Latent Probabilistic Topic Discovery for Text Documents Incorporating Segment Structure and Word Order

Shoaib Jameel

The Chinese University of Hong Kong
Department of Systems Engineering and Engineering Management

ADVISOR
Prof. LAM, Wai

June 23, 2014

One line summary of the thesis

Shows how maintaining the document structure such as paragraphs, sentences, and the word order helps improve the performance of topic models.

Contents

What do you do when you have these many pages?

Did you know?

The Indexed Web contains at least 4.96 billion pages (as of Wednesday, 11 June, 2014). – World-WideWebSize.com

4.96 billion

Take each one of them and read?

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

Nulla in ipsum. Praesent eros nulla, congue vitae, euismod ut, commodo a, wisi. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Aenean nonummy magna non leo. Sed felis erat, ullamcorper in, dictum non, ultricies ut, lectus. Proin vel arcu a odio lobortis euismod. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Proin ut est. Aliquam odio. Pellentesque massa turpis, cursus eu, euismod nec, tempor congue, nulla. Duis viverra gravida mauris. Cras tincidunt. Curabitur eros ligula, varius ut, pulvinar in, cursus faucibus, augue.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

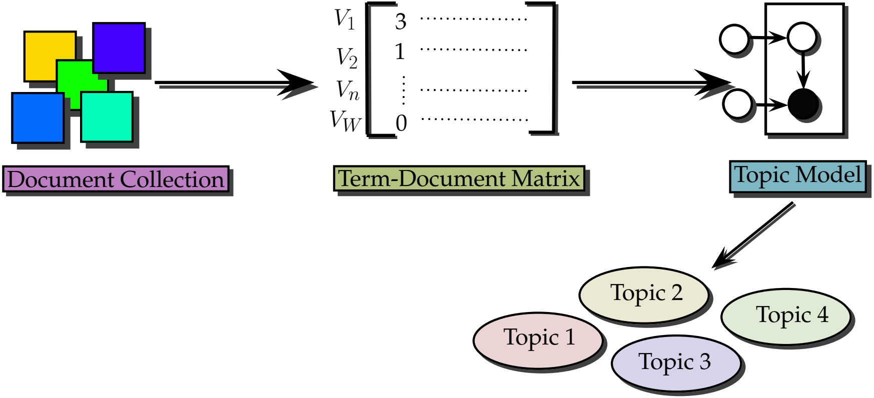
Remember!!

4.96 billion documents on the web.

Problem!!!

Even reading a small subset of such a huge collection is impossible for a human.

Or get gist of the data using statistical techniques



Word Overlaps

Example of text data: Titles of Some Technical Memos

c1:	*Human* machine *interface* for ABC *computer* applications
c2:	A *survey* of *user* opinion of *computer system response time*
c3:	The *EPS user interface* management *system*
c4:	*System* and *human system* engineering testing of *EPS*
c5:	Relation of *user* perceived *response time* to error measurement
m1:	The generation of random, binary, ordered *trees*
m2:	The intersection *graph* of paths in *trees*
m3:	*Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4:	*Graph minors*: A *survey*

What is so great about statistical techniques?

Term document matrix - High dimensional vector space.

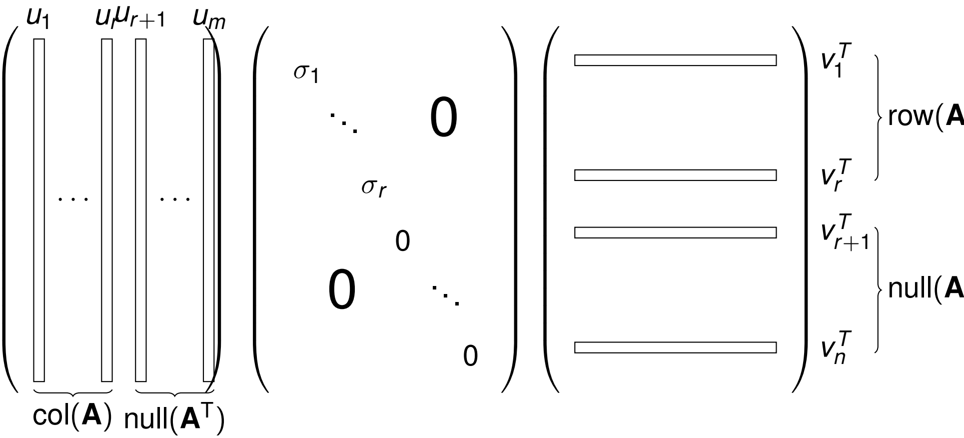
$$\begin{array}{c} \\ v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_W \end{array} \begin{array}{c} \begin{array}{ccccc} d_1 & d_2 & \cdot & \cdot & d_D \end{array} \\ \begin{pmatrix} 8 & 1 & 1 & 1 & 4 \\ 5 & 12 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Notations

W - Number of words in the vocabulary.
D - Number of documents in the collection.

Term document matrix, **A**, using Singular Value Decomposition is decomposed as:

$$\underbrace{\mathbf{A}}_{W\times D} = \underbrace{\mathbf{U}}_{W\times W} \times \underbrace{\mathbf{S}}_{W\times D} \times \underbrace{\mathbf{V}^{\mathrm{T}}}_{D\times D}$$

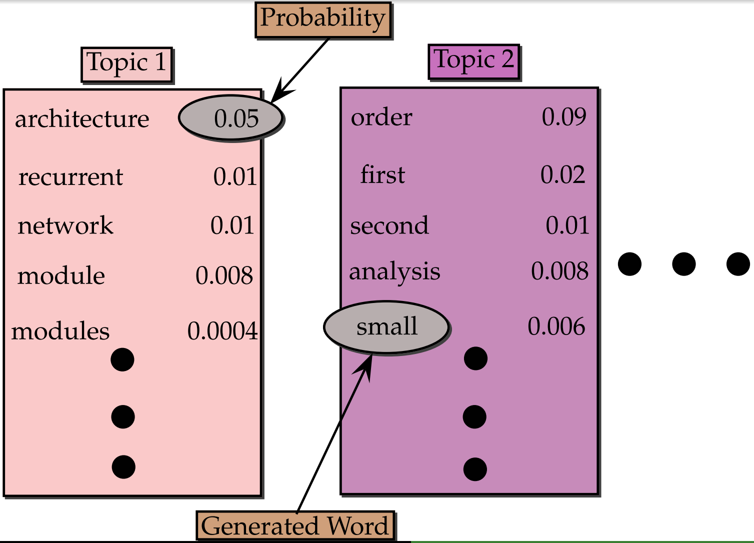


Topic Model as Matrix Factorization

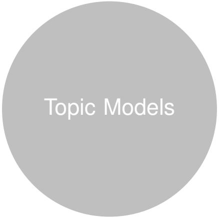
$$\text{Terms}\left\{\overbrace{\begin{array}{cc} & \begin{array}{ccc} \quad k_1 & \quad k_2 & \quad k_3 \end{array} \\ \begin{array}{c} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_W \end{array} & \begin{pmatrix} 1.00 & 0.91 & 1.00 \\ 0.44 & 0.57 & 0.84 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 0.00 & 0.00 & 0.47 \end{pmatrix} \end{array}}^{\text{Topics}}\right. \approx P(\mathbf{w}|\mathbf{z})$$

$$\text{Topics}\left\{\overbrace{\begin{array}{cc} & \begin{array}{ccccc} \quad d_1 & \quad d_2 & \cdot & \cdot & \quad d_D \end{array} \\ \begin{array}{c} k_1 \\ k_2 \\ k_3 \end{array} & \begin{pmatrix} 0.19 & 0.05 & \cdot & \cdot & 0.10 \\ 0.01 & 0.43 & \cdot & \cdot & 0.52 \\ 0.03 & 0.45 & \cdot & \cdot & 0.64 \end{pmatrix} \end{array}}^{\text{Documents}}\right. \approx P(\mathbf{d}|\mathbf{z})$$

This is how it works

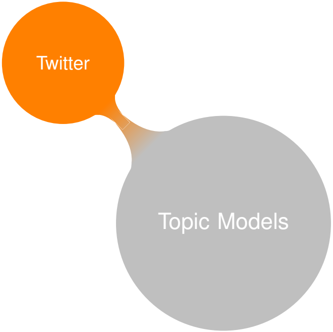


Applications of Topic Models



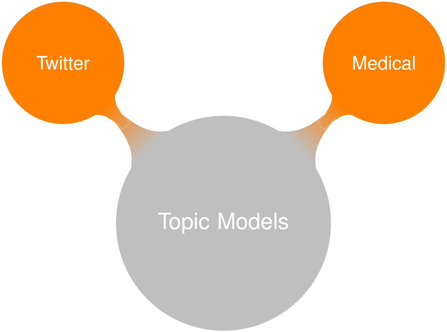
- Social Networks - Finding popular nodes in a graph.
- Gene expression analysis - Highlight the relationship between cell types, cellular processes, and gene expression.
- Image analysis - Image annotation.

Applications of Topic Models



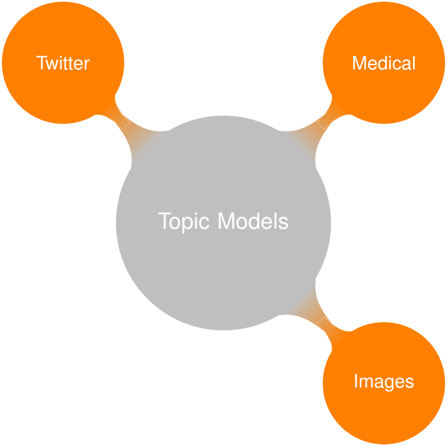
- Social Networks - Finding popular nodes in a graph.
- Gene expression analysis - Highlight the relationship between cell types, cellular processes, and gene expression.
- Image analysis - Image annotation.

Applications of Topic Models



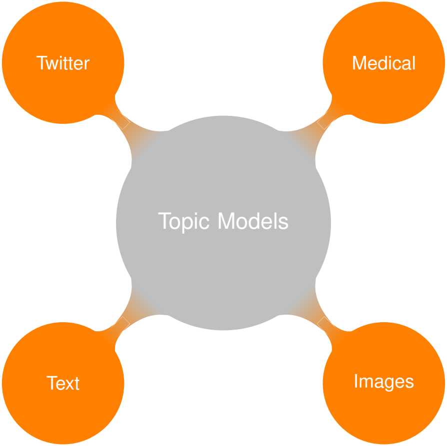
- Social Networks - Finding popular nodes in a graph.
- Gene expression analysis - Highlight the relationship between cell types, cellular processes, and gene expression.
- Image analysis - Image annotation.

