Statistical Methods for Mining Big Text Data

ChengXiang Zhai

Department of Computer Science
Graduate School of Library & Information Science
Institute for Genomic Biology
Department of Statistics
University of Illinois, Urbana-Champaign

http://www.cs.illinois.edu/homes/czhai  czhai@illinois.edu
Rapid Growth of Text Information

WWW

Desktop

Email

Intranet

Blog/Tweets

Literature

How to help people manage and exploit all the information?
Text Information Systems Applications

How to connect users with the right information at the right time?

How to discover patterns in text and turn text data into actionable knowledge?

Focus of this tutorial
Goal of the Tutorial

- Brief introduction to the emerging area of applying statistical topic models to text mining (TM)

- Targeted audience:
  - Practitioners working on developing intelligent text information systems who are interested in learning about cutting-edge text mining techniques
  - Researchers who are looking for new research problems in text mining, information retrieval, and natural language processing

- Emphasis is on basic concepts, principles, and major application ideas

- Accessible to anyone with basic knowledge of probability and statistics

Check out David Blei’s tutorials on this topic for a more complete coverage of advanced topic models: http://www.cs.princeton.edu/~blei/topicmodeling.html
Outline

1. Background
   - Text Mining (TM)
   - Statistical Language Models

2. Basic Topic Models
   - Probabilistic Latent Semantic Analysis (PLSA)
   - Latent Dirichlet Allocation (LDA)
   - Applications of Basic Topic Models to Text Mining

3. Advanced Topic Models
   - Capturing Topic Structures
   - Contextualized Topic Models
   - Supervised Topic Models

4. Summary
What is Text Mining?

• Data Mining View: Explore patterns in textual data
  – Find latent topics
  – Find topical trends
  – Find outliers and other hidden patterns

• Natural Language Processing View: Make inferences based on partial understanding of natural language text
  – Information extraction
  – Question answering
Applications of Text Mining

• Direct applications
  – Discovery-driven (Bioinformatics, Business Intelligence, etc): We have specific questions; how can we exploit data mining to answer the questions?
  – Data-driven (WWW, literature, email, customer reviews, etc): We have a lot of data; what can we do with it?

• Indirect applications
  – Assist information access (e.g., discover major latent topics to better summarize search results)
  – Assist information organization (e.g., discover hidden structures to link scattered information)
Text Mining Methods

- **Data Mining Style**: View text as high dimensional data
  - Frequent pattern finding
  - Association analysis
  - Outlier detection

- **Information Retrieval Style**: Fine granularity topical analysis
  - Topic extraction
  - Exploit term weighting and text similarity measures

- **Natural Language Processing Style**: Information Extraction
  - Entity extraction
  - Relation extraction
  - Sentiment analysis

- **Machine Learning Style**: Unsupervised or semi-supervised learning
  - Mixture models
  - Dimension reduction

This tutorial
Outline

1. Background
   - Text Mining (TM)
   - Statistical Language Models

2. Basic Topic Models
   - Probabilistic Latent Semantic Analysis (PLSA)
   - Latent Dirichlet Allocation (LDA)
   - Applications of Basic Topic Models to Text Mining

3. Advanced Topic Models
   - Capturing Topic Structures
   - Contextualized Topic Models
   - Supervised Topic Models

4. Summary
What is a Statistical Language Model?

• A probability distribution over word sequences
  – \( p(\text{“Today is Wednesday”}) \approx 0.001 \)
  – \( p(\text{“Today Wednesday is”}) \approx 0.0000000000001 \)
  – \( p(\text{“The eigenvalue is positive”}) \approx 0.00001 \)

• Context-dependent!

• Can also be regarded as a probabilistic mechanism for “generating” text, thus also called a “generative” model
Why is a LM Useful?

• Provides a principled way to quantify the uncertainties associated with natural language

• Allows us to answer questions like:
  
  – Given that we see “John” and “feels”, how likely will we see “happy” as opposed to “habit” as the next word?
    (speech recognition)
  
  – Given that we observe “baseball” three times and “game” once in a news article, how likely is it about “sports”?
    (text categorization, information retrieval)
  
  – Given that a user is interested in sports news, how likely would the user use “baseball” in a query?
    (information retrieval)
Source-Channel Framework for “Traditional” Applications of SLMs

When $X$ is text, $p(X)$ is a language model

Many Examples:
- Speech recognition: $X=$Word sequence, $Y=$Speech signal
- Machine translation: $X=$English sentence, $Y=$Chinese sentence
- OCR Error Correction: $X=$Correct word, $Y=$Erroneous word
- Information Retrieval: $X=$Document, $Y=$Query
- Summarization: $X=$Summary, $Y=$Document

This tutorial is about another type of applications of SLMs (i.e., topic mining)
The Simplest Language Model
(Unigram Model)

• Generate a piece of text by generating each word INDEPENDENTLY

• Thus, \( p(w_1 \ w_2 \ ... \ w_n) = p(w_1)p(w_2)\ldots p(w_n) \)

• Parameters: \( \{p(w_i)\} \) \( p(w_1)+\ldots+p(w_N)=1 \) (N is voc. size)

• A piece of text can be regarded as a sample drawn according to this word distribution

\[
P(\text{"today is Wed"}) = P(\text{"today"})p(\text{"is"})p(\text{"Wed"})
\]
\[
= 0.0002 \times 0.001 \times 0.000015
\]
Text Generation with Unigram LM

(Unigram) Language Model $\theta$

$$p(w|\theta)$$

Topic 1: Text mining

- text 0.2
- mining 0.1
- association 0.01
- clustering 0.02
- food 0.00001

... food 0.25
... nutrition 0.1
... healthy 0.05
... diet 0.02

Topic 2: Health

Sampling

Document $d$

Given $\theta$, $p(d|\theta)$ varies according to $d$
Given $d$, $p(d|\theta)$ varies according to $\theta$

Text mining paper

Food nutrition paper
Estimation of Unigram LM

(Unigram) Language Model $\theta$

$p(w|\theta) = ?$

Maximum Likelihood (ML) Estimator:
(maximizing the probability of observing document $D$)

$$p(w|\theta) = p(w|D) = \frac{c(w, D)}{\sum_{w'} c(w', D)}$$
Maximum Likelihood vs. Bayesian

• Maximum likelihood estimation
  – “Best” means “data likelihood reaches maximum”
  \[
  \hat{\theta} = \arg \max_{\theta} P(X \mid \theta)
  \]
  – Problem: small sample

• Bayesian estimation
  – “Best” means being consistent with our “prior” knowledge and explaining data well
  \[
  \hat{\theta} = \arg \max_{\theta} P(\theta \mid X) = \arg \max_{\theta} P(X \mid \theta)P(\theta)
  \]
  – Problem: how to define prior?

In general, we consider distribution of \( \theta \), so a point estimate can be obtained in potentially multiple ways (e.g. mean vs. mode)
Illustration of Bayesian Estimation

Prior: $p(\theta)$

Likelihood: $p(X|\theta)$
$X=(x_1,...,x_N)$

Posterior: $p(\theta|X) \propto p(X|\theta)p(\theta)$

$\theta_0$: prior mode
$\theta$: posterior mode
$\theta_{ml}$: ML estimate
Computation of Maximum Likelihood Estimate

\[ V = \{ w_1, w_2, \ldots, w_N \} \]

\[ \hat{\theta} = \text{argmax}_\theta p(d \mid \theta) \]

\[ = \text{argmax}_\theta p(w_1)^{c(w_1)} p(w_2)^{c(w_2)} \ldots p(w_N)^{c(w_N)} \]

\[ = \text{argmax}_\theta \prod_{i=1}^{N} p(w_i)^{c(w_i)} \]

\[ = \text{argmax}_\theta \prod_{i=1}^{N} \theta_i^{c(w_i)} \]

\[ = \text{argmax}_\theta \sum_{i=1}^{N} c(w_i) \log \theta_i \]

Maximize log-likelihood: \[ l(\theta \mid d) = \sum_{i=1}^{N} c(w_i) \log \theta_i \text{, subject to: } \sum_{i=1}^{N} \theta_i = 1 \]

Lagrange function: \[ l'(\theta \mid d) = \sum_{i=1}^{N} c(w_i) \log \theta_i + \lambda \left( \sum_{i=1}^{N} \theta_i - 1 \right) \]

\[ \frac{\partial l'}{\partial \theta_i} = \frac{c(w_i)}{\theta_i} + \lambda = 0 \quad \rightarrow \quad \theta_i = -\frac{c(w_i)}{\lambda} \quad \rightarrow \quad \lambda = -\sum_{i=1}^{N} c(w_i) \quad \rightarrow \quad \hat{\theta}_i = p(w_i \mid \hat{\theta}) = \frac{c(w_i)}{\sum_{i=1}^{N} c(w_i)} = \frac{c(w_i)}{|d|} \]

Set partial derivatives to zero

Use \[ \sum_{i=1}^{N} \theta_i = 1 \]
Computation of Bayesian Estimate

- ML estimator: \( \hat{\theta} = \arg\max_{\theta} p(d | \theta) \)
- Bayesian estimator:
  - First consider posterior: \( p(\theta | d) \propto p(d | \theta)p(\theta) \)
  - Then, consider the mean or mode of the posterior dist.
- \( p(d | \theta) \): Sampling distribution (of data)
- \( P(\theta)=p(\theta_1, \ldots, \theta_N) \): our prior on the model parameters
- conjugate = prior can be interpreted as “extra”/“pseudo” data
- Dirichlet distribution is a conjugate prior for multinomial sampling distribution

\[
\text{Dir}(\theta | \alpha_1, \ldots, \alpha_N) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_N)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_N)} \prod_{i=1}^{N} \theta_i^{\alpha_i-1}
\]

“extra”/“pseudo” word counts
Computation of Bayesian Estimate (cont.)

