	0,0029
excretion	0,0027	surface	0,0034	contributes	0,0032	workshops	0,0044	time	0,0027
involves	0,0027	exchange	0,0033	international	0,0030	furman	0,0044	development	0,0027

Topic 5		Topic 6		Topic 7		Topic 8		Topic 9	
research	0,0107	machinery	0,0107	data	0,0061	theory	0,0108	systems	0,0098
project	0,0070	changed	0,0095	using	0,0047	systems	0,0084	research	0,0098
data	0,0058	actively	0,0088	research	0,0043	research	0,0074	system	0,0080
events	0,0052	factors	0,0070	iterates	0,0042	methods	0,0068	data	0,0077
iterates	0,0046	packaged	0,0066	time	0,0041	iterates	0,0064	communications	0,0074
empirical	0,0046	visualize	0,0064	ferredoxin	0,0041	difference	0,0058	view	0,0073
information	0,0046	live	0,0058	surface	0,0038	models	0,0056	mediates	0,0059
skills	0,0045	assemble	0,0050	modeling	0,0036	project	0,0052	project	0,0055
models	0,0043	modeled	0,0049	properties	0,0036	mathematical	0,0045	based	0,0053
theory	0,0035	body	0,0049	map	0,0034	properties	0,0044	time	0,0052
postdoctoral	0,0034	specific	0,0048	modeled	0,0033	run	0,0042	algorithms	0,0050
time	0,0033	winning	0,0043	empirical	0,0031	empirical	0,0042	performance	0,0048
understanding	0,0033	slug	0,0039	models	0,0031	analysis	0,0042	applications	0,0046
characteristics	0,0031	research	0,0038	determined	0,0030	theoretical	0,0035	information	0,0045
analysis	0,0030	identified	0,0038	system	0,0030	time	0,0035	models	0,0043
involves	0,0028	structure	0,0038	systems	0,0030	dimensional	0,0035	techniques	0,0040
symmetry	0,0028	project	0,0036	project	0,0029	numerical	0,0035	boston	0,0039
development	0,0027	understanding	0,0036	composed	0,0028	structure	0,0034	methods	0,0039
modeled	0,0026	role	0,0035	furman	0,0028	applications	0,0034	analysis	0,0038
manipulate	0,0025	synthesis	0,0034	experiments	0,0027	superlattice	0,0034	develop	0,0037

Table 4.6: Topics extracted from NSF by HTMM

Chapter 5

Conclusion

The motivation behind the work presented in this thesis is to create a probabilistic graphical model that discovers the topics hidden in the text collections. The existing models (LSI, PLSI, LDA) focus on the same problem but they follow the bag-of-words assumption and neglect the topical correlation between terms in the same and closeby sentences. SBTM drops that assumption and exploits the semantic structure of a typical text, by modeling with a hidden Markov model the weak memory of topics across the successive sentences. The qualitative experiments show that SBTM extracts more meaningful topic distributions from text collections and annotate words in the same documents more consistently with their local meanings.

The evaluation results show the topic found by SBTM are more coherent, but LDA mixes up a few topics (like "Topic 2" on "UN/Middle East/Foreign Policy" and "Topic 7" on Cold War/Presidential Elections of topics extracted from AP dataset by LDA). Also, topic assignments to words and sentences made by LDA and SBTM, respectively, show SBTM's superiority qualitatively. Locally different meanings of the same terms (like "CV", "changes", "Washington" in the sample NSF document) are correctly identified by SBTM, but missed by LDA. On the other hand, HTMM fails to identify Markov topic transition probabilities, and this prevents it from building coherent topic distributions and topic-word assignments.

An author organizes his/her ideas into sections, each of which is further divided into paragraphs, each of which consists of related sentences. SBTM tries to capture physics underlying the writing habits and styles of authors. To achieve that idea, SBTM takes advantage of the sentence structure of a document. As a future work, new models to be developed can benefit from other type of semantic structures such as topically interconnect paragraphs, sections, chapters. Further segmentations within a sentence, paragraph, and document into semantic and syntactic elements can also increase the accuracy of topics discovered by the new models. Since text collections take the largest share in the data world, topic modeling will still stand among the most trend research topics.

Bibliography

[1] C. M. Bishop. *Pattern recognition and machine learning*. Information Science and Statistics. Springer, New York, 2006.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):138, 1977

[3] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391407, 1990.