Applications of Topic Models



- Social Networks - Finding popular nodes in a graph.
- Gene expression analysis - Highlight the relationship between cell types, cellular processes, and gene expression.
- Image analysis - Image annotation.

Applications of Topic Models



- Social Networks - Finding popular nodes in a graph.
- Gene expression analysis - Highlight the relationship between cell types, cellular processes, and gene expression.
- Image analysis - Image annotation.

But!!!!

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
architecture	order	connectionist	potential	prior
recurrent	first	role	membrane	bayesian
network	second	binding	current	data
module	analysis	structures	synaptic	evidence
modules	small	distributed	dendritic	experts

So what is the problem above?

Words in topics are not insightful.

Latent Dirichlet Allocation Model (LDA) [?]

Generative Process

1. Draw $\theta^{(d)}$ from **Dirichlet**(α), where each $\theta^{(d)}$ consists of topic distribution for document *d*
2. Draw ϕ from **Dirichlet**(β), where ϕ encompasses word distribution for topic
3. For every word in the document *d*
   1. Draw a topic $z_i^{(d)}$ from **Multinomial** $(\theta^{(d)})$
   2. Draw a word $w_i^{(d)}$ from **Multinomial** $(\phi_{z_i^{(d)}})$

Graphical Model

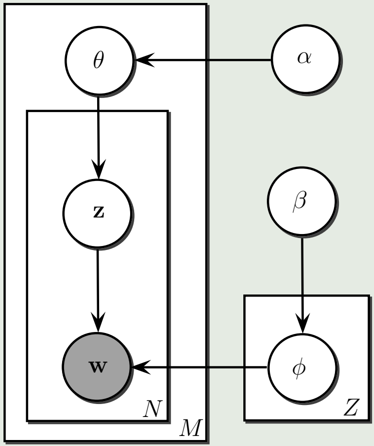
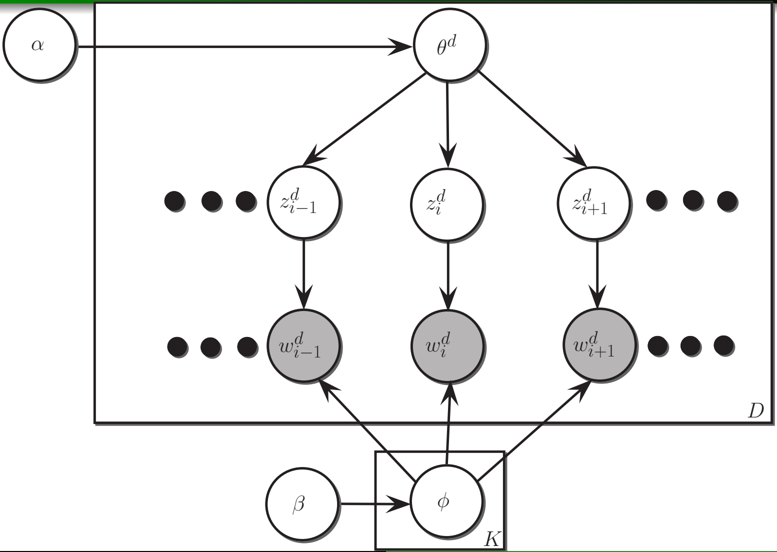


Plate Diagrams

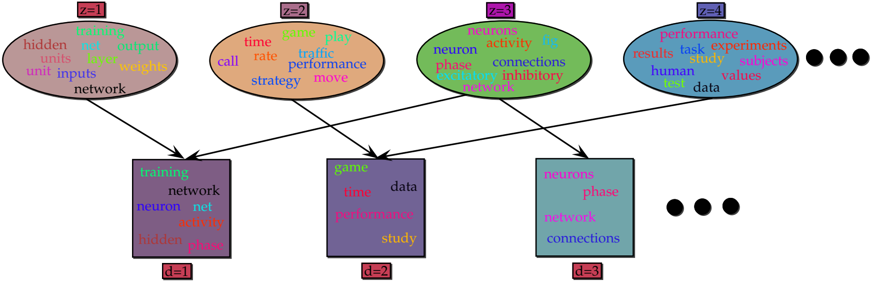


Topic Model as Matrix Factorization

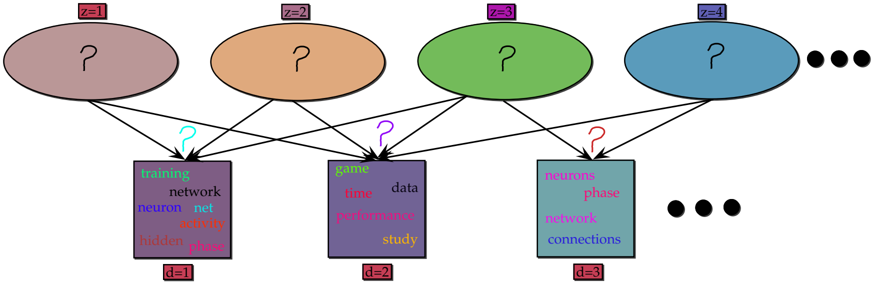
$$\text{Terms}\left\{\overbrace{\begin{array}{c c} & \begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\ \begin{array}{c} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_W \end{array} & \begin{pmatrix} 1.00 & 0.91 & 1.00 \\ 0.44 & 0.57 & 0.84 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 0.00 & 0.00 & 0.47 \end{pmatrix} \end{array}}^{\text{Topics}}\right. \approx P(\mathbf{w}|\mathbf{z})$$

$$\text{Topics}\left\{\overbrace{\begin{array}{c c} & \begin{array}{ccccc} d_1 & d_2 & \cdot & \cdot & d_D \end{array} \\ \begin{array}{c} k_1 \\ k_2 \\ k_3 \end{array} & \begin{pmatrix} 0.19 & 0.05 & \cdot & \cdot & 0.10 \\ 0.01 & 0.43 & \cdot & \cdot & 0.52 \\ 0.03 & 0.45 & \cdot & \cdot & 0.64 \end{pmatrix} \end{array}}^{\text{Documents}}\right. \approx P(\mathbf{d}|\mathbf{z})$$

What does a topic model do? - Generation Process



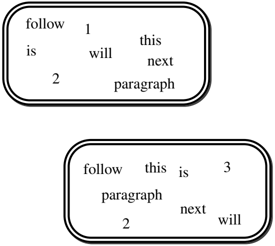
What does a topic model do? - Inference Process



Bag of Words in Topic Segmentation

- These models maintain document structure such as paragraphs or sentences.
- Assume that words within a segment (paragraph or a sentence) are exchangeable.
- Introduces the notion of super-topics and word-topics

Paragraph n in the document d



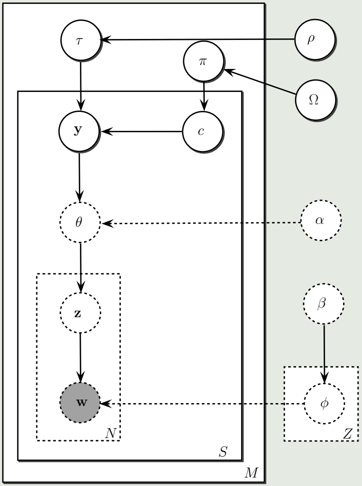
Paragraph $n + 1$ in the document d

A Topic Segmentation Model (LDSEG) [?]

Model Properties

- Performs topic segmentation
- Can work at paragraph and sentence level
- **c** a binary variable gives the change in topics segment-wise
- Segments come from a predefined number of super-topics
- The super-topics comprise of a mixture of word-topics

Graphical Model

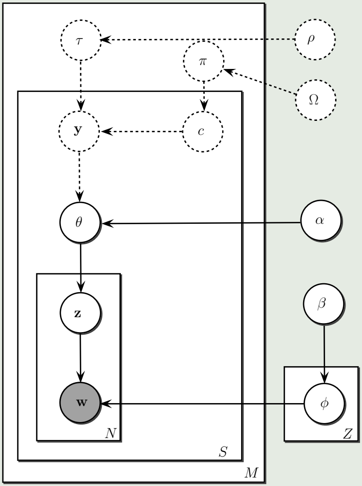


A Topic Segmentation Model (LDSEG)

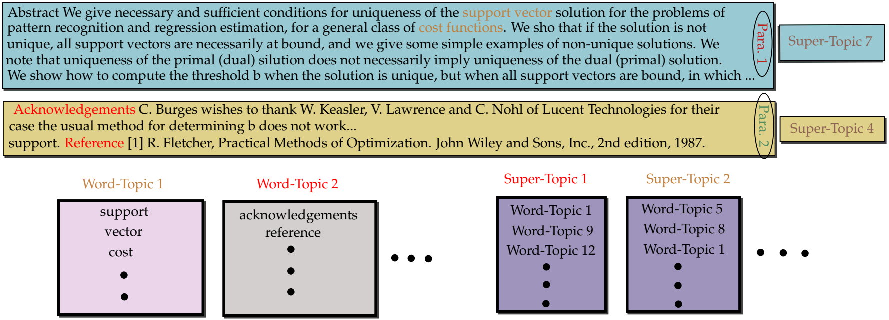
Model Properties

- This region is similar to the LDA model
- Segments exhibit multiple topics
- Words are generated from a predefined number of word-topics

Graphical Model



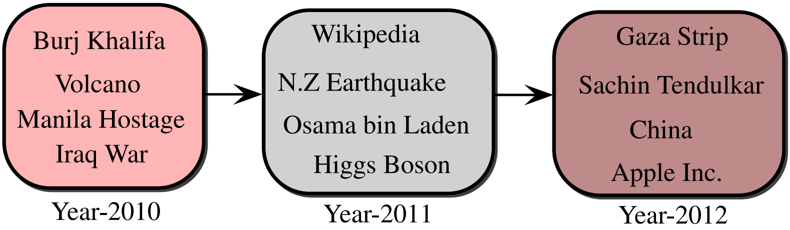
Topic Segmentation Illustration - LDSEG



- Performs topic segmentation
- Unigram words are assigned to the word-topics
- Segments are assigned to the document-topics or super-topics

Why capture topics over time?