Posterior distribution of parameters:

\[ p(\theta \mid d) = \text{Dir}(\theta \mid c(w_1) + \alpha_1, \ldots, c(w_N) + \alpha_N) \]

Property: If \( \theta \sim \text{Dir}(\theta \mid \alpha) \), then \( \mathbb{E}(\theta) = \left\{ \frac{\alpha_i}{\sum \alpha_i} \right\} \)

Thus the posterior mean estimate is:

\[ p(w_i \mid \hat{\theta}) = \int p(w_i \mid \theta) \text{Dir}(\theta \mid \alpha) d\theta \]

\[ = \frac{c(w_i) + \alpha_i}{|d| + \sum_{i=1}^{N} \alpha_i} \]

Compare this with ML estimate:

\[ p(w_i \mid \hat{\theta}) = \frac{c(w_i)}{\sum_{i=1}^{N} c(w_i)} = \frac{c(w_i)}{|d|} \]

Each word gets unequal extra “pseudo counts” based on prior

Total “pseudo counts” for all words
Unigram LMs for Topic Analysis

Background LM: $p(w|\theta_B)$

Collection LM: $p(w|\theta_C)$

Document LM: $p(w|\theta_d)$

Text mining papers

Computer Science Papers

General Background English Text
Unigram LMs for Association Analysis

What words are semantically related to “computer”?

Topic LM: \( p(w|\text{“computer”}) \)  
Normalized Topic LM: \( \frac{p(w|\text{“computer”})}{p(w|\theta_B)} \)

- computer 400
- software 150
- program 104
- …
- text 3.0
- …
- the 1.1
- a 0.99
- is 0.9
- we 0.8

Background LM: \( p(w|\theta_B) \)

- the 0.03
- a 0.02
- is 0.015
- we 0.01
- …
- computer 0.00001
- …

all the documents containing word “computer”

General Background English Text
More Sophisticated LMs

• Mixture of unigram language models
  - Assume multiple unigram LMs are involved in generating text data
  - Estimation of multiple unigram LMs “discovers” (recovers) latent topics in text

• Other sophisticated LMs (see [Jelinek 98, Manning & Schutze 99, Rosenfeld 00])
  - N-gram language models: \( p(w_1 w_2 \ldots w_n) = p(w_1)p(w_2|w_1)\ldots p(w_n|w_1 \ldots w_{n-1}) \)
  - Remote-dependence language models (e.g., Maximum Entropy model)
  - Structured language models (e.g., probabilistic context-free grammar)
Evaluation of SLMs

• Direct evaluation criterion: How well does the model fit the data to be modeled?
  – Example measures: Data likelihood, perplexity, cross entropy, Kullback-Leibler divergence (mostly equivalent)

• Indirect evaluation criterion: Does the model help improve the performance of the task?
  – Specific measure is task dependent
  – For retrieval, we look at whether a model is effective for a text mining task
  – We hope an “improvement” of a LM would lead to better task performance
Outline

1. Background
   - Text Mining (TM)
   - Statistical Language Models

2. Basic Topic Models
   - **Probabilistic Latent Semantic Analysis (PLSA)**
   - Latent Dirichlet Allocation (LDA)
   - Applications of Basic Topic Models to Text Mining

3. Advanced Topic Models
   - Capturing Topic Structures
   - Contextualized Topic Models
   - Supervised Topic Models

4. Summary
Document as a Sample of Mixed Topics

How can we discover these topic word distributions?

Many applications would be enabled by discovering such topics

- Summarize themes/aspects
- Facilitate navigation/browsing
- Retrieve documents
- Segment documents
- Many other text mining tasks

[ Criticism of government response to the hurricane primarily consisted of criticism of its response to the approach of the storm and its aftermath, specifically in the delayed response] to the [flooding of New Orleans. ... 80% of the 1.3 million residents of the greater New Orleans metropolitan area evacuated] ...[Over seventy countries pledged monetary donations or other assistance]. ...
Simplest Case: 1 topic + 1 “background”

Assume words in d are from two distributions: 1 topic + 1 background (rather than just one)

How can we “get rid of” the common words from the topic to make it more discriminative?

Background LM: \( p(w|\theta_B) \)

Document LM: \( p(w|\theta_d) \)
The Simplest Case: One Topic + One Background Model

Assume $p(w|\theta_B)$ and $\lambda$ are known

$\lambda = \text{assumed percentage of background words in } d$

$P(Topic)$  
\[ \lambda \]  
\[ \text{Background words} \]  
$P(w|\theta_B)$  
$\lambda$  
$P(Topic)$  
\[ 1-\lambda \]  
\[ \text{Topic words} \]  
$P(w|\theta)$  
$1-\lambda$  
$\sum_{w \in V} c(w,d) \log[\lambda p(w|\theta_B) + (1-\lambda) p(w|\theta)]$

Maximum Likelihood  
$\hat{\theta} = \arg \max_{\theta} \log p(d | \theta)$
Understanding a Mixture Model

Suppose each model would be selected with equal probability \( \lambda = 0.5 \)

The probability of observing word “text”:
\[
\lambda p(\text{“text”}|\theta_B) + (1- \lambda)p(\text{“text”}|\theta) = 0.5 * 0.0001 + 0.5 * p(\text{“text”}|\theta)
\]

The probability of observing word “the”:
\[
\lambda p(\text{“the”}|\theta_B) + (1- \lambda)p(\text{“the”}|\theta) = 0.5 * 0.2 + 0.5 * p(\text{“the”}|\theta)
\]

The probability of observing “the” & “text” (likelihood)
\[
[0.5*0.0001 + 0.5* p(\text{“text”}|\theta)] 
\times [0.5*0.2 + 0.5* p(\text{“the”}|\theta)]
\]

How to set \( p(\text{“the”}|\theta) \) and \( p(\text{“text”}|\theta) \) so as to maximize this likelihood?

assume \( p(\text{“the”}|\theta) + p(\text{“text”}|\theta) = \text{constant} \)

\( \Rightarrow \) give \( p(\text{“text”}|\theta) \) a higher probability than \( p(\text{“the”}|\theta) \) (why?)

\( \theta_B \) and \( \theta \) are competing for explaining words in document \( d \)!
Simplest Case Continued: How to Estimate $\theta$?

**Known Background**

$p(w|\theta_B)$

- the 0.2
- a 0.1
- we 0.01
- to 0.02
- text 0.0001
- mining 0.00005
- ...

**Unknown query topic**

$p(w|\theta) = ?$

- text = ?
- mining = ?
- association = ?
- word = ?
- ...

**“Text mining”**

Suppose we know the identity/label of each word ...

**Observed words**

$\lambda = 0.7$

$\lambda = 0.3$

**ML Estimator**
Can We Guess the Identity?

Identity ("hidden") variable: $z_i \in \{1 \text{ (background)}, 0 \text{ (topic)}\}$

Suppose the parameters are all known, what’s a reasonable guess of $z_i$?
- depends on $\lambda$ (why?)
- depends on $p(w|\theta_B)$ and $p(w|\theta)$ (how?)

$$p(z_i = 1 | w_i) = \frac{p(z_i = 1)p(w | z_i = 1)}{p(z_i = 1)p(w | z_i = 1) + p(z_i = 0)p(w | z_i = 0)}$$

$$= \frac{\lambda p(w | \theta_B)}{\lambda p(w | \theta_B) + (1 - \lambda)p^{\text{current}}(w | \theta)}$$  \hspace{1cm} \text{E-step}$$

$$p^\text{new}(w_i | \theta) = \frac{c(w_i,d)(1 - p(z_i = 1 | w_i))}{\sum_{w' \in V} c(w',d)(1 - p(z_i = 1 | w'))}$$  \hspace{1cm} \text{M-step}$$

Initially, set $p(w|\theta)$ to some random values, then iterate ...
An Example of EM Computation

\[ p^{(n)}(z_i = 1 | w_i) = \frac{\lambda p(w_i | \theta_B)}{\lambda p(w_i | \theta_B) + (1 - \lambda) p^{(n)}(w_i | \theta)} \]

\[ p^{(n+1)}(w_i | \theta) = \frac{c(w_i, d)(1 - p^{(n)}(z_i = 1 | w_i))}{\sum_{w_j \in \text{vocabulary}} c(w_j, d)(1 - p^{(n)}(z_j = 1 | w_j))} \]

**Assume \( \lambda = 0.5 \)**

| Word  | #  | P(w||θ_B) | Iteration 1 | Iteration 2 | Iteration 3 |
|-------|----|-----------|-------------|-------------|-------------|
|       |    |           | P(w|θ)     | P(z=1)     | P(w|θ)      | P(z=1)      | P(w|θ)      | P(z=1)      |
| The   | 4  | 0.5       | 0.25       | 0.67       | 0.20       | 0.71       | 0.18       | 0.74       |
| Paper | 2  | 0.3       | 0.25       | 0.55       | 0.14       | 0.68       | 0.10       | 0.75       |
| Text  | 4  | 0.1       | 0.25       | 0.29       | 0.44       | 0.19       | 0.50       | 0.17       |
| Mining| 2  | 0.1       | 0.25       | 0.29       | 0.22       | 0.31       | 0.22       | 0.31       |
| Log-Likelihood | | | -16.96 | -16.13 | -16.02 |

**Expectation-Step:**
Augmenting data by guessing hidden variables

**Maximization-Step**
With the “augmented data”, estimate parameters using maximum likelihood
Discover Multiple Topics in a Collection

\[ p_d(w) = \lambda_B p(w \mid \theta_B) + (1 - \lambda_B) \sum_{j=1}^{k} \pi_{d,j} p(w \mid \theta_j) \]

\[ \log p(d) = \sum_{w \in V} c(w,d) \log[\lambda_B p(w \mid \theta_B) + (1 - \lambda_B) \sum_{j=1}^{k} \pi_{d,j} p(w \mid \theta_j)] \]

\[ \log p(C \mid \Lambda) = \sum_{d \in C} \sum_{w \in V} c(w,d) \log[\lambda_B p(w \mid \theta_B) + (1 - \lambda_B) \sum_{j=1}^{k} \pi_{d,j} p(w \mid \theta_j)] \]