[4] T. K. Landauer and S. T. Dumais. A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211 240, 1997

[5] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR 99, pages 5057, New York, NY, USA, 1999. ACM.

[6] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42:177196, January 2001.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:9931022, March 2003.

[8] William M. Darling. A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling. December, 2011.

[9] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. In *In Advances in Neural Information Processing Systems 17*, volume 17, pages 537544, 2005.

[10] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):52285235, 2004.

[11] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum. Topics in semantic representation. *Psychological Review*, 114(2):211 244, 2007.

[12] M. Steyvers and T. Griffiths. *Probabilistic Topic Models*. Lawrence Erlbaum Associates, 2007.

[13] D. Blei, L. Carin, and D. Dunson. Probabilistic topic models. *Signal Processing Magazine, IEEE*, 27(6):5565, 2010.

[14] D. Blei and J. Lafferty. *Text Mining: Theory and Applications*, chapter Topic models. Taylor and Francis, 2009.

[15] T. M. Department, T. Minka, and J. Lafferty. Expectation-propagation for the generative aspect model. In *In Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 352359. Morgan Kaufmann, 2002.

[16] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov chain Monte Carlo in practice*. Interdisciplinary Statistics. Chapman and Hall, London, 1996.

[17] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum. Topics in semantic representation. *Psychological Review*, 114(2):211 244, 2007

[18] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183233, 1999. 10.1023/A:1007665907178.

[19] M. Andrews, G. Vigliocco. The Hidden Markov Topic Model: A Probabilistic Model of Semantic Representation. *Topics in Cognitive Science 2*, (2010) 101-113.

[20] A. Gruber, M. Rosen-Zvi and Y. Weiss. Hidden Topic Markov Models. *Artificial Intelligence and Statistics (AISTATS)*, San Juan, Puerto Rico, March 2007.

[21] Heinrich, Gregor. 2008. Parameter estimation for text analysis. Version 2.4. Unpublished note.

[22] T. Griffiths. Gibbs sampling in the generative model of Latent Dirichlet Allocation.

[23] B. Carpenter. Integrating Out Multinomial Parameters in Latent Dirichlet Allocation and Naive Bayes for Collapsed Gibbs Sampling. Revision 1.4, September 27, 2010.

[24] A Highly Accurate Sentence and Paragraph Breaker, `http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector`.

Appendix A

Full conditional distributions for the collapsed Gibbs sampler

For the first sentence of each document, we define

$$\begin{aligned}
P(K_{m,1} = k | \underset{\sim}{K}^{-(m,1)}) = \underset{\sim}{k}^{-(m,1)}, \underset{\sim}{S} = \underset{\sim}{s}, \alpha, \beta, \gamma\} \\
&\propto P(K_{m,1} = k, \underset{\sim}{K}^{-(m,1)}) = \underset{\sim}{k}^{-(m,1)}, \underset{\sim}{S} = \underset{\sim}{s} | \alpha, \beta, \gamma\} \\
&\propto \frac{\Delta(\alpha + e_m^{-(m,1),k})}{\Delta(\alpha)} \prod_{l=1}^{K} \frac{\Delta(\beta + f_l^{-(m,1),k})}{\Delta(\beta)} \prod_{l=1}^{K} \frac{\Delta(\gamma + g_{m,l}^{-(m,1),k})}{\Delta(\gamma)} \\
&\propto \Delta(\alpha + e_m^{-(m,1),k}) \times \prod_{l=1}^{K} \left(\Delta(\beta + f_l^{-(m,1),k}) \Delta(\gamma + g_{m,l}^{-(m,1),k}) \right),
\end{aligned}$$

and for every sentences $t = 2, \ldots$ of each document, we define

$$\begin{aligned}
P(K_{m,t} = k | \underset{\sim}{K}^{-(m,t)}) = \underset{\sim}{k}^{-(m,t)}, \underset{\sim}{S} = \underset{\sim}{s}, \alpha, \beta, \gamma\} \\
&\propto P(K_{m,t} = k, \underset{\sim}{K}^{-(m,t)}) = \underset{\sim}{k}^{-(m,t)}, \underset{\sim}{S} = \underset{\sim}{s} | \alpha, \beta, \gamma\} \\
&\propto \prod_{l=1}^{K} \frac{\Delta(\beta + f_l^{-(m,t),k})}{\Delta(\beta)} \prod_{l=1}^{K} \frac{\Delta(\gamma + g_{m,l}^{-(m,t),k})}{\Delta(\gamma)} \\
&\propto \prod_{l=1}^{K} \left(\Delta(\beta + f_l^{-(m,t),k}) \Delta(\gamma + g_{m,l}^{-(m,t),k}) \right),
\end{aligned}$$