1. We know that data evolves over time.
2. What people are talking today may not be talking tomorrow or an year after.



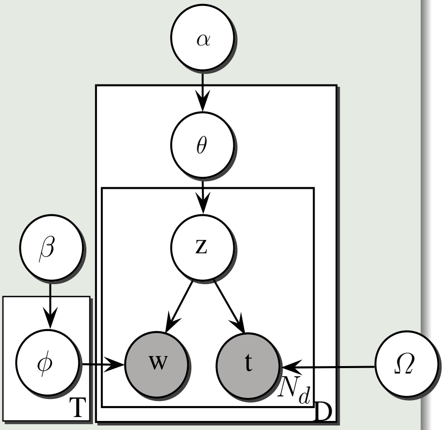
3. Models such as LDA cannot capture such temporal characteristics in data.

Topics Over Time (TOT) [?]

Generative Process

1. Draw T multinomials ϕ_z from a Dirichlet Prior β, one for each topic z
2. For each document d, draw a multinomial $\theta^{(d)}$ from a Dirichlet prior α; then for each word $w_i^{(d)}$ in the document d
   1. Draw a topic z_i^d from Multinomial $\theta^{(d)}$
   2. Draw a word $w_i^{(d)}$ from Multinomial $\phi_{z_i^{(d)}}$
   3. Draw a timestamp $t_i^{(d)}$ from Beta $\Omega_{z_i^{(d)}}$

Topics Over Time Model (TOT)

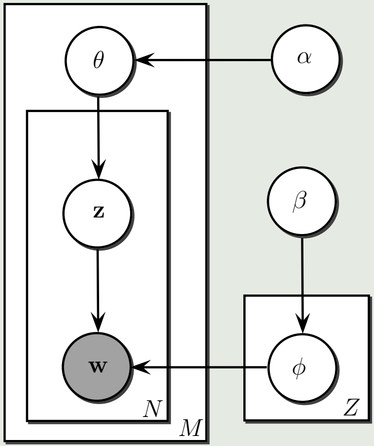


Latent Dirichlet Allocation Model (LDA) [?]

Generative Process

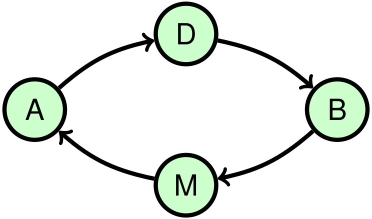
1. Draw $\theta^{(d)}$ from **Dirichlet**(α), where each $\theta^{(d)}$ consists of topic distribution for document *d*
2. Draw ϕ from **Dirichlet**(β), where ϕ encompasses word distribution for topic
3. For every word in the document *d*
    1. Draw a topic $z_i^{(d)}$ from **Multinomial** ($\theta^{(d)}$)
    2. Draw a word $w_i^{(d)}$ from **Multinomial** ($\phi_{z_i^{(d)}}$)

Graphical Model



Topics Over Time Model (TOT)

1. The model assumes a continuous distribution over time associated with each topic.
2. Topics are responsible for generating both observed time-stamps and also words.
3. The model does not capture the sequence of state changes with a Markov assumption.



Relaxing the Bag-of-Words Assumption in a Topic Model

Can the bag-of-words assumption be relaxed in a topic model?
This makes more sense as this is how documents are written by humans and also read.

Word Order

Fermat's Last Theorem states that

$$x^n + y^n = z^n$$

has no non-zero integer solutions for x, y and z when $n > 2$.

Bag-of-Words

last, states, has, when, integer, non, zero, solutions, x, n, 2, z, fermat, that,=,+

Something "NOT" very useful!

Figure: Illustration of Word Order

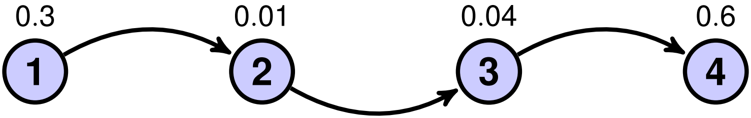


Figure: Illustration of n-gram generation using topic modeling approach

A sentence can be a segment.

Figure: Illustration of a segment

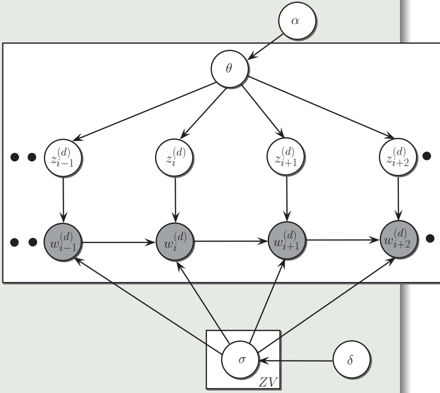
Relaxing the Bag-of-Words Assumption

Bigram Topic Model (BTM) [?]

Some Properties of the model

- Word is generated by both the topic and the previous word
- Inspired by the Hierarchical Dirichlet Language Model
- Better empirical results than the LDA model
- A limitation of the model
  - Always generates bigrams in a topic

Graphical Model of BTM



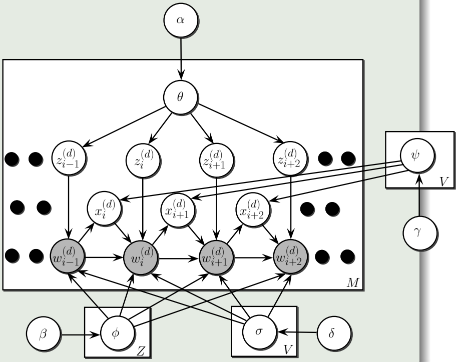
Relaxing the Bag-of-Words Assumption

LDA-Collocation Model (LDACOL) [?]

Some Properties of the model

- Word is generated by the topic, previous word and a binary bigram status variable
- Each word has a topic assignment and a collocation assignment
- Can generate both unigrams and bigrams
- A limitation of the model
  - Only the first word in a bigram has a topic assignment

Graphical Model of LDACOL



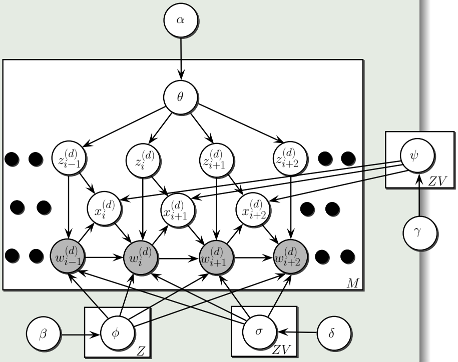
Relaxing the Bag-of-Words Assumption

Topical N-Gram Model (TNG) [?]

Some Properties of the model

- Extends LDACOL
- Each word has a topic assignment and a collocation assignment
- Can form longer order phrases
- Can generate both unigrams and bigrams
- A limitation of the model
  - Words in a bigram may have different topic assignments

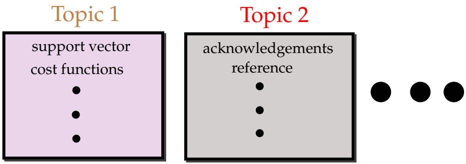
Graphical Model of TNG



What Topic N-gram models do - An Illustration

Abstract We give necessary and sufficient conditions for uniqueness of the support vector solution for the problems of pattern recognition and regression estimation, for a general class of cost functions. We sho that if the solution is not unique, all support vectors are necessarily at bound, and we give some simple examples of non-unique solutions. We note that uniqueness of the primal (dual) silution does not necessarily imply uniqueness of the dual (primal) solution. We show how to compute the threshold b when the solution is unique, but when all support vectors are bound, in which ... Para. 1

Acknowledgements C. Burges wishes to thank W. Keasler, V. Lawrence and C. Nohl of Lucent Technologies for their case the usual method for determining b does not work...
support. Reference [1] R. Fletcher, Practical Methods of Optimization. John Wiley and Sons, Inc., 2nd edition, 1987. Para. 2

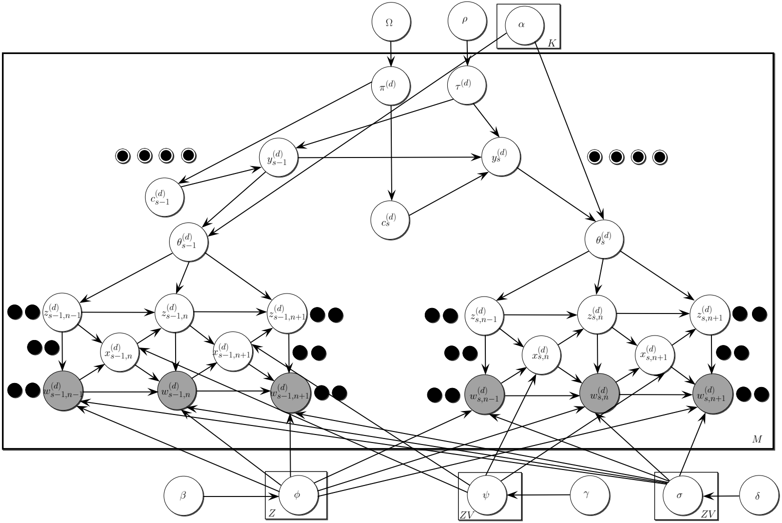


- Consider the document as a whole
- Find topical n-grams in the document

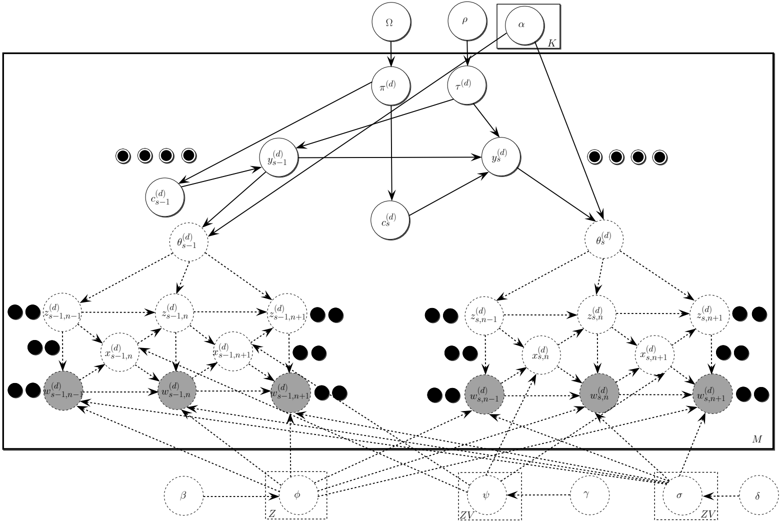
Main Contributions in this Thesis

- A model that maintains the document structure such as paragraphs and sentences (SIGIR-2013 [**?**]).
- Detection and coordination two topic granularity levels
  - Segment-Topics
  - Word-Topics
- Temporal dynamics in text data with n-grams (ECIR-2013 [**?**]).
- Proposed new models with word order to solve different tasks, for example, readability problem in IR (COLING-2012 [**?**], CIKM-2011 [**?**], WI-2012 [**?**], SKG-2012 [**?**], JCDL-2012 [**?**]), Bayesian nonparametrics (AIRS-2013 [**?**]).
- Derivation of the posterior inference schemes.