Parameters: \( \Lambda=(\lambda_B, \{\pi_{d,j}\}, \{\theta_j\}) \)

Can be estimated using ML Estimator

"Generating" word \( w \) in doc \( d \) in the collection

### Topic coverage in document \( d \)

- **Topic \( \theta_1 \)**
  - warning
  - system

- **Topic \( \theta_2 \)**
  - aid
  - donation
  - support

- **Topic \( \theta_k \)**
  - statistics
  - loss
  - dead

- **Background \( \theta_B \)**
  - is
  - the
  - a

Percentage of background words
Coverage of topic \( \theta_j \) in doc \( d \)
Prob. of word \( w \) in topic \( \theta_j \)
Probabilistic Latent Semantic Analysis/Indexing (PLSA/PLSI) [Hofmann 99a, 99b]

• Mix k multinomial distributions to generate a document

• Each document has a potentially different set of mixing weights which captures the topic coverage

• When generating words in a document, each word may be generated using a DIFFERENT multinomial distribution (this is in contrast with the document clustering model where, once a multinomial distribution is chosen, all the words in a document would be generated using the same multinomial distribution)

• By fitting the model to text data, we can estimate (1) the topic coverage in each document, and (2) word distribution for each topic, thus achieving “topic mining”
How to Estimate Multiple Topics? (Expectation Maximization)

Known Background
\[ p(w \mid \theta_B) \]

Unknown topic model
\[ p(w \mid \theta_1) = ? \]
"Text mining"

Unknown topic model
\[ p(w \mid \theta_2) = ? \]
"information retrieval"

E-Step:
Predict topic labels using Bayes Rule

M-Step:
Max. Likelihood Estimator based on "fractional counts"
Parameter Estimation

**E-Step:**

Word w in doc d is generated
- from cluster j
- from background

\[ p(z_{d,w} = j) = \frac{\pi_{d,j}^{(n)} p^{(n)}(w | \theta_j)}{\sum_{j'=1}^{k} \pi_{d,j'}^{(n)} p^{(n)}(w | \theta_{j'})} \]

\[ p(z_{d,w} = B) = \frac{\lambda_B p(w | \theta_B)}{\lambda_B p(w | \theta_B) + (1 - \lambda_B) \sum_{j=1}^{k} \pi_{d,j}^{(n)} p^{(n)}(w | \theta_j)} \]

**M-Step:**

Re-estimate
- mixing weights
- topic LM

\[ \pi_{d,j}^{(n+1)} = \frac{\sum_{w \in V} c(w, d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j)}{\sum_{j'} \sum_{w \in V} c(w, d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j')} \]

\[ p^{(n+1)}(w | \theta_j) = \frac{\sum_{d \in C} c(w, d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j)}{\sum_{w' \in V} \sum_{d \in C} c(w', d)(1 - p(z_{d,w'} = B)) p(z_{d,w'} = j)} \]

Sum over all docs in the collection

Fractional counts contributing to
- using cluster j in generating d
- generating w from cluster j
How the Algorithm Works

\[
\begin{align*}
\pi_{d1,1} (P(\theta_1|d_1)) & \\
\pi_{d1,2} (P(\theta_2|d_1)) & \\
\pi_{d2,1} (P(\theta_1|d_2)) & \\
\pi_{d2,2} (P(\theta_2|d_2)) & \\
\end{align*}
\]

Iteration 3, 4, 5, … Until converging

Topic coverage

1 - \lambda_B

\lambda_B

1 - \lambda_B

\lambda_B

P(w|\theta)

Topic 1

Topic 2

Initial value

\[ c(w,d)(1 - p(z_{d,w} = B))p(z_{d,w} = j) \]

\[ c(w,d)p(z_{d,w} = B) \]
PLSA with Prior Knowledge

• Users have some domain knowledge in mind, e.g.,
  – We expect to see “retrieval models” as a topic in IR literature
  – We want to see aspects such as “battery” and “memory” for opinions about a laptop
  – One topic should be fixed to model background words (infinitely strong prior!)

• We can easily incorporate such knowledge as priors of PLSA model
Adding Prior: Maximum a Posteriori (MAP) Estimation

$$\Lambda^* = \arg \max_{\Lambda} p(\Lambda)p(Data | \Lambda)$$

Most likely $\Lambda$

- Topic $\theta_1$
  - warning 0.3
  - system 0.2

- Topic $\theta_2$
  - aid 0.1
  - donation 0.05
  - support 0.02

- Topic $\theta_k$
  - statistics 0.2
  - loss 0.1
  - dead 0.05

- Background $\theta_B$
  - is 0.05
  - the 0.04
  - a 0.03

Parameters:
- $\lambda_B$ = noise-level (manually set)
- $\theta$’s and $\pi$’s are estimated with Maximum A Posteriori (MAP)

Prior can be placed on $\pi$ as well (more about this later)

"Generating" word $w$ in doc $d$ in the collection

$1 - \lambda_B$
Adding Prior as Pseudo Counts

Known Background
\[ p(w \mid B) \]

- the 0.2
- a 0.1
- we 0.01
- to 0.02

- text
- mining
- association
- word

Unknown topic model
\[ p(w \mid \theta_1) = \]

- “Text mining”

Unknown topic model
\[ p(w \mid \theta_2) = \]

- “Information retrieval”

Suppose, we know the identity of each word ...

Observed Doc(s)

MAP Estimator

Pseudo Doc

Size = \( \mu \)
Maximum A Posterior (MAP) Estimation

\[
p(z_{d,w} = j) = \frac{\pi_{d,j}^{(n)} p^{(n)}(w | \theta_j)}{\sum_{j'=1}^{k} \pi_{d,j'}^{(n)} p^{(n)}(w | \theta_{j'})}
\]

\[
p(z_{d,w} = B) = \frac{\lambda_B p(w | \theta_B)}{\lambda_B p(w | \theta_B) + (1 - \lambda_B) \sum_{j=1}^{k} \pi_{d,j}^{(n)} p^{(n)}(w | \theta_j)}
\]

\[
\pi_{d,j}^{(n+1)} = \frac{\sum_{w \in V} c(w, d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j)}{\sum_{j'} \sum_{w \in V} c(w, d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j')}
\]

\[
p^{(n+1)}(w | \theta_j) = \frac{\sum_{d \in C} c(w, d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j)}{\sum_{w \in V} \sum_{d \in C} c(w', d)(1 - p(z_{d,w'} = B)) p(z_{d,w'} = j)} + \mu p(w | \theta_j) + \mu
\]

\[
\text{Sum of all pseudo counts}
\]

What if \( \mu = 0? \) What if \( \mu = +\infty? \)

A consequence of using conjugate prior is that the prior can be converted into “pseudo data” which can then be “merged” with the actual data for parameter estimation.
A General Introduction to EM

Data: \( X \) (observed) + \( H \) (hidden)  Parameter: \( \theta \)

“Incomplete” likelihood: \( L(\theta) = \log p(X \mid \theta) \)

“Complete” likelihood: \( L_c(\theta) = \log p(X,H \mid \theta) \)

EM tries to iteratively maximize the incomplete likelihood:

Starting with an initial guess \( \theta^{(0)} \),

1. E-step: compute the **expectation** of the complete likelihood

\[
Q(\theta; \theta^{(n-1)}) = E_{\theta^{(n-1)}} [L_c(\theta) \mid X] = \sum_{h_i} p(H = h_i \mid X, \theta^{(n-1)}) \log P(X, h_i)
\]

2. M-step: compute \( \theta^{(n)} \) by **maximizing** the Q-function

\[
\theta^{(n)} = \arg \max_{\theta} Q(\theta; \theta^{(n-1)}) = \arg \max_{\theta} \sum_{h_i} p(H = h_i \mid X, \theta^{(n-1)}) \log P(X, h_i)
\]
Convergence Guarantee

Goal: maximizing “Incomplete” likelihood: \( L(\theta) = \log p(X|\theta) \)
I.e., choosing \( \theta^{(n)} \), so that \( L(\theta^{(n)}) - L(\theta^{(n-1)}) \geq 0 \)

Note that, since \( p(X,H|\theta) = p(H|X,\theta) p(X|\theta) \), \( L(\theta) = Lc(\theta) - \log p(H|X,\theta) \)
\( L(\theta^{(n)}) - L(\theta^{(n-1)}) = Lc(\theta^{(n)}) - Lc(\theta^{(n-1)}) + \log \frac{p(H|X,\theta^{(n-1)})}{p(H|X,\theta^{(n)})} \)

Taking expectation w.r.t. \( p(H|X,\theta^{(n-1)}) \),
\( L(\theta^{(n)}) - L(\theta^{(n-1)}) = Q(\theta^{(n)};\theta^{(n-1)}) - Q(\theta^{(n-1)};\theta^{(n-1)}) + D(p(H|X,\theta^{(n-1)})||p(H|X,\theta^{(n)})) \)

Doesn’t contain \( H \)

**EM chooses \( \theta^{(n)} \) to maximize \( Q \)**

**KL-divergence, always non-negative**

Therefore, \( L(\theta^{(n)}) \geq L(\theta^{(n-1)}) \)!
EM as Hill-Climbing: converging to a local maximum

Likelihood $p(X|\theta)$

$E$-step = computing the lower bound

$M$-step = maximizing the lower bound
Outline

1. Background
   - Text Mining (TM)
   - Statistical Language Models

2. Basic Topic Models
   - Probabilistic Latent Semantic Analysis (PLSA)
   - Latent Dirichlet Allocation (LDA)
   - Applications of Basic Topic Models to Text Mining

3. Advanced Topic Models
   - Capturing Topic Structures
   - Contextualized Topic Models
   - Supervised Topic Models

4. Summary
Deficiency of PLSA

• Not a generative model
  – Can’t compute probability of a new document
  – Heuristic workaround is possible, though

• Many parameters $\Rightarrow$ high complexity of models
  – Many local maxima
  – Prone to overfitting

• Not necessary a problem for text mining (only interested in fitting the “training” documents)
Latent Dirichlet Allocation (LDA) [Blei et al. 02]

• Make PLSA a generative model by imposing a Dirichlet prior on the model parameters ➔
  – LDA = Bayesian version of PLSA
  – Parameters are regularized

• Can achieve the same goal as PLSA for text mining purposes
  – Topic coverage and topic word distributions can be inferred using Bayesian inference
LDA = Imposing Prior on PLSA

PLSA:
Topic coverage \( \pi_{d,j} \) is specific to the “training documents”, thus can’t be used to generate a new document.