where f_l is the total count of words assigned to topic l, $e_{m,l}$ is the topic indicator of the first sentence of document m, and $g_{m,l}$ is the total count of topic transitions

out of topic l in document m, and

$$f_l^{-(m,t),k} = (f_{l,1}^{-(m,t),k}, \ldots, f_{l,N}^{-(m,t),k}),$$

where

$$f_{l,s}^{-(m,t),k} = \left\{ \begin{array}{ll} f_{l,s} & s \notin \underset{\sim}{s}_{m,t} \\ f_{l,s} + (1_{l,k} - 1_{l,k_{m,t}}) c_{m,t,s} & s \in \underset{\sim}{s}_{m,t} \end{array} \right\}, l = 1, \cdots, K,$$

$$\begin{aligned} e_{m,l}^{-(m,1),k} &= (e_{m,1}^{-(m,1),k}, \ldots, e_{m,K}^{-(m,1),k}) \\ &= e_{m,l} - 1_{l,k_{m,1}} + 1_{l,k} \text{ for } l = 1, \ldots, K \\ , g_{m,l}^{-(m,t),k} &= (g_{m,l,1}^{-(m,t),k}, \ldots, g_{m,l,K}^{-(m,t),k}), \end{aligned}$$

and

$$\begin{aligned} g_{m,l,\tilde{l}}^{-(m,t),k} =& g_{m,l,\tilde{l}} && \text{for } l \neq k_{m,t-1} \text{ and } \tilde{l} \neq k_{m,t+1}, \\ g_{m,l,k_{m,t+1}}^{-(m,t),k} =& g_{m,l,k_{m,t+1}} - 1_{l,k_{m,t}} + 1_{l,k} && \text{for } l = 1, \ldots, K, \\ g_{m,k_{m,t-1},l}^{-(m,t),k} =& g_{m,k_{m,t-1},l} - 1_{l,k_{m,t}} + 1_{l,k} && \text{for } l = 1, \ldots, K. \end{aligned}$$

Appendix B

Derivations of full conditional distributions in (3.1)-(3.6)

We have

$$P(K_{m,1} = k|\underset{\sim}{K}^{-(m,1)}) = \underset{\sim}{k}^{-(m,1)}, \underset{\sim}{S} = \underset{\sim}{s}, \alpha, \beta, \gamma\}$$
$$\propto \Delta(\alpha + e_m^{-(m,1),k}) \times \prod_{l=1}^{K} \left(\Delta(\beta + f_l^{-(m,1),k}) \Delta(\gamma + g_{m,l}^{-(m,1),k}) \right),$$

and for every sentences $t = 2, \ldots$ of each document, we define

$$P(K_{m,t} = k|\underset{\sim}{K}^{-(m,t)}) = \underset{\sim}{k}^{-(m,t)}, \underset{\sim}{S} = \underset{\sim}{s}, \alpha, \beta, \gamma\}$$
$$\propto \prod_{l=1}^{K} \left(\Delta(\beta + f_l^{-(m,t),k}) \Delta(\gamma + g_{m,l}^{-(m,t),k}) \right),$$