Our Proposed Model (NTSeg)



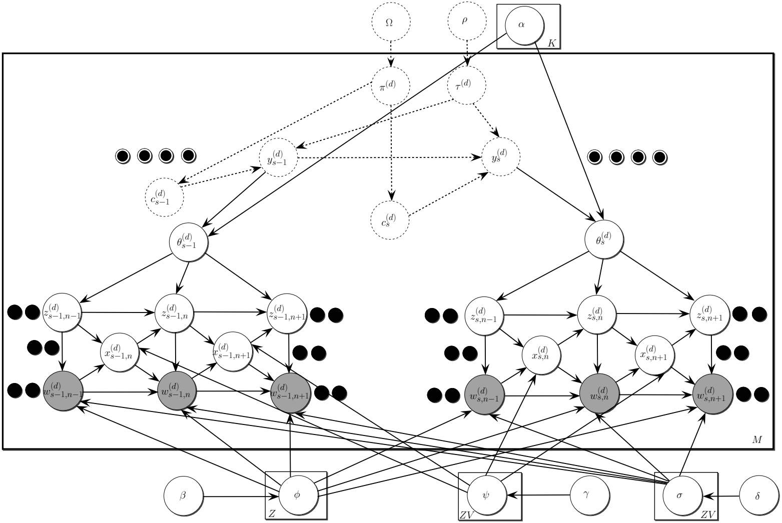
Our Proposed Model (NTSeg)



Few Properties of NTSeg

- Segments are assigned to the segment-topics.
- Assumes a Markov property on the segment-topics $y_s^{(d)}$.
- $c_s^{(d)}$ denotes the segment-topic change-points.
- Segments can be taken as a paragraphs or sentences.

Our Proposed Model (NTSeg)



Some properties of NTSeg

- Does not break the order of the words
- Can form unigrams, bigrams and higher order phrases (using **x**) variable
- *The phrases share the same topic*

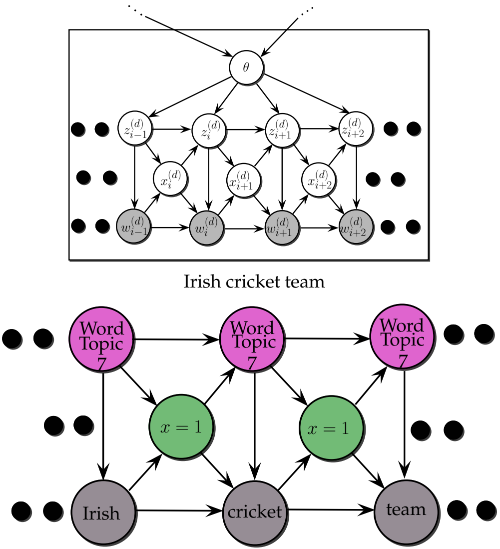


Figure: This is how we form longer phrases

Posterior Inference
Gibbs Sampling

Sampling word-topic assignments

$$P(z_{si}^{(d)}, x_{si}^{(d)} | \mathbf{w}, z_{\neg si}^{(d)}, x_{\neg si}^{(d)}, \mathbf{y}, \mathbf{c}, \alpha, \beta, \gamma, \delta, \rho, \Omega) \propto$$

$$\underbrace{\left(\alpha_{y_s^{(d)} z_{si}^{(d)}} + h_{s z_{si}^{(d)}}^{(d)} - 1 \right)}_{\text{Document topic proportions}} \times \underbrace{\left(\gamma_{x_{si}^{(d)}} + p_{z_{s,i-1}^{(d)} w_{s,i-1}^{(d)} x_{si}^{(d)}} - 1\right)}_{\text{Bigram status variable}}$$

$$\times \begin{cases} \underbrace{\dfrac{\beta_{w_{si}^{(d)}} + n_{z_{si}^{(d)} w_{si}^{(d)}} - 1}{\sum_{v=1}^{V} \left(\beta_v + n_{z_{si}^{(d)} v}\right) - 1}}_{\text{Prob. of a unigram in a topic}} & \text{if } x_{si}^{(d)} = 0 \\ \underbrace{\dfrac{\delta_{w_{si}^{(d)}} + m_{w_{si}^{(d)} w_{s,i-1}^{(d)} z_{si}^{(d)}} - 1}{\sum_{v=1}^{V} \left(\delta_v + m_{w_{s,i-1}^{(d)} v z_{si}^{(d)}}\right) - 1}}_{\text{Bigram probability}} & \text{if } x_{si}^{(d)} = 1 \ \& \ \underbrace{z_{si}^{(d)} = z_{s,i-1}^{(d)}}_{\text{Share same topic}} \end{cases}$$

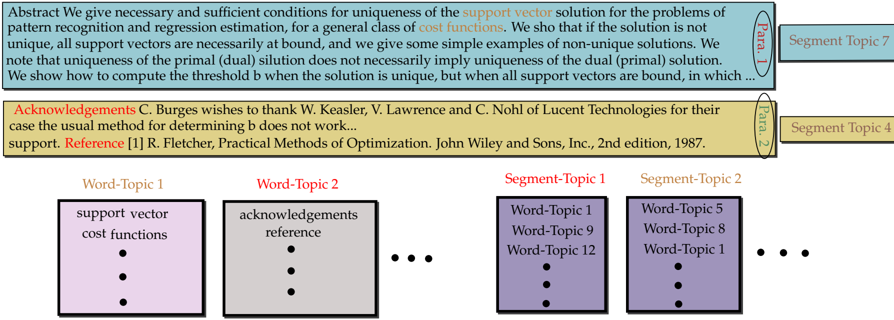
Posterior Inference
Gibbs Sampling

Sampling segment-topic assignments

$$P(y_s^{(d)}, c_s^{(d)} | \mathbf{z}, y_{\neg s}^{(d)}, c_{\neg s}^{(d)}, \mathbf{w}, \mathbf{x}, \alpha, \beta, \gamma, \delta, \rho, \Omega) \propto$$

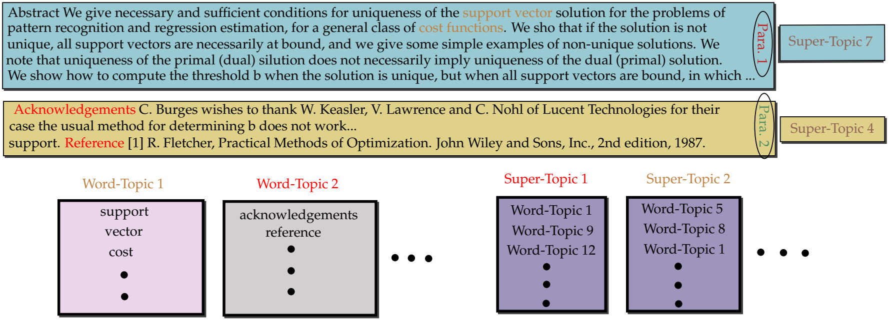
$$\begin{cases} (\underbrace{\rho_{y_s^{(d)}} + b_{y_s^{(d)}}^{(d)} - 1)}_{\text{Segment-topic mixtures}} \times (\underbrace{\alpha_{y_s^{(d)} z_{si}^{(d)}} + h_{s z_{si}^{(d)}}^{(d)} - 1}_{\text{Word and Segment topic mixtures}}) \times \\ \Bigg(\underbrace{\frac{\kappa_{c_s,0}^{(d)} + \Omega_0}{\sum_{x=0}^{1} \kappa_{c_s,x}^{(d)} + \Omega_0 + \Omega_1}}_{\text{Segment changepoints status update}} \Bigg) & \text{if } c_s^{(d)} = 0 \\ \left(\frac{\kappa_{c_s,1}^{(d)} + \Omega_1}{\sum_{x=0}^{1} \kappa_{c_s,x}^{(d)} + \Omega_0 + \Omega_1} \right) & \text{if } c_s^{(d)} = 1 \;\&\; s > 1 \;\&\; y_s^{(d)} = y_{(s-1)}^{(d)} \end{cases}$$

NTSeg Word-Topic and Segment-Topic Illustration



- Performs document segmentation based on topic
- N-gram words are assigned to the word-topics
- Segments are assigned to the segment-topics

Topic Segmentation Illustration - LDSEG

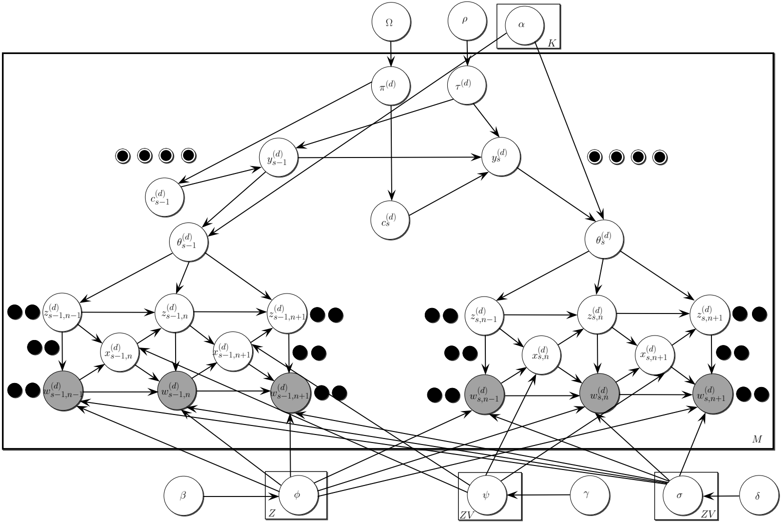


- Performs topic segmentation
- Unigram words are assigned to the word-topics
- Segments are assigned to the document-topics or super-topics