LDA:
Topic coverage distribution \( \{\pi_{d,j}\} \) for any document is sampled from a Dirichlet distribution, allowing for generating a new doc.

\[
p(\tilde{\pi}_d) = \text{Dirichlet}(\tilde{\alpha})
\]

In addition, the topic word distributions \( \{\theta_j\} \) are also drawn from another Dirichlet prior.

\[
p(\tilde{\theta}_i) = \text{Dirichlet}(\tilde{\beta})
\]

\( \pi_{d,1} \) \( \pi_{d,2} \) \( \pi_{d,k} \) \( \theta_1 \) \( \theta_2 \) \( \theta_k \)

\{\pi_{d,j}\} \) are free for tuning

“Generating” word \( w \) in doc \( d \) in the collection

\{\pi_{d,j}\} \) are regularized

Magnitudes of \( \alpha \) and \( \beta \) determine the variances of the prior, thus also the strength of prior (larger \( \alpha \) and \( \beta \) \( \Rightarrow \) stronger prior)
Equations for PLSA vs. LDA

PLSA

\[ p_d(w \mid \{\theta_j\}, \{\pi_{d,j}\}) = \sum_{j=1}^{k} \pi_{d,j} p(w \mid \theta_j) \]

\[ \log p(d \mid \{\theta_j\}, \{\pi_{d,j}\}) = \sum_{w \in V} c(w, d) \log \left( \sum_{j=1}^{k} \pi_{d,j} p(w \mid \theta_j) \right) \]

\[ \log p(C \mid \{\theta_j\}, \{\pi_{d,j}\}) = \sum_{d \in C} \log p(d \mid \{\theta_j\}, \{\pi_{d,j}\}) \]

LDA

\[ p_d(w \mid \{\theta_j\}, \{\pi_{d,j}\}) = \sum_{j=1}^{k} \pi_{d,j} p(w \mid \theta_j) \]

\[ \log p(d \mid \tilde{\alpha}, \{\theta_j\}) = \int \left[ \sum_{w \in V} c(w, d) \log \left( \sum_{j=1}^{k} \pi_{d,j} p(w \mid \theta_j) \right) \right] p(\tilde{\pi}_d \mid \tilde{\alpha}) d\tilde{\pi}_d \]

\[ \log p(C \mid \tilde{\alpha}, \tilde{\beta}) = \int \sum_{d \in C} \log p(d \mid \tilde{\alpha}, \{\theta_j\}) \prod_{j=1}^{k} p(\theta_j \mid \tilde{\beta}) d\theta_1 \ldots d\theta_k \]

Core assumption in all topic models

PLSA component

Added by LDA
Parameter Estimation & Inferences in LDA

\[ p_d(w \mid \{\theta_j\}, \{\pi_{d,j}\}) = \sum_{j=1}^{k} \pi_{d,j} p(w \mid \theta_j) \]

\[ \log p(d \mid \tilde{\alpha}, \{\theta_j\}) = \int \sum_{w \in V} c(w, d) \log \left[ \sum_{j=1}^{k} \pi_{d,j} p(w \mid \theta_j) \right] p(\tilde{\pi}_d \mid \tilde{\alpha}) d \tilde{\pi}_d \]

\[ \log p(C \mid \tilde{\alpha}, \tilde{\beta}) = \int \sum_{d \in C} \log p(d \mid \tilde{\alpha}, \{\theta_j\}) \prod_{j=1}^{k} p(\theta_j \mid \tilde{\beta}) d \theta_1 \ldots d \theta_k \]

Parameter estimation can be done in the same way as in PLSA:
Maximum Likelihood Estimator:

\[ (\hat{\alpha}, \hat{\beta}) = \arg \max_{\alpha, \beta} \log p(C \mid \tilde{\alpha}, \tilde{\beta}) \]

However, \( \{\theta_j\}, \{\pi_{d,j}\} \) must now be computed using posterior inference:

\[ p(\{\theta_j\} \mid C, \tilde{\alpha}, \tilde{\beta}) = \frac{p(C \mid \{\theta_j\}, \tilde{\alpha}, \tilde{\beta}) p(\{\theta_j\} \mid \tilde{\alpha}, \tilde{\beta})}{p(C \mid \tilde{\alpha}, \tilde{\beta})} = \frac{p(C \mid \{\theta_j\}, \tilde{\alpha}) p(\{\theta_j\} \mid \tilde{\beta})}{p(C \mid \tilde{\alpha}, \tilde{\beta})} \]

\[ p(\pi_d \mid C, \tilde{\alpha}, \tilde{\beta}) = \frac{p(C \mid \pi_d, \tilde{\alpha}, \tilde{\beta}) p(\pi_d \mid \tilde{\alpha}, \tilde{\beta})}{p(C \mid \tilde{\alpha}, \tilde{\beta})} = \frac{p(C \mid \pi_d, \tilde{\alpha}, \tilde{\beta}) p(\pi_d \mid \tilde{\alpha})}{p(C \mid \tilde{\alpha}, \tilde{\beta})} \]

Computationally intractable, must resort to approximate inference!
LDA as a graph model [Blei et al. 03a]

Dirichlet priors

distribution over topics for each document
(same as $\pi_d$ on the previous slides)

$\theta^{(d)} \sim \text{Dirichlet}(\alpha)$

topic assignment for each word

$z_i \sim \text{Discrete}(\theta^{(d)})$

word generated from assigned topic

$w_i \sim \text{Discrete}(\phi^{(z_i)})$

Most approximate inference algorithms aim to infer $p(z_i|\tilde{w}, \tilde{\alpha}, \tilde{\beta})$
from which other interesting variables can be easily computed
Approximate Inferences for LDA

• Many different ways; each has its pros & cons

• Deterministic approximation
  – variational EM [Blei et al. 03a]
  – expectation propagation [Minka & Lafferty 02]

• Markov chain Monte Carlo
  – full Gibbs sampler [Pritchard et al. 00]
  – collapsed Gibbs sampler [Griffiths & Steyvers 04]

Most efficient, and quite popular, but can only work with conjugate prior
The collapsed Gibbs sampler

[Griffiths & Steyvers 04]

• Using conjugacy of Dirichlet and multinomial distributions, integrate out continuous parameters

\[
P(z) = \int_{\Delta_T^D} P(z \mid \Theta)p(\Theta)d\Theta = \prod_{d=1}^{D} \frac{\Gamma(n^{(d)}_j + \alpha)}{\Gamma(\alpha)^T} \frac{\Gamma(T\alpha)}{\Gamma(\sum_j n^{(d)}_j + \alpha)}
\]

\[
P(w \mid z) = \int_{\Delta_W^T} P(w \mid z, \Phi)p(\Phi)d\Phi = \prod_{j=1}^{T} \frac{\Gamma(n^{(j)}_w + \beta)}{\Gamma(\beta)^W} \frac{\Gamma(W\beta)}{\Gamma(\sum_w n^{(j)}_w + \beta)}
\]

• Defines a distribution on discrete ensembles \( z \)

\[
P(z \mid w) = \frac{P(w \mid z)P(z)}{\sum_z P(w \mid z)P(z)}
\]
The collapsed Gibbs sampler
[Griffiths & Steyvers 04]

• Sample each $z_i$ conditioned on $z_{-i}$

$$P(z_i \mid w, z_{-i}) \propto \frac{n_{w_i}^{(z_i)} + \beta}{n_{z_i}^{(z_i)} + W\beta} \frac{n_j^{(d_i)} + \alpha}{n_{d_i}^{(d_i)} + T\alpha}$$

• This is nicer than your average Gibbs sampler:
  – memory: counts can be cached in two sparse matrices
  – optimization: no special functions, simple arithmetic
  – the distributions on $\Phi$ and $\Theta$ are analytic given $z$ and $w$, and can later be found for each sample
### Gibbs sampling in LDA

#### Update 1

<table>
<thead>
<tr>
<th>( i )</th>
<th>( w_i )</th>
<th>( d_i )</th>
<th>( z_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>
Gibbs sampling in LDA

<table>
<thead>
<tr>
<th>$i$</th>
<th>$w_i$</th>
<th>$d_i$</th>
<th>$z_i$</th>
<th>$\tilde{z}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
# Gibbs sampling in LDA

## Iteration 1 vs. Iteration 2

<table>
<thead>
<tr>
<th></th>
<th>$w_i$</th>
<th>$d_i$</th>
<th>$z_i$</th>
<th>$z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

**Words in $d_i$ assigned with topic $j$**

**Count of instances where $w_i$ is assigned with topic $j$**

**Count of all words assigned with topic $j$**

**Words in $d_i$ assigned with any topic**

\[
P(z_i = j \mid z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(i)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,j}^{(d_i)} + T\alpha}
\]
## Gibbs sampling in LDA

### Iteration 1 and 2

<table>
<thead>
<tr>
<th>$i$</th>
<th>$w_i$</th>
<th>$d_i$</th>
<th>$z_i$</th>
<th>$z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

What’s the most likely topic for $w_i$ in $d_i$?

How likely would $d_i$ choose topic $j$?