where f_l is the total number of sentences assigned to topic l in the entire

collection, and

$$
\begin{aligned}
&\Delta(\beta + f_l^{-(m,t),k}) \\
&\quad = \frac{\prod_{s=1}^{N} \Gamma(\beta_s + f_{l,s}^{-(m,t),k})}{\Gamma(\sum_{s=1}^{N}(\beta_s + f_{l,s}^{-(m,t),k}))} \\
&\quad = \frac{\left(\prod_{s \notin s_{m,t}} \Gamma(\beta_s + f_{l,s})\right)\left(\prod_{s \in s_{m,t}} \Gamma(\beta_s + f_{l,s} + (1_{l,k} - 1_{l,k_{m,t}})c_{m,t,s})\right)}{\Gamma\left(\sum_{s \notin s_{m,t}} (\beta_s + f_{l,s}) + \sum_{s \in s_{m,t}} (\beta_s + f_{l,s} + (1_{l,k} - 1_{l,k_{m,t}})c_{m,t,s})\right)} \\
&\quad = \frac{\prod_{s=1}^{N} \Gamma(\beta_s + f_{l,s})}{\Gamma(\sum_{s=1}^{N}(\beta_s + f_{l,s}))} \cdot \frac{\Gamma(\sum_{s=1}^{N}(\beta_s + f_{l,s}))}{\prod_{s \in s_{m,t}} \Gamma(\beta_s + f_{l,s})} \\
&\qquad \times \frac{\prod_{s \in s_{m,t}} \Gamma(\beta_s + f_{l,s} + (1_{l,k} - 1_{l,k_{m,t}})c_{m,t,s})}{\Gamma(\sum_{s=1}^{N}(\beta_s + f_{l,s}) + \sum_{s \in s_{m,t}} (1_{l,k} - 1_{l,k_{m,t}})c_{m,t,s})} \\
&\quad = \Delta(\beta + f_l) \prod_{s \in s_{m,t}} \frac{\Gamma(\beta_s + f_{l,s} + (1_{l,k} - 1_{l,k_{m,t}})c_{m,t,s})}{\Gamma(\beta_s + f_{l,s})} \\
&\qquad \times \frac{\Gamma(\sum_{s=1}^{N}(\beta_s + f_{l,s}))}{\Gamma(\sum_{s=1}^{N}(\beta_s + f_{l,s}) + (1_{l,k} - 1_{l,k_{m,t}}) \sum_{s \in s_{m,t}} c_{m,t,s})}.
\end{aligned}
$$

If $k = k_{m,t}$, then

$$
\begin{aligned}
f_l^{-(m,t),k_{m,t}} &= f_l, \\
\Delta(\beta + f_l^{-(m,t),k_{m,t}}) &= \Delta(\beta + f_l)
\end{aligned}
$$

for every $l = 1, \ldots, \kappa$.

If $k \neq k_{m,t}$, then

$$\Delta(\beta + f_l^{-(m,t),k}) = \begin{cases} \Delta(\beta + f_k) \dfrac{\prod\limits_{s \in \underset{\sim}{s}_{m,t}} P_{c_{m,t,s}}^{\beta_s + f_{k,s} - 1 + c_{m,t,s}}}{P_{N_{m,t}}^{\sum_{s=1}^{N} (\beta_s + f_{k,s}) - 1 + N_{m,t}}}, & l = k \\ \Delta(\beta + f_{k_{m,t}}) \dfrac{P_{N_{m,t}}^{\sum_{s=1}^{N} (\beta_s + f_{k_{m,t},s}) - 1}}{\prod\limits_{s \in \underset{\sim}{s}_{m,t}} P_{c_{m,t,s}}^{\beta_s + f_{k_{m,t},s} - 1}}, & l = k_{m,t} \\ \Delta(\beta + f_l), & \text{otherwise} \end{cases}$$

Moreover,$g_{m,l}$ is the topic transition counts from sentence l, and for $l \neq k_{m,t-1}$,

$$\begin{aligned} &\Delta(\gamma + g_{m,l}^{-(m,t),k}) \\ &= \frac{\prod_{\tilde{l}=1}^{K} \Gamma(\gamma_{\tilde{l}} + g_{m,l,\tilde{l}}^{-(m,t),k})}{\Gamma(\sum_{\tilde{l}=1}^{K} (\gamma_{\tilde{l}} + g_{m,l,\tilde{l}}^{-(m,t),k}))} \\ &= \frac{\left(\prod_{\tilde{l} \neq k_{m,t+1}} \Gamma(\gamma_{\tilde{l}} + g_{m,l,\tilde{l}}) \right) \Gamma\left(\underset{\sim}{\gamma}_{k_{m,t+1}} + g_{m,l,k_{m,t+1}} - 1_{l,k_{m,t}} + 1_{l,k} \right)}{\Gamma\left(\sum_{\tilde{l} \neq k_{m,t+1}} (\gamma_{\tilde{l}} + g_{m,l,\tilde{l}}) + \gamma_{k_{m,t+1}} + g_{m,l,k_{m,t+1}} - 1_{l,k_{m,t}} + 1_{l,k} \right)} \\ &= \frac{\prod_{\tilde{l}=1}^{K} \Gamma(\gamma_{\tilde{l}} + g_{m,l,\tilde{l}})}{\Gamma(\sum_{\tilde{l}=1}^{K} (\gamma_{\tilde{l}} + g_{m,l,\tilde{l}}))} \cdot \frac{\Gamma(\tilde{\gamma}_{k_{m,t+1}} + g_{m,l,k_{m,t+1}} - 1_{l,k_{m,t}} + 1_{l,k})}{\Gamma(\tilde{\gamma}_{k_{m,t+1}} + g_{m,l,k_{m,t+1}})} \\ &\quad \times \frac{\Gamma(\sum_{\tilde{l}=1}^{K} (\gamma_{\tilde{l}} + g_{m,l,\tilde{l}}))}{\Gamma(\sum_{\tilde{l}=1}^{K} (\gamma_{\tilde{l}} + g_{m,l,\tilde{l}}) - 1_{l,k_{m,t}} + 1_{l,k})} \end{aligned}$$