Word-topic and Segment-topic Correlation Graph

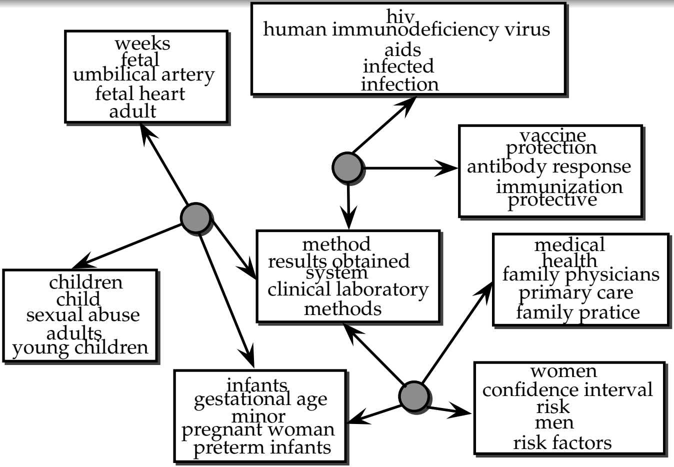
- Used a large dataset - OHSUMED
  - OHSUMED consists of 348,566 medical abstracts
- The idea is to show the discovery of n-gram words of topics via the correlation graph

Our Proposed Model (NTSeg)

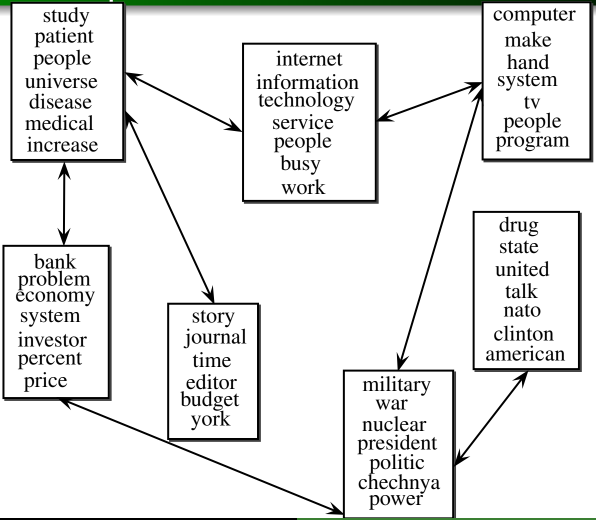


Word-Topic and Segment-Topic Correlation Graph

Result of NTSeg



Topic Correlation Graph
Correlation Graph from GD-LDA



Topic Segmentation Experiment

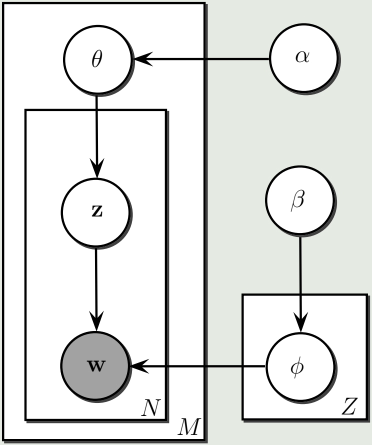
- Used two benchmark datasets - Books and Lectures
  - Books dataset - Medical text book, 140 sentences, 227 chapters
  - Lectures dataset - Undergraduate lecture recording of Physics and AI classes, 90 min lecture, 700 sentences, 8500 words
- Comparative method - **TopicTiling** Algorithm [**?**]
- Used two commonly used evaluation metrics
  - **Pk** - Probability that the two segments drawn randomly from a document are incorrectly identified as belonging to the same topic
  - **WinDiff** - Moves a sliding window across the text and counts the number of times the hypothesized and referenced segment boundaries are different from within the window
- These two evaluation metrics give an error estimate, so **the lower, the better**

Latent Dirichlet Allocation Model (LDA) [?]

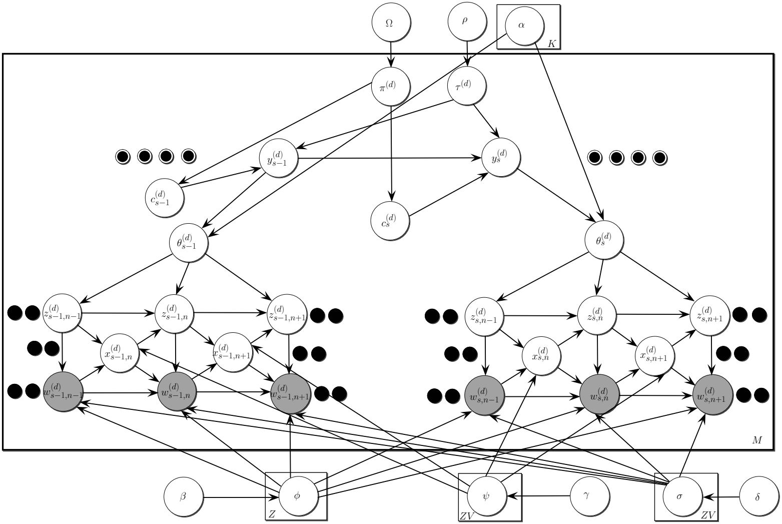
Generative Process

1. Draw $\theta^{(d)}$ from **Dirichlet**(α), where each $\theta^{(d)}$ consists of topic distribution for document *d*
2. Draw ϕ from **Dirichlet**(β), where ϕ encompasses word distribution for topic
3. For every word in the document *d*
   1. Draw a topic $z_i^{(d)}$ from **Multinomial** ($\theta^{(d)}$)
   2. Draw a word $w_i^{(d)}$ from **Multinomial** ($\phi_{z_i^{(d)}}$)

Graphical Model

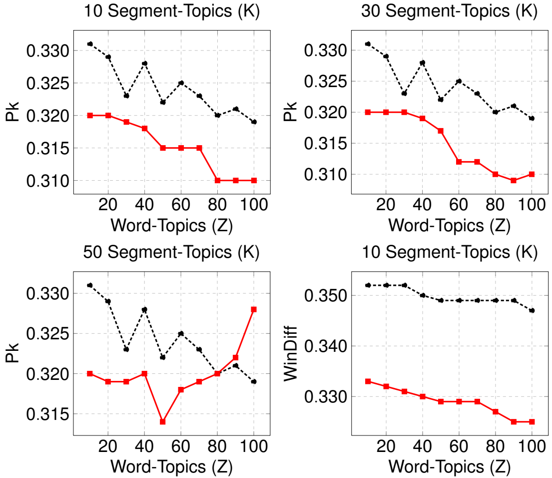


Our Proposed Model (NTSeg)



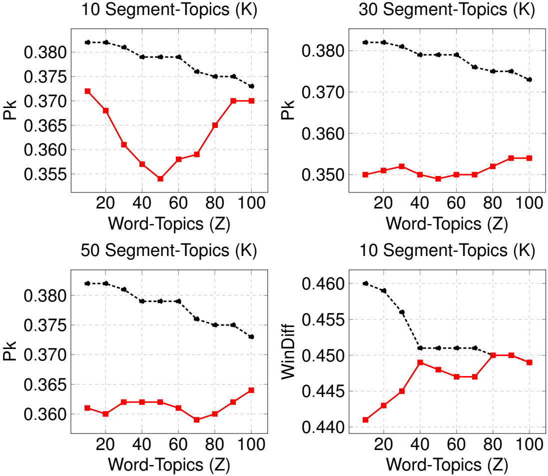
Topic Segmentation
Results

Books dataset



Topic Segmentation
Results

Lectures dataset



Document Classification Experiment
Dataset

- Generate four datasets from 20 Newsgroups data
- The datasets are:
  - Computer
  - Politics
  - Sports
  - Science
- Each dataset comprises of equal number of documents of several classes. For example, the Computer dataset consists of the following classes:
  - Graphics
  - Hardware
  - X Windows
  - Mac
  - Microsoft Windows

Document Classification Experiment
Experimental Setup

- Split each dataset into training and test set maintaining the class distribution
  - We used 75% training and 25% testing in our experiments
- For each class, we generate a topic model using the training set
- During classification, compute the likelihood of each document in the test set in **each** topic model
- The test document gets classified to that class where the likelihood is maximum
- Evaluation Metrics
  - Standard Precision, Recall and F-Measure for each class
  - Adopted Macro-Averaging scheme

Precision, Recall and F-Measure

In document classification:

Precision for a class:

The number of true positives divided by the total number of documents predicted to that class.

Recall is:

Recall is defined as the number of true positives divided by the total number of elements that actually belong to that class in the gold standard.

F-Measure is:

F-measure is the harmonic mean of precision and recall.

Document Classification Experiment
Comparative Methods

- Latent Dirichlet Segmentation Method (Word-Topics and Super-Topics) - **LDSEG** ([**?**])
- Pachinko Allocation Model (Super-Topics and Word-Topics) - **PAM** ([**?**])
- LDA Collocation Model (N-gram Topic Model) - **LDACOL** ([**?**])
- Topical N-gram Model (N-gram Topic Model) - **TNG** ([**?**])
- Phrase Discovery Topic Model based on Pitman-Yor Process - **PDLDA** ([**?**])

Document Classification Experiment
Results

	Precision	Recall	F-Measure	Precision	Recall	F-Measure
LDSEG	0.580	0.420	0.487	0.440	0.400	0.419
PAM	0.550	0.450	0.495	0.500	0.330	0.398
LDACOL	0.400	0.300	0.343	0.420	0.370	0.393
TNG	0.490	0.420	0.452	0.560	0.470	0.511
PDLDA	0.580	0.500	0.537	0.580	0.510	0.543
NTSeg	**0.640**	**0.520**	**0.574**	**0.620**	**0.560**	**0.588**

Computer dataset Science dataset

	Precision	Recall	F-Measure	Precision	Recall	F-Measure
LDSEG	0.390	0.320	0.352	0.330	0.320	0.325
PAM	0.540	0.490	0.514	0.368	0.360	0.363
LDACOL	0.550	0.410	0.470	0.200	0.180	0.189
TNG	0.550	0.450	0.495	0.340	0.290	0.313
PDLDA	0.590	0.410	0.484	0.380	0.210	0.271
NTSeg	**0.620**	**0.570**	**0.594**	**0.420**	**0.380**	**0.399**

Politics dataset Sports dataset

Document Modeling Experiment

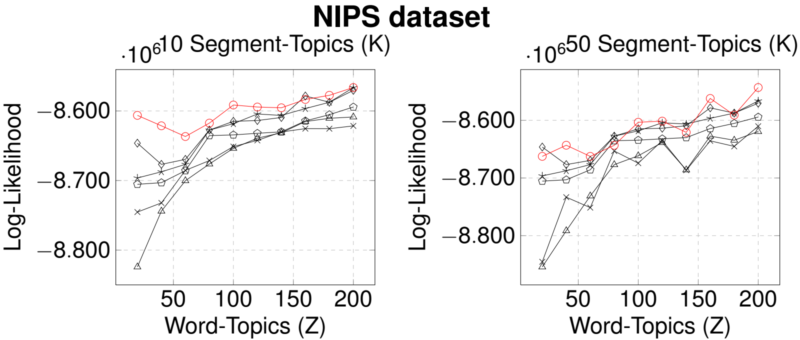


Figure: NTSeg (**??**) LDSEG (**??**), PAM (**??**), LDACOL (**??**), TNG (**??**), and PDLDA (**??**).