How likely would topic $j$ generate word $w_i$?

$$P(z_i = j | z_{-i}, w) \propto \frac{n^{(w_i)} + \beta}{n^{(i)} + W \beta} \frac{n^{(d_i)} + \alpha}{n^{(d_i)} + T \alpha}$$
### Gibbs sampling in LDA

**Iteration**

<table>
<thead>
<tr>
<th>(i)</th>
<th>(w_i)</th>
<th>(d_i)</th>
<th>(z_i)</th>
<th>(z_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

\[
P(z_i = j | z_{-i}, w) \propto \frac{n^{(w_i)}_{-i,j} + \beta}{n^{(i)}_{-i,j} + W\beta} \frac{n^{(d_i)}_{-i,j} + \alpha}{n^{(d_i)}_{-i,j} + T\alpha}
\]
Gibbs sampling in LDA

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>(w_i)</td>
</tr>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
</tr>
</tbody>
</table>

\[
P(z_i = j | z_{-i}, w) \propto \frac{n^{(w_i)}_{-i,j} + \beta}{n^{(i)}_{-i,j} + W\beta} \frac{n^{(d_i)}_{-i,j} + \alpha}{n^{(d_i)}_{-i,j} + T\alpha}
\]
Gibbs sampling in LDA

<table>
<thead>
<tr>
<th>$i$</th>
<th>$w_i$</th>
<th>$d_i$</th>
<th>$z_i$</th>
<th>$z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

$P(z_i = j | z_{-i}, w) \propto \frac{n^{(w_i)}_{-i,j} + \beta}{n^{(i)}_{-i,j} + W \beta} \frac{n^{(d_i)}_{-i,j} + \alpha}{n^{(d_i)}_{-i,} + T \alpha}$
### Gibbs sampling in LDA

<table>
<thead>
<tr>
<th>$i$</th>
<th>$w_i$</th>
<th>$d_i$</th>
<th>$z_i$</th>
<th>$z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

\[
P(z_i = j | z_{-i}, w) \propto \frac{n(w_i)}{n_{-i,j}} + \beta \frac{n(d_i)}{n_{-i,j} + W \beta} \frac{n_{-i,j} + \alpha}{n_{-i,i} + \alpha} + T \alpha
\]
Gibbs sampling in LDA

<table>
<thead>
<tr>
<th>iteration</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>$w_i$</td>
<td>$d_i$</td>
<td>$z_i$</td>
<td>$z_i$</td>
</tr>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

$$P(z_i = j | z_{-i}, w) \propto \frac{n(w_i)}{n_{-i,j}} + \beta \frac{n(d_i)}{n_{-i,i} + \alpha}$$
Outline

1. Background
   - Text Mining (TM)
   - Statistical Language Models

2. Basic Topic Models
   - Probabilistic Latent Semantic Analysis (PLSA)
   - Latent Dirichlet Allocation (LDA)
   - Applications of Basic Topic Models to Text Mining

3. Advanced Topic Models
   - Capturing Topic Structures
   - Contextualized Topic Models
   - Supervised Topic Models

4. Summary
Applications of Topic Models for Text Mining: Illustration with 2 Topics

\[
p(d \mid \theta_1 \oplus \theta_2) = \prod_{w \in V} \left[ \lambda p(w \mid \theta_1) + (1 - \lambda) p(w \mid \theta_2) \right]^{c(w,d)}
\]

**Likelihood:**

\[
\log p(d \mid \theta_1 \oplus \theta_2) = \sum_{w \in V} c(w,d) \log[\lambda p(w \mid \theta_1) + (1 - \lambda) p(w \mid \theta_2)]
\]

Application Scenarios:

- \(p(w \mid \theta_1) \) & \(p(w \mid \theta_2)\) are known; estimate \(\lambda\) → The doc is about text mining and food nutrition, how much percent is about text mining?

- \(p(w \mid \theta_1)\) & \(\lambda\) are known; estimate \(p(w \mid \theta_2)\) → 30% of the doc is about text mining, what’s the rest about?

- \(p(w \mid \theta_1)\) is known; estimate \(\lambda\) & \(p(w \mid \theta_2)\) → The doc is about text mining, is it also about some other topic, and if so to what extent?

- \(\lambda\) is known; estimate \(p(w \mid \theta_1)\) & \(p(w \mid \theta_2)\) → 30% of the doc is about one topic and 70% is about another, what are these two topics?

- Estimate \(\lambda, p(w \mid \theta_1), p(w \mid \theta_2)\) → The doc is about two subtopics, find out what these two subtopics are and to what extent the doc covers each.
Use PLSA/LDA for Text Mining

• Both PLSA and LDA would be able to generate
  – Topic coverage in each document: \( p(\pi_d = j) \)
  – Word distribution for each topic: \( p(w|\theta_j) \)
  – Topic assignment at the word level for each document
  – The number of topics must be given in advance

• These probabilities can be used in many different ways
  – \( \theta_j \) naturally serves as a word cluster
  – \( \pi_{d,j} \) can be used for document clustering \( j^* = \arg \max \pi_{d,j} \)
  – Contextual text mining: Make these parameters conditioned on context, e.g.,
    • \( p(\theta_j \mid \text{time}) \), from which we can compute/plot \( p(\text{time} \mid \theta_j) \)
    • \( p(\theta_j \mid \text{location}) \), from which we can compute/plot \( p(\text{loc} \mid \theta_j) \)
Sample Topics from TDT Corpus [Hofmann 99b]

<table>
<thead>
<tr>
<th>“plane”</th>
<th>“space shuttle”</th>
<th>“family”</th>
<th>“Hollywood”</th>
</tr>
</thead>
<tbody>
<tr>
<td>plane</td>
<td>space</td>
<td>home</td>
<td>film</td>
</tr>
<tr>
<td>airport</td>
<td>shuttle</td>
<td>family</td>
<td>movie</td>
</tr>
<tr>
<td>crash</td>
<td>mission</td>
<td>like</td>
<td>music</td>
</tr>
<tr>
<td>flight</td>
<td>astronauts</td>
<td>love</td>
<td>new</td>
</tr>
<tr>
<td>safety</td>
<td>launch</td>
<td>kids</td>
<td>best</td>
</tr>
<tr>
<td>aircraft</td>
<td>station</td>
<td>mother</td>
<td>hollywood</td>
</tr>
<tr>
<td>air</td>
<td>crew</td>
<td>life</td>
<td>love</td>
</tr>
<tr>
<td>passenger</td>
<td>nasa</td>
<td>happy</td>
<td>actor</td>
</tr>
<tr>
<td>board</td>
<td>satellite</td>
<td>friends</td>
<td>entertainment</td>
</tr>
<tr>
<td>airline</td>
<td>earth</td>
<td>cnn</td>
<td>star</td>
</tr>
</tbody>
</table>

Table 1: Four factors from a 128 factor decomposition of the TDT-1 corpus. Factor are represented by their 10 most probable words, i.e., the words are ordered according to \( P(w|z) \).
How to Help Users Interpret a Topic Model?

[Mei et al. 07b]

- Use top words
  - automatic, but hard to make sense

  Term, relevance, weight, feedback

- Human generated labels
  - Make sense, but cannot scale up

Retrieval Models

<table>
<thead>
<tr>
<th>term</th>
<th>0.16</th>
</tr>
</thead>
<tbody>
<tr>
<td>relevance</td>
<td>0.08</td>
</tr>
<tr>
<td>weight</td>
<td>0.07</td>
</tr>
<tr>
<td>feedback</td>
<td>0.04</td>
</tr>
<tr>
<td>independence</td>
<td>0.03</td>
</tr>
<tr>
<td>model</td>
<td>0.02</td>
</tr>
<tr>
<td>frequent</td>
<td>0.02</td>
</tr>
<tr>
<td>probabilistic</td>
<td>0.02</td>
</tr>
<tr>
<td>document</td>
<td>0.02</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Question: Can we automatically generate understandable labels for topics?
What is a Good Label?

Retrieval models

<table>
<thead>
<tr>
<th>term</th>
<th>weight</th>
<th>relevance</th>
<th>feedback</th>
<th>independence</th>
<th>model</th>
<th>frequent</th>
<th>probabilistic</th>
<th>document</th>
</tr>
</thead>
<tbody>
<tr>
<td>term</td>
<td>0.1599</td>
<td>0.0752</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>relevance</td>
<td>0.0660</td>
<td>0.0372</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>weight</td>
<td>0.0311</td>
<td>0.0310</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>feedback</td>
<td>0.0233</td>
<td>0.0310</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>independence</td>
<td>0.0188</td>
<td>0.0173</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Semantically close (relevance)
- Understandable – phrases?
- High coverage inside topic
- Discriminative across topics
- iPod Nano
- じょうほうけんさく
- Information Retrieval

A topic from [Mei & Zhai 06b]
Automatic Labeling of Topics [Mei et al. 07b]

Statistical topic models

Collection (Context)

Relevance Score

database system, clustering algorithm, r tree, functional dependency, iceberg cube, concurrency control, index structure ...

Candidate label pool

Multinomial topic models

Coverage; Discrimination

clustering algorithm; distance measure; ...