$$= \Delta(\tilde{\gamma} + g_{m,l})(\gamma_{k_{m,t+1}} + g_{m,l,k_{m,t+1}} - 1_{l,k_{m,t}})^{1_{l,k}-1_{l,k_{m,t}}}$$
$$\times (\sum_{\tilde{l}=1}^{K}(\gamma_{\tilde{l}} + g_{m,l,\tilde{l}}) - 1_{l,k_{m,t}})^{1_{l,k_{m,t}}-1_{l,k}}$$
$$= \Delta(\tilde{\gamma} + g_{m,l})\left(\frac{\gamma_{k_{m,t+1}} + g_{m,l,k_{m,t+1}} - 1_{l,k_{m,t}}}{\prod_{\tilde{l}=1}^{K}(\gamma_{\tilde{l}} + g_{m,l,\tilde{l}}) - 1_{l,k_{m,t}}}\right)^{1_{l,k}-1_{l,k_{m,t}}},$$

and for $l = k_{m,t-1}$,

$$\Delta(\gamma + g^{-(m,t),k}_{m,k_{m,t-1}}) = \frac{\prod_{\tilde{l}=1}^{K}\Gamma(\gamma_{\tilde{l}} + g^{-(m,t),k}_{m,k_{m,t-1},\tilde{l}})}{\Gamma(\sum_{\tilde{l}=1}^{K}(\gamma_{\tilde{l}} + g^{-(m,t),k}_{m,k_{m,t-1},\tilde{l}}))}.$$

For $k = k_{m,t}$, we have

$$g^{-(m,t),k_{m,t}}_{m,k_{m,t-1},\tilde{l}} = g_{m,k_{m,t-1},\tilde{l}}$$

for every $\tilde{l} = 1, \cdots, \kappa$ and

$$\Delta(\gamma + g^{-(m,t),k_{m,t}}_{m,k_{m,t-1}}) = \Delta(\gamma + g_{m,k_{m,t-1}}).$$

Now, for $k \neq k_{m,t}$,

$$\Delta(\gamma + g^{-(m,t),k}_{m,k_{m,t-1}})$$
$$= \frac{\left(\prod_{\tilde{l}\notin\{k_{m,t},k\}}\Gamma(\gamma_{\tilde{l}} + g_{m,k_{m,t-1},\tilde{l}})\right)\Gamma\left(\gamma_{k_{m,t}} + g_{m,k_{m,t-1},k_{m,t}} - 1\right)\Gamma\left(\gamma_k + g_{m,k_{m,t-1},k} + 1\right)}{\Gamma\left(\sum_{\tilde{l}\notin\{k_{m,t},k\}}(\tilde{\gamma}_{\tilde{l}} + g_{m,k_{m,t-1},\tilde{l}}) + \gamma_{k_{m,t}} + g_{m,k_{m,t-1},k_{m,t}} - 1 + \gamma_k + g_{m,k_{m,t-1},k} + 1\right)}$$
$$= \frac{\prod_{\tilde{l}=1}^{K}\Gamma(\gamma_{\tilde{l}} + g_{m,k_{m,t-1},\tilde{l}})}{\Gamma(\sum_{\tilde{l}=1}^{K}(\tilde{\gamma}_{\tilde{l}} + g_{m,k_{m,t-1},\tilde{l}}))} \cdot \frac{\Gamma(\gamma_k + g_{m,k_{m,t-1},k} + 1)}{\Gamma(\gamma_k + g_{m,k_{m,t-1},k})} \cdot \frac{\Gamma(\gamma_{k_{m,t}} + g_{m,k_{m,t-1},k_{m,t}} - 1)}{\Gamma(\gamma_{k_{m,t}} + g_{m,k_{m,t-1},k})}$$
$$= \Delta(\gamma + g_{m,k_{m,t-1}})\frac{\gamma_k + g_{m,k_{m,t-1},k}}{\gamma_{k_{m,t}} + g_{m,k_{m,t-1},k_{m,t}} - 1}$$
$$= \Delta(\gamma + g_{m,k_{m,t-1}})\frac{\gamma_k + g_{m,k_{m,t-1},k} - 1_{k,k_{m,t}}}{\gamma_{k_{m,t}} + g_{m,k_{m,t-1},k_{m,t}} - 1}.$$