Document Modeling Experiment
Results

OHSUMED dataset (348,566 medical abstracts)

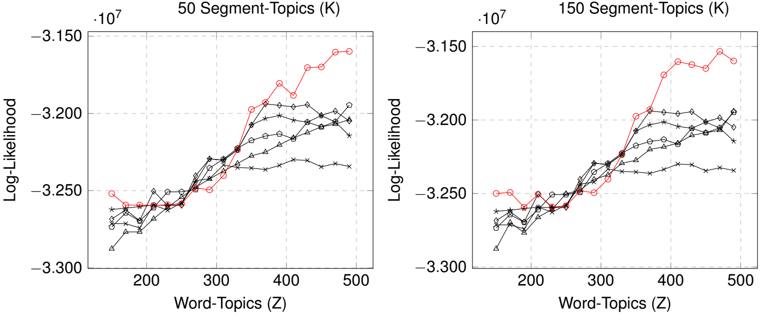


Figure: NTSeg (**??**) LDSEG (**??**), PAM (**??**), LDACOL (**??**), TNG (**??**), and PDLDA (**??**).

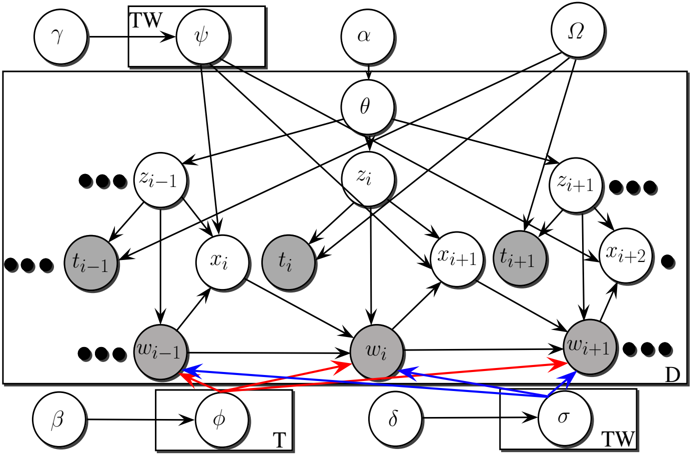
Our Model

N-gram Topics Over Time Model

1. The model assumes a continuous distribution over time associated with each topic.
2. Topics are responsible for generating both observed time-stamps and also words.
3. The model does not capture the sequence of state changes with a Markov assumption.
4. Maintains the order of words during topic generation process.
5. Generates words as unigrams, bigrams, etc. in topics.
6. Results in more interpretable topics.

Graphical Model

N-gram Topics Over Time Model



Generative Process

N-gram Topics Over Time Model

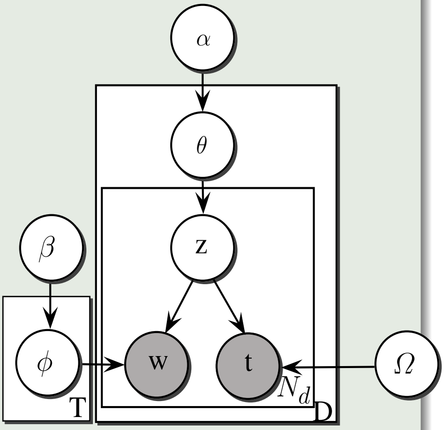
Draw **Discrete**(ϕ_z) from **Dirichlet**(β) for each topic z;
Draw **Bernoulli**(ψ_{zw}) from **Beta**(γ) for each topic z and each word w;
Draw **Discrete**(σ_{zw}) from **Dirichlet**(δ) for each topic z and each word w;
For every document d, draw **Discrete**($\theta^{(d)}$) from **Dirichlet**(α);
foreach *word* $w_i^{(d)}$ *in document* d **do**
  Draw $x_i^{(d)}$ from **Bernoulli**($\psi_{z_{i-1}^{(d)} w_{i-1}^{(d)}}$);
  Draw $z_i^{(d)}$ from **Discrete**($\theta^{(d)}$);
  Draw $w_i^{(d)}$ from **Discrete**($\sigma_{z_i^{(d)} w_{i-1}^{(d)}}$) if $x_i^{(d)} = 1$;
  Otherwise, Draw $w_i^{(d)}$ from **Discrete**($\phi_{z_i^{(d)}}$);
  Draw a time-stamp $t_i^{(d)}$ from **Beta**($\Omega_{z_i^{(d)}}$);
end

Topics Over Time (TOT) [?]

Generative Process

1. Draw T multinomials ϕ_z from a Dirichlet Prior β, one for each topic z
2. For each document d, draw a multinomial $\theta^{(d)}$ from a Dirichlet prior α; then for each word $w_i^{(d)}$ in the document d
   1. Draw a topic z_i^d from Multinomial $\theta^{(d)}$
   2. Draw a word $w_i^{(d)}$ from Multinomial $\phi_{z_i^{(d)}}$
   3. Draw a timestamp $t_i^{(d)}$ from Beta $\Omega_{z_i^{(d)}}$

Topics Over Time Model (TOT)



Posterior Inference
Collapsed Gibbs Sampling

$$P(z_i^{(d)}, x_i^{(d)} | \mathbf{w}, \mathbf{t}, \mathbf{x}_{\neg \mathbf{i}}^{(\mathbf{d})}, \mathbf{z}_{\neg \mathbf{i}}^{(\mathbf{d})}, \alpha, \beta, \gamma, \delta, \Omega) \propto \underbrace{\left(\gamma_{x_i^{(d)}} + p_{z_{i-1}^{(d)} w_{i-1}^{(d)} x_i} - 1\right)}_{\text{Bigram status update}} \times \underbrace{\left(\alpha_{z_i^{(d)}} + q_{d z_i^{(d)}} - 1 \right)}_{\text{Document topic prob. update}} \times$$

$$\underbrace{\frac{(1 - t_i^{(d)})^{\Omega_{z_i^{(d)}1} - 1} t_i^{(d)\Omega_{z_i^{(d)}2} - 1}}{B(\Omega_{z_i^{(d)}1}, \Omega_{z_i^{(d)}2})}}_{\text{Temporal information}} \times \begin{cases} \dfrac{\beta_{w_i^{(d)}} + n_{z_i^{(d)} w_i^{(d)}} - 1}{\sum_{v=1}^{W} (\beta_v + n_{z_i^{(d)} v}) - 1} & \text{if } x_i^{(d)} = 0 \\ \dfrac{\delta_{w_i^{(d)}} + m_{z_i^{(d)} w_{i-1}^{(d)} w_i^{(d)}} - 1}{\sum_{v=1}^{W} (\delta_v + m_{z_i^{(d)} w_{i-1}^{(d)} v}) - 1} & \text{if } x_i^{(d)} = 1 \end{cases} \quad (2)$$

Posterior Estimates

$$\hat{\theta}_z^{(d)} = \frac{\alpha_z + q_{dz}}{\sum_{t=1}^{T} (\alpha_t + q_{dt})} \quad (3) \qquad \hat{\phi}_{zw} = \frac{\beta_w + n_{zw}}{\sum_{v=1}^{W} (\beta_v + n_{zv})} \quad (4) \qquad \hat{\psi}_{zwk} = \frac{\gamma_k + p_{zwk}}{\sum_{k=0}^{1} (\gamma_k + p_{zwk})} \quad (5)$$

$$\hat{\sigma}_{zwv} = \frac{\delta_v + m_{zwv}}{\sum_{v=1}^{W} (\delta_v + m_{zwv})} \quad (6) \qquad \hat{\Omega}_{z1} = \bar{t_z}\left(\frac{\bar{t_z}(1 - \bar{t_z})}{s_z^2} - 1\right) \quad (7) \qquad \hat{\Omega}_{z2} = (1 - \bar{t_z})\left(\frac{\bar{t_z}(1 - \bar{t_z})}{s_z^2} - 1\right) \quad (8)$$

Empirical Evaluation
Data Sets

We have conducted experiments on two datasets

1. U.S. Presidential State-of-the-Union[1] speeches from 1790 to 2002.
2. NIPS conference papers - The original raw NIPS dataset[2] consists of 17 years of conference papers. But we supplemented this dataset by including some new raw NIPS documents[3] and it has 19 years of papers in total.

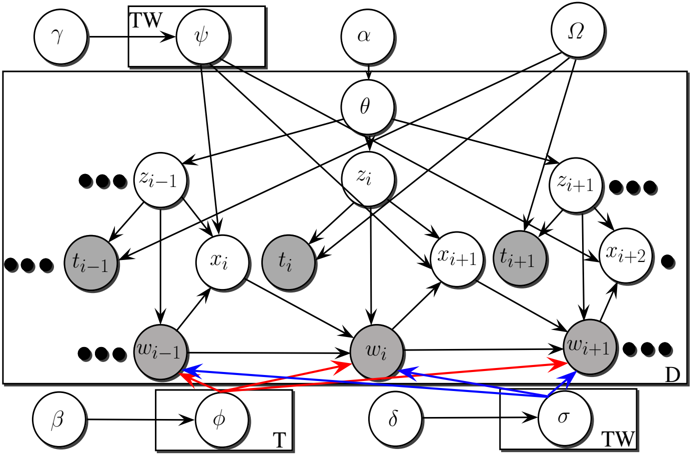
Preprocessing

1. Removed stopwords.
2. Did not perform word stemming.

[1] http://infomotions.com/etexts/gutenberg/dirs/etext04/suall11.txt
[2] http://www.cs.nyu.edu/~roweis/data.html
[3] http://ai.stanford.edu/~gal/Data/NIPS/

Graphical Model

N-gram Topics Over Time Model

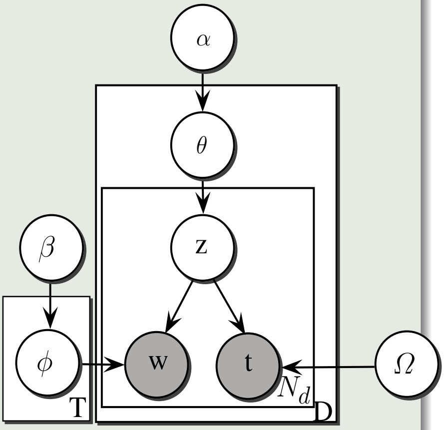


Topics Over Time (TOT) [?]