Ranked List of Labels

NLP Chunker Ngram stat.
Relevance: the Zero-Order Score

• Intuition: prefer phrases well covering top words

\[ p("clustering" | \theta) = 0.4 \]
\[ p("dimensional" | \theta) = 0.3 \]
\[ p("shape" | \theta) = 0.01 \]
\[ p("body" | \theta) = 0.001 \]

Latent Topic \( \theta \)

Clustering

dimensional

algorithm

birch

shape

body

Good Label \( (l_1) \): "clustering algorithm"

\[ \frac{p(clustering + algorithm | \theta)}{p(clustering + algorithm)} > \frac{p(body + shape | \theta)}{p(body + shape)} \]

Bad Label \( (l_2) \): "body shape"
Relevance: the First-Order Score

- Intuition: prefer phrases with similar context (distribution)

\[ \sum_{w} p(w | \theta) \text{PMI}(w, l | C) \]

\[ \infty \sum_{w} p(w | \theta) \text{PMI}(w, l | C) \]

\[ D(\theta | \text{clustering algorithm}) < D(\theta | \text{hash join}) \]
Results: Sample Topic Labels

- sampling 0.06
- estimation 0.04
- approximate 0.04
- histograms 0.03
- selectivity 0.03
- histogram 0.02
- answers 0.02
- accurate 0.02

**selectivity estimation ...**

- the, of, a, and, to, data, > 0.02
- clustering 0.02
- time 0.01
- clusters 0.01
- databases 0.01
- large 0.01
- performance 0.01
- quality 0.005

**iran contra**

- north 0.02
- case 0.01
- trial 0.01
- iran 0.01
- documents 0.01
- walsh 0.009
- reagan 0.009
- charges 0.007

**r tree b tree ...**

- tree 0.09
- trees 0.08
- spatial 0.08
- b 0.05
- r 0.04
- disk 0.02
- array 0.01
- cache 0.01

**clustering algorithm clustering structure ...**

- large data, data quality, high data, data application, ...
Results: Contextual-Sensitive Labeling

Context: Database
(SIGMOD Proceedings)

Context: IR
(SIGIR Proceedings)

sampling
estimation
approximation
histogram
selectivity
histograms

dependencies
functional
cube
multivalued
iceberg
buc

selectivity estimation;
random sampling;
approximate answers;
multivalue dependency
functional dependency
Iceberg cube
distributed retrieval;
parameter estimation;
mixture models;
term dependency;
independence assumption;
Using PLSA to Discover Temporal Topic Trends [Mei & Zhai 05]

[Graph showing normalized strength of themes over time with labels for gene, expressions, probability, microarray, biology data, web information, time series, classification, association rule, clustering, business, marketing, customer, model, business rules, association, support, and rules.]
Construct Theme Evolution Graph [Mei & Zhai 05]

1999 2000 2001 2002 2003 2004 T

SVM 0.007
criteria 0.007
classification 0.006
linear 0.005
...

decision 0.006
tree 0.006
classifier 0.005
class 0.005
Bayes 0.005
...

web 0.009
classification 0.007
features 0.006
topic 0.005
...

mixture 0.005
random 0.006
cluster 0.006
clustering 0.005
variables 0.005
...

topic 0.010
mixture 0.008
LDA 0.006
semantic 0.005
...

Classifica
tion 0.015
text 0.013
unlabeled 0.012
document 0.008
labeled 0.008
learning 0.007
...

Information
ollection 0.012
web 0.010
social 0.008
retrieval 0.007
distance 0.005
networks 0.004
...

...
Use PLSA to Integrate Opinions [Lu & Zhai 08]

**Input**
- Topic: iPod
- Expert review with aspects
- Text collection of ordinary opinions, e.g., Weblogs

**Output**
- Integrated Summary
- Design
  - Battery
  - Price
- Similar opinions
  - cute... tiny...
  - last many hrs
  - could afford it
- Supplementary opinions
  - ..thicker..
  - die out soon
  - still expensive
- Extra Aspects Review Aspects
  - iTunes: ... easy to use...
  - Warranty: ...better to extend..
Methods

• Semi-Supervised Probabilistic Latent Semantic Analysis (PLSA)
  – The aspects extracted from expert reviews serve as clues to define a conjugate prior on topics
  – Maximum a Posteriori (MAP) estimation
  – Repeated applications of PLSA to integrate and align opinions in blog articles to expert review
Results: Product (iPhone)

- Opinion Integration with review aspects

<table>
<thead>
<tr>
<th>Review article</th>
<th>Similar opinions</th>
<th>Supplementary opinions</th>
</tr>
</thead>
<tbody>
<tr>
<td>You can make emergency calls, but you can't use any other functions...</td>
<td>N/A</td>
<td>... methods for unlocking the iPhone have emerged on the Internet in the past few weeks, although they involve tinkering with the iPhone hardware...</td>
</tr>
<tr>
<td>rated battery life, 24 hours talk time, 24 hours of music playback, 7 hours of video playback, and 6 hours on Internet use.</td>
<td>Up to 8 Hours of Talk Time, 6 Hours of Internet Use, 7 Hours of Video Playback or 24 Hours of Audio Playback</td>
<td>Playing relatively high bitrate VGA H.264 videos, our iPhone lasted almost exactly 9 freaking hours of continuous playback with cell and WiFi on (but Bluetooth off).</td>
</tr>
</tbody>
</table>

**Unlock/hack iPhone**

**Confirm the opinions from the review**

**Activation**

**Battery**

**Additional info under real usage**
Results: Product (iPhone)

• Opinions on extra aspects

<table>
<thead>
<tr>
<th>support</th>
<th>Supplementary opinions on extra aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>You may have heard of <strong>iASign</strong>, an iPhone Dev Utility that allows you to <strong>activate</strong> your phone without going through the iTunes rigamarole.</td>
</tr>
<tr>
<td>13</td>
<td><strong>Cisco</strong> has owned the <strong>trademark</strong> on the name &quot;<strong>iPhone</strong>&quot; since 2000, when it acquired <strong>InfoGear Technology Corp.</strong>, which originally registered the name.</td>
</tr>
<tr>
<td>13</td>
<td>With the imminent availability of the iPhone, a look at 10 things current smartphones like the <strong>Nokia N95</strong> have been able to do for a while and that the iPhone can't currently match...</td>
</tr>
</tbody>
</table>

*Another way to activate iPhone*

*iPhone trademark originally owned by Cisco*

*A better choice for smart phones?*
Results: Product (iPhone)

• Support statistics for review aspects

- People care about price
- Controversy: activation requires contract with AT&T
- People comment a lot about the unique wi-fi feature
Comparison of Task Performance of PLSA and LDA [Lu et al. 11]

- Three text mining tasks considered
  - Topic model for text clustering
  - Topic model for text categorization (topic model is used to obtain low-dimensional representation)
  - Topic model for smoothing language model for retrieval

- Conclusions
  - PLSA and LDA generally have similar task performance for clustering and retrieval
  - LDA works better than PLSA when used to generate low-dimensional representation (PLSA suffers from overfitting)
  - Task performance of LDA is very sensitive to setting of hyperparameters
  - Multiple local maxima problem of PLSA didn’t seem to affect task performance much
Outline

1. Background
   - Text Mining (TM)
   - Statistical Language Models

2. Basic Topic Models
   - Probabilistic Latent Semantic Analysis (PLSA)
   - Latent Dirichlet Allocation (LDA)
   - Applications of Basic Topic Models to Text Mining

3. Advanced Topic Models
   - Capturing Topic Structures
   - Contextualized Topic Models
   - Supervised Topic Models

4. Summary
Overview of Advanced Topic Models

• There are MANY variants and extensions of the basic PLSA/LDA topic models!

• Selected major lines to cover in this tutorial
  – Capturing Topic Structures
  – Contextualized Topic Models
  – Supervised Topic Models
Capturing Topic Structure: Learning topic hierarchies

- Fixed hierarchies: [Hofmann 99c]
- Learning hierarchies: [Blei et al 03b]
Learning topic hierarchies

The topics in each document form a path from root to leaf

- Fixed hierarchies: [Hofmann 99c]
- Learning hierarchies: [Blei et al. 03b]
Twelve Years of NIPS [Blei et al. 03b]
Capturing Topic Structures: Correlated Topic Model (CTM) [Blei & Lafferty 05]

- Draw topic proportions from a logistic normal, where topic occurrences can exhibit correlation.

- Use for:
  - Providing a “map” of topics and how they are related
  - Better prediction via correlated topics

No conjugate prior on topic proportions
Sample Result of CTM
1. Background
   - Text Mining (TM)
   - Statistical Language Models
2. Basic Topic Models
   - Probabilistic Latent Semantic Analysis (PLSA)
   - Latent Dirichlet Allocation (LDA)
   - Applications of Basic Topic Models to Text Mining
3. Advanced Topic Models
   - Capturing Topic Structures
   - Contextualized Topic Models
   - Supervised Topic Models
4. Summary

We are here
Contextual Topic Mining

• Documents are often associated with context (metadata)
  – Direct context: time, location, source, authors,…
  – Indirect context: events, policies, …

• Many applications require “contextual text analysis”:
  – Discovering topics from text in a context-sensitive way
  – Analyzing variations of topics over different contexts
  – Revealing interesting patterns (e.g., topic evolution, topic variations, topic communities)
Example: Comparing News Articles

Common Themes | “Vietnam” specific | “Afghan” specific | “Iraq” specific
--- | --- | --- | ---
United nations | … | … | …
Death of people | … | … | …
… | … | … | …

What’s in common? What’s unique?
More Contextual Analysis Questions

• What positive/negative aspects did people say about X (e.g., a person, an event)? Trends?
• How does an opinion/topic evolve over time?
• What are emerging research topics in computer science? What topics are fading away?
• How can we mine topics from literature to characterize the expertise of a researcher?
• How can we characterize the content exchanges on a social network?
• …
Contextual Probabilistic Latent Semantics Analysis [Mei & Zhai 06b]

**Themes**
- government
- donation
- New Orleans

**View1**
- July 2005

**View2**
- sociologist

**View3**
- government 0.3
  - response 0.2
- donate 0.1
- relief 0.05
- help 0.02
- city 0.2
- new 0.1
- orleans 0.05

Choose a view

Choose a theme

Criticism of government response to the hurricane primarily consisted of criticism of its response to ... The total shut-in oil production from the Gulf of Mexico ... approximately 24% of the annual production and the shut-in gas production ... Over seventy countries pledged monetary donations or other assistance. ...
Comparing News Articles [Zhai et al. 04]
Iraq War (30 articles) vs. Afghan War (26 articles)

The common theme indicates that “United Nations” is involved in both wars

<table>
<thead>
<tr>
<th>Common Theme</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>united nations</td>
<td>0.042</td>
<td>killed</td>
<td>0.035</td>
</tr>
<tr>
<td>month</td>
<td>0.032</td>
<td>month</td>
<td>0.032</td>
</tr>
<tr>
<td>deaths</td>
<td>0.023</td>
<td>deaths</td>
<td>0.023</td>
</tr>
</tbody>
</table>