The last term $e_m^{-(m,1),k}$ determines the topic assignment of the first sentence. For $k = k_{m,1}$,

$$e_m^{-(m,1),k_{m,1}} = e_m = (0, \cdots, 0, 1, 0, \cdots, 0),$$

where only $k_{m,1}$th place in e_m distribution equals one. For $k \neq k_{m,1}$,

$$e_m^{-(m,1),k} = (0, \cdots, 0, 1, 0, \cdots, 0),$$

where only kth place in e_m distribution equals one. Then

$$\begin{aligned}
&\Delta(\alpha + e_m^{-(m,1),k}) \\
&\quad = \Delta(\alpha + e_m)\frac{\Delta(\alpha + e_m^{-(m,1),k})}{\Delta(\alpha + e_m)} \\
&\quad = \Delta(\alpha + e_m)\frac{\Gamma(\alpha_k)\left(\prod_{l \neq k}\Gamma(\alpha_l)\right)\Gamma(\alpha_k + 1)}{\Gamma(\alpha_k)\Gamma(\sum_{l \neq k}\alpha_l + \alpha_k + 1)} \cdot \frac{\Gamma(\alpha_{k_{m,1}})\Gamma(\sum_{l \neq k_{m,1}}\alpha_l + \alpha_{k_{m,1}} + 1)}{\Gamma(\alpha_{k_{m,1}})\left(\prod_{l \neq k_{m,1}}\Gamma(\alpha_l)\right)\Gamma(\alpha_{k_{m,1}} + 1)} \\
&\quad = \Delta(\alpha + e_m)\frac{\Gamma(\alpha_k + 1)}{\Gamma(\alpha_k)} \cdot \frac{\Gamma(\alpha_{k_{m,1}})}{\Gamma(\alpha_{k_{m,1}} + 1)} \\
&\quad = \Delta(\alpha + e_m)\frac{\alpha_k}{\alpha_{k_{m,1}}}.
\end{aligned}$$

The equations (3.1)-(3.6) are obtained by replacing those three terms with the proper derivations above according to the location of the desired sentence in the document.

Also regarding the variables

$$\begin{aligned}
e_m^{-(m,1),k_{m,1}} &= e_m, \\
f_l^{-(m,t),k_{m,t}} &= f_l, \\
g_{m,l}^{-(m,t),k_{m,t}} &= g_{m,l},
\end{aligned}$$

we have

$$\begin{aligned}
P(K_{m,1} = k_{m,1} | \underset{\sim}{K}^{-(m,1)}) &= \underset{\sim}{k}^{-(m,1)}, \underset{\sim}{S} = \underset{\sim}{s}, \alpha, \beta, \gamma\} \\
&\propto \Delta(\alpha + e_m) \times \prod_{l=1}^{K}\left(\Delta(\beta + f_l)\Delta(\gamma + g_{m,l})\right)
\end{aligned}$$

and for $t = 2, \ldots, T_m$ of each document

$$P(K_{m,t} = k_{m,t} | \underset{\sim}{K}^{-(m,t)}) = \underset{\sim}{k}^{-(m,t)}, \underset{\sim}{S} = \underset{\sim}{s}, \alpha, \beta, \gamma\}$$
$$\propto \prod_{l=1}^{K} \left(\Delta(\beta + f_l) \Delta(\gamma + g_{m,l}) \right).$$