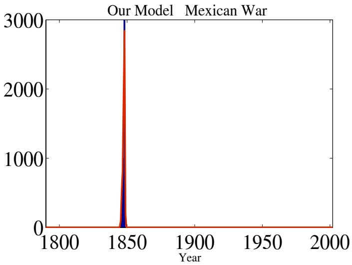
Generative Process

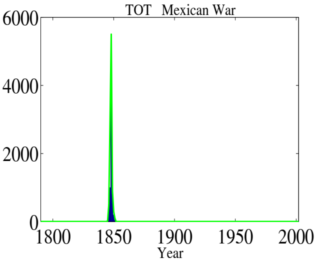
1. Draw T multinomials ϕ_z from a Dirichlet Prior β, one for each topic z
2. For each document d, draw a multinomial $\theta^{(d)}$ from a Dirichlet prior α; then for each word $w_i^{(d)}$ in the document d
    1. Draw a topic z_i^d from Multinomial $\theta^{(d)}$
    2. Draw a word $w_i^{(d)}$ from Multinomial $\phi_{z_i^{(d)}}$
    3. Draw a timestamp $t_i^{(d)}$ from Beta $\Omega_{z_i^{(d)}}$

Topics Over Time Model (TOT)



Qualitative Results



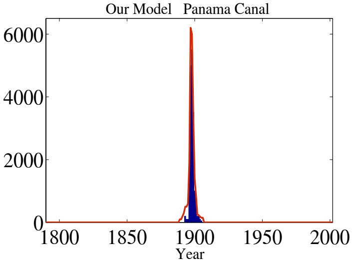


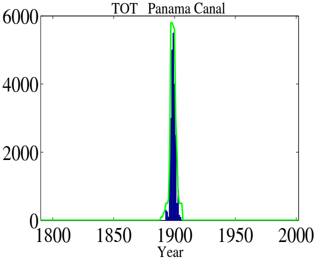
1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

Qualitative Results

Topics changes over time





1. panama canal	8. united states senate	1. government	8. spanish
2. isthmian canal	9. french canal company	2. cuba	9. island
3. isthmus panama	10. caribbean sea	3. islands	10. act
4. republic panama	11. panama canal bonds	4. international	11. commissio
5. united states government	12. panama	5. powers	12. officers
6. united states	13. american control	6. gold	13. spain
7. state panama	14. canal	7. action	14. rico

Qualitative Results
Topics changes over time - TOT

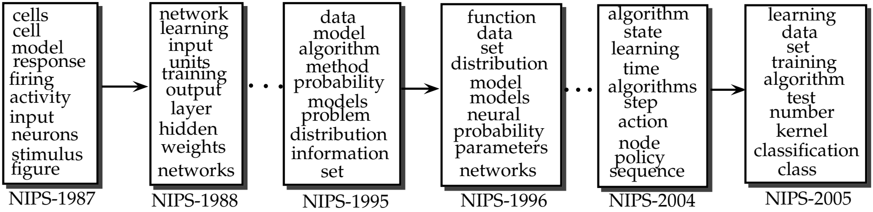


Figure: Top ten probable phrases from the posterior inference in NIPS year-wise.

Qualitative Results

Topics changes over time - Our Model

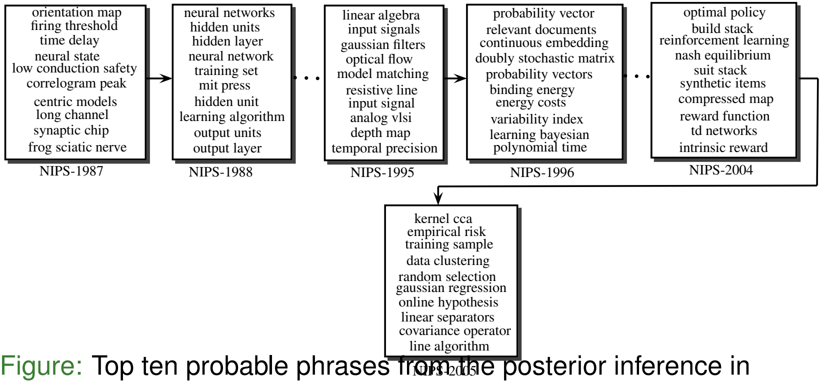
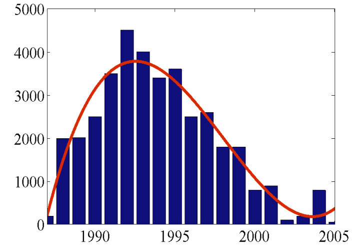
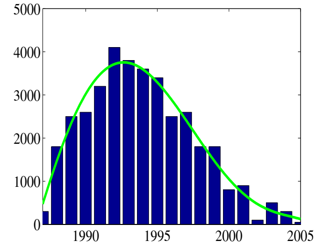


Figure: Top ten probable phrases from the posterior inference in NIPS year-wise.

Qualitative Results

Topics changes over time





1. hidden unit	6. learning algorithms
2. neural net	7. error signals
3. input layer	8. recurrent connections
4. recurrent network	9. training pattern
5. hidden layers	10. recurrent cascade

1. state	6. sequences
2. time	7. recurrent
3. sequence	8. models
4. states	9. markov
5. model	10. transition

Figure: A topic related to “recurrent NNs” comprising of *n*-gram words obtained from both the models.

Quantitative Results
Predicting decade on State-of-the-Union dataset

1. Computed the time-stamp prediction performance.
2. Learn a model on some subset of the data randomly sampled from the collection.
3. Given a new document, compute the likelihood of the decade prediction.

	L1 Error	E(L1)	Accuracy
Our Model	1.60	1.65	0.25
TOT	1.95	1.99	0.20

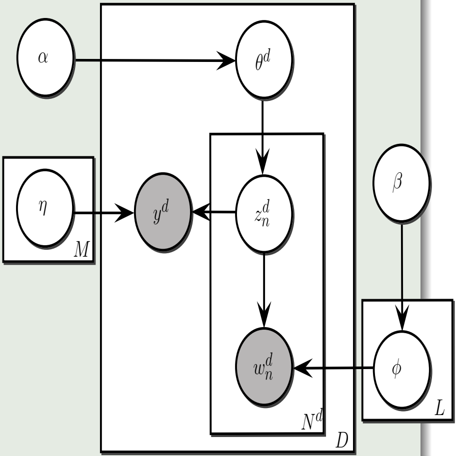
Table: Results of decade prediction in the State-of-the-Union speeches dataset.

MedLDA Topic Model [?]

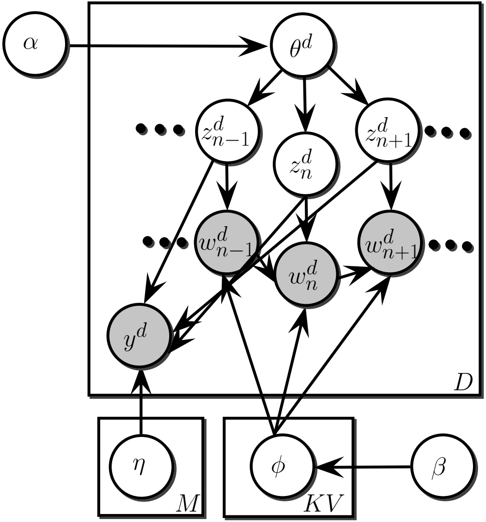
Properties

1. Considers side information during learning.
2. Side information, for example, class labels.
3. Side information can help generate more fine-grained topics.
4. Assumes a document as a bag-of-words.
5. Problem
    - Cannot capture the semantic storyline in the document.

Graphical Model



Our Supervised Topic Model



Results

Dataset	20 Newsgroups															
Models	`Our Model`	`gMedLDA`	`vMedLDA`	`sLDA`	`DiscLDA`	`LDA`	`LDA+SVM`	`BTM`	`BTM+SVM`	`LDACOL`	`LDACOL+SVM`	`TNG`	`TNG+SVM`	`NTSeg`	`NTSeg+SVM`	`SVM`
Topics	80	50	30	60	70	50	50	80	80	60	70	70	60	60	60	
Pre	0.945	0.869	0.865	0.805	0.756	0.859	0.835	0.877	0.835	0.843	0.845	0.845	0.832	0.766	0.869	0.825
Rec	0.916	0.869	0.865	0.812	0.780	0.859	0.920	0.848	0.920	0.914	0.932	0.932	0.866	0.905	0.845	0.910
F1	0.930	0.868	0.857	0.799	0.741	0.858	0.862	0.862	0.862	0.862	0.864	0.865	0.861	0.866	0.858	0.852
Dataset	OHSUMED-23															
Models	`Our Model`	`gMedLDA`	`vMedLDA`	`sLDA`	`DiscLDA`	`LDA`	`LDA+SVM`	`BTM`	`BTM+SVM`	`LDACOL`	`LDACOL+SVM`	`TNG`	`TNG+SVM`	`NTSeg`	`NTSeg+SVM`	`SVM`
Topics	70	40	60	60	70	40	40	60	40	50	50	60	60	40	40	
Pre	0.496	0.456	0.489	0.456	0.402	0.465	0.463	0.422	0.545	0.534	0.534	0.432	0.442	0.531	0.522	0.483
Rec	0.915	0.814	0.821	0.802	0.735	0.801	0.798	0.767	0.776	0.742	0.744	0.711	0.710	0.779	0.765	0.903
F1	0.643	0.633	0.629	0.620	0.587	0.626	0.631	0.610	0.622	0.630	0.625	0.623	0.620	0.634	0.623	0.630

Concluding Remarks

- We have presented a topic segmentation model that:
  - Maintains the document structure such as paragraphs and sentences
  - Keeps the order of the words intact
- We have applied our model in multitudes of text mining tasks
  - We have obtained good improvement over the state-of-the-art models

References I

[1] D. M. Blei.
Probabilistic topic models.
Commun. ACM, 55(4):77–84, Apr. 2012.