Iraq Theme

<table>
<thead>
<tr>
<th></th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>0.03</td>
<td>troops</td>
</tr>
<tr>
<td>Weapons</td>
<td>0.024</td>
<td>hoon</td>
</tr>
<tr>
<td>Inspections</td>
<td>0.023</td>
<td>sanches</td>
</tr>
</tbody>
</table>

Afghan Theme

<table>
<thead>
<tr>
<th></th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern alliance</td>
<td>0.04</td>
<td>taleban</td>
</tr>
<tr>
<td>Kabul</td>
<td>0.03</td>
<td>rumsfeld</td>
</tr>
<tr>
<td>Taleban</td>
<td>0.025</td>
<td>hotel</td>
</tr>
<tr>
<td>Aid</td>
<td>0.02</td>
<td>front</td>
</tr>
</tbody>
</table>

Collection-specific themes indicate different roles of “United Nations” in the two wars
Spatiotemporal Patterns in Blog Articles

[Mei et al. 06a]

- Query = “Hurricane Katrina”

- Topics in the results:

<table>
<thead>
<tr>
<th>Government Response</th>
<th>New Orleans</th>
<th>Oil Price</th>
<th>Praying and Blessing</th>
<th>Aid and Donation</th>
<th>Personal</th>
</tr>
</thead>
<tbody>
<tr>
<td>bush 0.071</td>
<td>city 0.063</td>
<td>price 0.077</td>
<td>god 0.141</td>
<td>donate 0.120</td>
<td>i 0.405</td>
</tr>
<tr>
<td>president 0.061</td>
<td>orleans 0.054</td>
<td>oil 0.064</td>
<td>pray 0.047</td>
<td>relief 0.076</td>
<td>my 0.116</td>
</tr>
<tr>
<td>federal 0.051</td>
<td>new 0.034</td>
<td>gas 0.045</td>
<td>prayer 0.041</td>
<td>red 0.070</td>
<td>me 0.060</td>
</tr>
<tr>
<td>government 0.047</td>
<td>louisiana 0.023</td>
<td>increase 0.020</td>
<td>love 0.030</td>
<td>cross 0.065</td>
<td>am 0.029</td>
</tr>
<tr>
<td>fema 0.047</td>
<td>flood 0.022</td>
<td>product 0.020</td>
<td>life 0.025</td>
<td>help 0.050</td>
<td>think 0.015</td>
</tr>
<tr>
<td>administrate 0.023</td>
<td>evacuate 0.021</td>
<td>fuel 0.018</td>
<td>bless 0.025</td>
<td>victim 0.036</td>
<td>feel 0.012</td>
</tr>
<tr>
<td>response 0.020</td>
<td>storm 0.017</td>
<td>company 0.018</td>
<td>lord 0.017</td>
<td>organize 0.022</td>
<td>know 0.011</td>
</tr>
<tr>
<td>brown 0.019</td>
<td>resident 0.016</td>
<td>energy 0.017</td>
<td>jesus 0.016</td>
<td>effort 0.020</td>
<td>something 0.007</td>
</tr>
<tr>
<td>blame 0.017</td>
<td>center 0.016</td>
<td>market 0.016</td>
<td>will 0.013</td>
<td>fund 0.019</td>
<td>guess 0.007</td>
</tr>
<tr>
<td>governor 0.014</td>
<td>rescue 0.012</td>
<td>gasoline 0.012</td>
<td>faith 0.012</td>
<td>volunteer 0.019</td>
<td>myself 0.006</td>
</tr>
</tbody>
</table>

- Spatiotemporal patterns
Theme Life Cycles ("Hurricane Katrina")

(a) Theme life cycles in Texas (Hurricane Katrina)

(b) Theme "New Orleans" over states (Hurricane Katrina)

Oil Price

- price 0.0772
- oil 0.0643
- gas 0.0454
- increase 0.0210
- product 0.0203
- fuel 0.0188
- company 0.0182

New Orleans

- city 0.0634
- orleans 0.0541
- new 0.0342
- louisiana 0.0235
- flood 0.0227
- evacuate 0.0211
- storm 0.0177
Theme Snapshots (“Hurricane Katrina”)

Week 1: The theme is the strongest along the Gulf of Mexico.

Week 2: The discussion moves towards the north and west.

Week 3: The theme distributes more uniformly over the states.

Week 4: The theme is again strong along the east coast and the Gulf of Mexico.

Week 5: The theme fades out in most states.
## Multi-Faceted Sentiment Summary

**[Mei et al. 07a]**

(query=“Da Vinci Code”)

<table>
<thead>
<tr>
<th>Facet 1: Movie</th>
<th>Neutral</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>... Ron Howards selection of Tom Hanks to play Robert Langdon.</td>
<td>Tom Hanks stars in the movie, who can be mad at that?</td>
<td>But the movie might get delayed, and even killed off if he loses.</td>
<td></td>
</tr>
<tr>
<td>Directed by: Ron Howard Writing credits: Akiva Goldsman ...</td>
<td>Tom Hanks, who is my favorite movie star act the leading role.</td>
<td>protesting ... will lose your faith by ... watching the movie.</td>
<td></td>
</tr>
<tr>
<td>After watching the movie I went online and some research on ...</td>
<td>Anybody is interested in it?</td>
<td>... so sick of people making such a big deal about a FICTION book and movie.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Facet 2: Book</th>
<th>Neutral</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>I remembered when i first read the book, I finished the book in two days.</td>
<td>Awesome book.</td>
<td>... so sick of people making such a big deal about a FICTION book and movie.</td>
<td></td>
</tr>
<tr>
<td>I'm reading “Da Vinci Code” now. ...</td>
<td>So still a good book to past time.</td>
<td>This controversy book cause lots conflict in west society.</td>
<td></td>
</tr>
</tbody>
</table>
Separate Theme Sentiment Dynamics

“book”

“religious beliefs”
Event Impact Analysis: IR Research

[Mei & Zhai 06b]

Theme: retrieval models

term 0.1599
relevance 0.0752
weight 0.0660
feedback 0.0372
independence 0.0311
model 0.0310
frequent 0.0233
probabilistic 0.0188
document 0.0173
...

vector 0.0514
concept 0.0298
extend 0.0297
model 0.0291
space 0.0236
boolean 0.0151
function 0.0123
feedback 0.0077
...

xml 0.0678
e-mail 0.0197
model 0.0191
collect 0.0187
judgment 0.0102
rank 0.0097
subtopic 0.0079
...

Publication of the paper “A language modeling approach to information retrieval”

SIGIR papers

1992

Starting of the TREC conferences

1998

model 0.1687
language 0.0753
estimate 0.0520
parameter 0.0281
distribution 0.0268
probable 0.0205
smooth 0.0198
markov 0.0137
likelihood 0.0059
...

Theme: retrieval models

term 0.1599
relevance 0.0752
weight 0.0660
feedback 0.0372
independence 0.0311
model 0.0310
frequent 0.0233
probabilistic 0.0188
document 0.0173
...

vector 0.0514
concept 0.0298
extend 0.0297
model 0.0291
space 0.0236
boolean 0.0151
function 0.0123
feedback 0.0077
...

xml 0.0678
e-mail 0.0197
model 0.0191
collect 0.0187
judgment 0.0102
rank 0.0097
subtopic 0.0079
...

Publication of the paper “A language modeling approach to information retrieval”

SIGIR papers

1992

Starting of the TREC conferences

1998

model 0.1687
language 0.0753
estimate 0.0520
parameter 0.0281
distribution 0.0268
probable 0.0205
smooth 0.0198
markov 0.0137
likelihood 0.0059
...
The Author-Topic model
[Rosen-Zvi et al. 04]

Each author has a distribution over topics

\[ \theta^{(a)} \sim \text{Dirichlet}(\alpha) \]

\[ \phi^{(j)} \sim \text{Dirichlet}(\beta) \]

The author of each word is chosen uniformly at random

\[ x_i \sim \text{Uniform}(A^{(d)}) \]

\[ z_i \sim \text{Discrete}(\theta^{(x_i)}) \]