[2] M. Shafiei and E. Milios.
Latent Dirichlet co-clustering.
In *ICDM*, pages 542–551, Dec 2006.

[3] X. Wang and A. McCallum.
Topics over time: a non-Markov continuous-time model of topical trends.
In *SIGKDD*, pages 424–433, 2006.

[4] H. M. Wallach.
Topic modeling: beyond bag-of-words.
In *ICML*, pages 977–984, 2006.

[5] T. L. Griffiths, M. Steyvers, J. B. Tenenbaum, et al.
Topics in semantic representation.
Psychological Review, 114(2):211, 2007.

References II

[6] X. Wang, A. McCallum, and X. Wei.
Topical n-grams: Phrase and topic discovery, with an application to information retrieval.
In *ICDM*, pages 697–702. 2007.

[7] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei.
Hierarchical Dirichlet processes.
JASA, 101(476):1566–1581, 2006.

[8] R. V. Lindsey, W. P. Headden, III, and M. J. Stipicevic.
A phrase-discovering topic model using hierarchical Pitman-Yor processes.
In *EMNLP*,pages 214–222, 2012.

[9] W. Li and A. McCallum.
Pachinko Allocation: DAG-structured mixture models of topic correlations.
In *ICML*, pages 577–584, 2006.

[10] Shoaib Jameel and W. Lam.
An n-gram topic model for time-stamped documents.
In *ECIR*, pages 292–304, 2013.

References III

[11] Shoaib Jameel and W. Lam.
An unsupervised topic segmentation model incorporating word order.
In *SIGIR*, pages 472–479.

[12] Shoaib Jameel and Wai Lam.
A nonparametric n-gram topic model with interpretable latent topics.
In *Proceedings of the 9th Asia Information Retrieval Societies Conference*, pages 74–85. Springer, 2013.

[13] Shoaib Jameel, Wai Lam, Ching-man Au Yeung, and Sheaujiun Chyan.
An unsupervised ranking method based on a technical difficulty terrain.
In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1989–1992. ACM, 2011.

[14] Shoaib Jameel, Wai Lam, and Xiaojun Qian.
Ranking text documents based on conceptual difficulty using term embedding and sequential discourse cohesion.
In *Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 145–152. IEEE Computer Society, 2012.

References IV

[15] Shoaib Jameel, Wai Lam, Xiaojun Qian, and Ching-man Au Yeung.
An unsupervised technical difficulty ranking model based on conceptual terrain in the latent space.
In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 351–352. ACM, 2012.

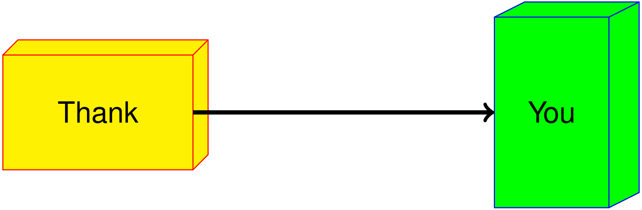
[16] Shoaib Jameel and Xiaojun Qian.
An unsupervised technical readability ranking model by building a conceptual terrain in LSI.
In *Proceedings of the 8th International Conference on Semantics, Knowledge and Grids*, pages 39–46. IEEE, 2012.

[17] Shoaib Jameel, Xiaojun Qian, and Wai Lam.
N-gram fragment sequence based unsupervised domain-specific document readability.
In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 1309–1326. ACL, 2012.

References V

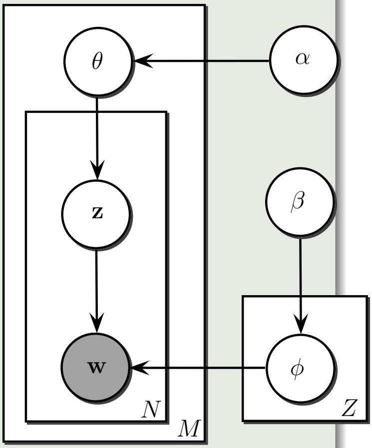
[18] Jun Zhu, Amr Ahmed, and Eric P Xing.
MedLDA: maximum margin supervised topic models for regression and classification.
In *ICML*, pages 1257–1264. ACM, 2009.

[19] Martin Riedl and Chris Biemann.
Topictiling: a text segmentation algorithm based on LDA.
In *ACL*, pages 37–42. Association for Computational Linguistics, 2012.


Thank
You

Bayesian Nonparametrics - Remember this?

Graphical Model - LDA



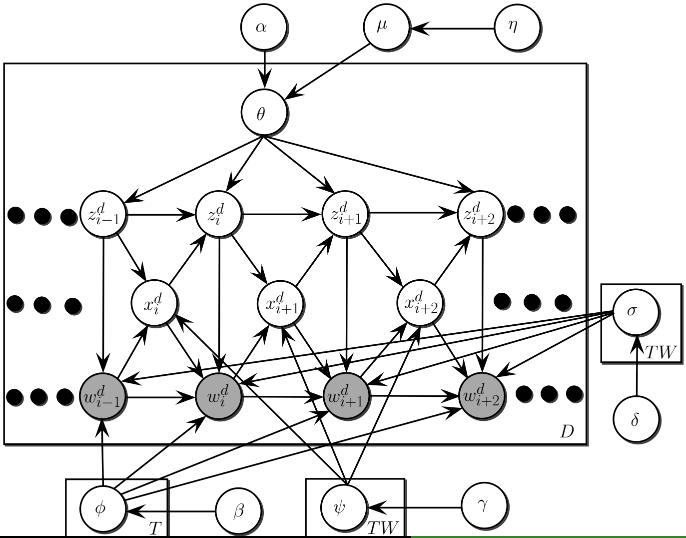
One Limitation

The variable Z has to be explicitly pre-defined.

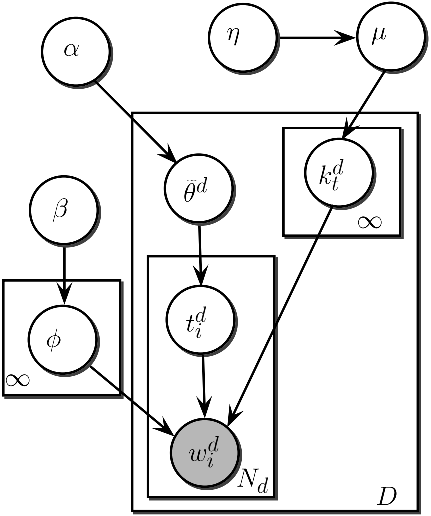
Number of latent topics

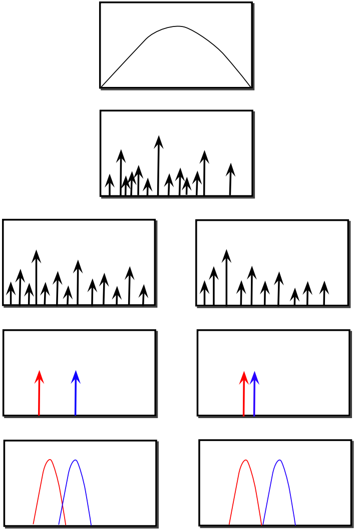
Variable Z

Nonparametric N-gram Model



Hierarchical Dirichlet Processes (HDP) [?]





Document Classification Results

Models	Precision	Recall	F-Measure
LDA	0.514	0.476	0.501
BTM	0.501	0.466	0.499
LDACOL	0.518	0.472	0.509
TNG	0.520	0.469	0.509
HDP	0.518	0.476	0.504
NHDP	0.496	0.491	0.483
NNTM−1	0.526	0.499	0.513
NNTM−2	0.501	0.438	0.509

Table: Computer Dataset

Models	Precision	Recall	F-Measure
LDA	0.416	0.392	0.392
BTM	0.401	0.376	0.376
LDACOL	0.405	0.322	0.394
TNG	0.411	0.339	0.399
HDP	0.416	0.401	0.405
NHDP	0.408	0.366	0.372
NNTM−1	0.415	0.405	0.405
NNTM−2	0.420	0.409	0.410

Table: Science Dataset

Document Classification Results

Models	Precision	Recall	F-Measure
LDA	0.412	0.401	0.376
BTM	0.415	0.401	0.398
LDACOL	0.416	0.402	0.389
TNG	0.411	0.399	0.399
HDP	0.418	0.401	0.405
NHDP	0.402	0.380	0.401
NNTM-1	0.416	0.401	0.402
NNTM-2	0.418	0.405	0.410

Table: Politics Dataset

Models	Precision	Recall	F-Measure
LDA	0.301	0.296	0.294
BTM	0.299	0.299	0.295
LDACOL	0.301	0.294	0.299
TNG	0.308	0.301	0.302
HDP	0.309	0.302	0.286
NHDP	0.302	0.296	0.292
NNTM-1	0.302	0.299	0.293
NNTM-2	0.303	0.301	0.303

Table: Sports Dataset

Readability - Some Traditional Readability Methods

The Flesch reading ease score is given by the following formula:

$$206.835 - 1.015 \times \underbrace{\frac{\text{Number of words}}{\text{Number of sentences}}}_{\text{Syntactic component}} - 84.6 \times \underbrace{\frac{\text{Number of syllables}}{\text{Number of words}}}_{\text{Semantic component}}$$

The Flesch-Kincaid reading ease formula is given by:

$$0.39 \times \underbrace{\frac{\text{Number of words}}{\text{Number of sentences}}}_{\text{Syntactic component}} + 11.8 \times \underbrace{\frac{\text{Number of syllables}}{\text{Number of words}}}_{\text{Semantic component}} - 15.59$$

Our Approach - Terrain Model

$$\mathbf{W} \approx \hat{\mathbf{W}} = \mathbf{USV}^T$$

Word difficulty scores:

$$\underset{[\gamma_n^x]}{\text{minimize}} \quad ||\vec{\hat{r}}_x - [\gamma_n^x]^T \boldsymbol{\Delta}_x||$$

$$\text{subject to} \quad \sum_{n=1}^{N^d} \gamma_n^x = 1, \gamma_n^x \geq 0$$

Cohesion:

$$\zeta_j = \frac{\sum_{s=1}^{S_j - 1} \nu(\vec{\Delta}_s, \vec{\Delta}_{s+1})}{S_j} \tau$$