\[ w_i \sim \text{Discrete}(\phi^{(z_i)}) \]
Four example topics from NIPS

<table>
<thead>
<tr>
<th>TOPIC 19</th>
<th>TOPIC 24</th>
<th>TOPIC 29</th>
<th>TOPIC 87</th>
</tr>
</thead>
<tbody>
<tr>
<td>WORD</td>
<td>PROB.</td>
<td>WORD</td>
<td>PROB.</td>
</tr>
<tr>
<td>LIKELIHOOD</td>
<td>0.0539</td>
<td>RECOGNITION</td>
<td>0.0400</td>
</tr>
<tr>
<td>MIXTURE</td>
<td>0.0509</td>
<td>CHARACTER</td>
<td>0.0336</td>
</tr>
<tr>
<td>EM</td>
<td>0.0470</td>
<td>CHARACTERS</td>
<td>0.0250</td>
</tr>
<tr>
<td>DENSITY</td>
<td>0.0398</td>
<td>TANGENT</td>
<td>0.0241</td>
</tr>
<tr>
<td>GAUSSIAN</td>
<td>0.0349</td>
<td>HANDWRITTEN</td>
<td>0.0169</td>
</tr>
<tr>
<td>ESTIMATION</td>
<td>0.0314</td>
<td>DIGITS</td>
<td>0.0159</td>
</tr>
<tr>
<td>LOG</td>
<td>0.0263</td>
<td>IMAGE</td>
<td>0.0157</td>
</tr>
<tr>
<td>MAXIMUM</td>
<td>0.0254</td>
<td>DISTANCE</td>
<td>0.0153</td>
</tr>
<tr>
<td>PARAMETERS</td>
<td>0.0209</td>
<td>DIGIT</td>
<td>0.0149</td>
</tr>
<tr>
<td>ESTIMATE</td>
<td>0.0204</td>
<td>HAND</td>
<td>0.0126</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AUTHOR</th>
<th>PROB.</th>
<th>AUTHOR</th>
<th>PROB.</th>
<th>AUTHOR</th>
<th>PROB.</th>
<th>AUTHOR</th>
<th>PROB.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tresp_V</td>
<td>0.0333</td>
<td>Simard_P</td>
<td>0.0694</td>
<td>Singh_S</td>
<td>0.1412</td>
<td>Smola_A</td>
<td>0.1033</td>
</tr>
<tr>
<td>Singer_Y</td>
<td>0.0281</td>
<td>Martin_G</td>
<td>0.0394</td>
<td>Barto_A</td>
<td>0.0471</td>
<td>Scholkopf_B</td>
<td>0.0730</td>
</tr>
<tr>
<td>Jebara_T</td>
<td>0.0207</td>
<td>LeCun_Y</td>
<td>0.0359</td>
<td>Sutton_R</td>
<td>0.0430</td>
<td>Burges_C</td>
<td>0.0489</td>
</tr>
<tr>
<td>Ghahramani_Z</td>
<td>0.0196</td>
<td>Denker_J</td>
<td>0.0278</td>
<td>Dayan_P</td>
<td>0.0324</td>
<td>Vapnik_V</td>
<td>0.0431</td>
</tr>
<tr>
<td>Ueda_N</td>
<td>0.0170</td>
<td>Henderson_D</td>
<td>0.0256</td>
<td>Parr_R</td>
<td>0.0314</td>
<td>Chapelle_O</td>
<td>0.0210</td>
</tr>
<tr>
<td>Jordan_M</td>
<td>0.0150</td>
<td>Revw_M</td>
<td>0.0229</td>
<td>Dietterich_T</td>
<td>0.0231</td>
<td>Cristianini_N</td>
<td>0.0185</td>
</tr>
<tr>
<td>Roweis_S</td>
<td>0.0123</td>
<td>Platt_J</td>
<td>0.0226</td>
<td>Tsitsiklis_J</td>
<td>0.0194</td>
<td>Ratsch_G</td>
<td>0.0172</td>
</tr>
<tr>
<td>Schuster_M</td>
<td>0.0104</td>
<td>Keeler_J</td>
<td>0.0192</td>
<td>Randlov_J</td>
<td>0.0167</td>
<td>Laskov_P</td>
<td>0.0169</td>
</tr>
<tr>
<td>Xu_L</td>
<td>0.0098</td>
<td>Rashid_M</td>
<td>0.0182</td>
<td>Bradtke_S</td>
<td>0.0161</td>
<td>Tipping_M</td>
<td>0.0153</td>
</tr>
<tr>
<td>Saul_L</td>
<td>0.0094</td>
<td>Sackinger_E</td>
<td>0.0132</td>
<td>Schwartz_A</td>
<td>0.0142</td>
<td>Sollich_P</td>
<td>0.0141</td>
</tr>
</tbody>
</table>
Dirichlet-multinomial Regression (DMR)  
[Mimno & McCallum 08]

Allows arbitrary features to be used to influence choice of topics

Figure 3: The Dirichlet-multinomial Regression (DMR) topic model. Unlike all previous models, the prior distribution over topics, $\alpha$, is a function of observed document features, and is therefore specific to each distinct combination of metadata feature values.
Outline

1. Background
   - Text Mining (TM)
   - Statistical Language Models

2. Basic Topic Models
   - Probabilistic Latent Semantic Analysis (PLSA)
   - Latent Dirichlet Allocation (LDA)
   - Applications of Basic Topic Models to Text Mining

3. Advanced Topic Models
   - Capturing Topic Structures
   - Contextualized Topic Models
   - Supervised Topic Models

4. Summary
Supervised LDA [Blei & McAuliffe 07]

1. Draw topic proportions $\theta | \alpha \sim \text{Dir}(\alpha)$.
2. For each word
   - Draw topic assignment $z_n | \theta \sim \text{Mult}(\theta)$.
   - Draw word $w_n | z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_n})$.
3. Draw response variable $y | Z_{1:N}, \eta, \sigma^2 \sim \text{N}(\eta^T \bar{Z}, \sigma^2)$, where
   \[
   \bar{Z} = \frac{1}{N} \sum_{n=1}^{N} z_n.
   \]
Sample Results of Supervised LDA

- 10-topic sLDA model on movie reviews (Pang and Lee, 2005).
- Response: number of stars associated with each review
Latent Aspect Rating Analysis [Wang et al. 11]

• Given a set of review articles about a topic with overall ratings (ratings as “supervision signals”)

• Output
  – Major aspects commented on in the reviews
  – Ratings on each aspect
  – Relative weights placed on different aspects by reviewers

• Many applications
  – Opinion-based entity ranking
  – Aspect-level opinion summarization
  – Reviewer preference analysis
  – Personalized recommendation of products
  – …
An Example of LARA

Hotel Palomar Chicago: Traveler Reviews

"Great location + spacious room = happy traveler"

- leos_10 • 3 contributions
- Boston
- Jul 11, 2010 | Trip type: Couples

Stayed for a weekend in July. Walked everywhere, enjoyed the comfy bed and quiet hallways. more

"Terrific service and gorgeous facility"

- a_hickling • 1 contribution
- Greensboro, North Carolina
- Jul 7, 2010 | Trip type: Family

I stayed at the Palomar with my young daughter for three nights June 17-20, 2010 and absolutely loved the hotel. The room was one of the nicest I’ve ever stayed in (My daughter loved the Fuji jetted tub so much that she wanted to take 2 baths a day!) in terms of decor, design, and size. (It compared favorably to... more

How to infer aspect ratings?

My ratings for this hotel

- Value
- Rooms
- Location
- Cleanliness

How to infer aspect weights?

Value | Location | Service

---|---|---

...... Value | Location | Service

......
Excellent location in walking distance to Tiananmen Square and shopping streets. That's the best part of this hotel! The rooms are getting really old. Bathroom was nasty. The fixtures were falling off, lots of cracks and everything looked dirty. I don’t think it worth the price. Service was the most disappointing part, especially the door men. this is not how you treat guests, this is not hospitality.

“Spend your money elsewhere”
Excellent location in walking distance to Tiananmen Square and shopping streets. That’s the best part of this hotel! The rooms are getting really old. Bathroom was nasty. The fixtures were falling off, lots of cracks and everything looked dirty. I don’t think it worth the price. Service was the most disappointing part, especially the door men. This is not how you treat guests, this is not hospitality.
Aspect Identification

- Amazon reviews: no guidance

<table>
<thead>
<tr>
<th>Table 2: Topical Aspects Learned on MP3 Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low Overall Ratings</strong></td>
</tr>
<tr>
<td>unit</td>
</tr>
<tr>
<td>usb</td>
</tr>
<tr>
<td>battery</td>
</tr>
<tr>
<td>charger</td>
</tr>
<tr>
<td>reset</td>
</tr>
<tr>
<td>time</td>
</tr>
<tr>
<td>hours</td>
</tr>
<tr>
<td>work</td>
</tr>
<tr>
<td>thing</td>
</tr>
<tr>
<td>wall</td>
</tr>
</tbody>
</table>

**battery life accessory service file format volume video**
Network Supervised Topic Modeling [Mei et al. 08]

• Probabilistic topic modeling as an optimization problem (e.g., PLSA/LDA: Maximum Likelihood):

\[ O(Collection \mid Model) = \log(P(Collection \mid Model)) \]

• Regularized objective function with network constrains
  – Topic distribution are smoothed over adjacent vertices

\[ O(Collection, Network \mid Model) = \log(P(Collection \mid Model)) \oplus Regularizer(Model, Network) \]

\[ ModelParams = \arg \max_{\text{params}} O(Collection[, Network] \mid Model) \]

• Flexibility in selecting topic models and regularizers
Instantiation: NetPLSA

- **Basic Assumption: Neighbors have similar topic distribution**

\[
O(C,G) = (1 - \lambda) \cdot \left( \sum_d \sum_w c(w,d) \log \sum_{j=1}^k p(\theta_j | d) p(w | \theta_j) \right) \\
+ \lambda \left( -\frac{1}{2} \sum_{(u,v) \in E} w(u,v) \sum_{j=1}^k (p(\theta_j | u) - p(\theta_j | v))^2 \right)
\]

Graph Harmonic Regularizer, Generalization of [Zhu ’03],

\[
= \frac{1}{2} \sum_{j=1,...,k} f_j^T \Delta f_j, \text{where } f_{j,u} = p(\theta_j | u)
\]
### Topical Communities with PLSA

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>term</td>
<td>peer</td>
<td>visual</td>
<td>interface</td>
</tr>
<tr>
<td>question</td>
<td>patterns</td>
<td>analog</td>
<td>towards</td>
</tr>
<tr>
<td>protein</td>
<td>mining</td>
<td>neurons</td>
<td>browsing</td>
</tr>
<tr>
<td>training</td>
<td>clusters</td>
<td>vlsi</td>
<td>xml</td>
</tr>
<tr>
<td>weighting</td>
<td>stream</td>
<td>motion</td>
<td>generation</td>
</tr>
<tr>
<td>multiple</td>
<td>frequent</td>
<td>chip</td>
<td>design</td>
</tr>
<tr>
<td>recognition</td>
<td>e</td>
<td>natural</td>
<td>engine</td>
</tr>
<tr>
<td>relations</td>
<td>page</td>
<td>cortex</td>
<td>service</td>
</tr>
<tr>
<td>library</td>
<td>gene</td>
<td>spike</td>
<td>social</td>
</tr>
</tbody>
</table>

#### Noisy community assignment
Topical Communities with NetPLSA

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieval</td>
<td>mining</td>
<td>neural</td>
<td>web</td>
</tr>
<tr>
<td>information</td>
<td>data</td>
<td>learning</td>
<td>services</td>
</tr>
<tr>
<td>document</td>
<td>discovery</td>
<td>networks</td>
<td>semantic</td>
</tr>
<tr>
<td>query</td>
<td>databases</td>
<td>recognition</td>
<td>services</td>
</tr>
<tr>
<td>text</td>
<td>rules</td>
<td>analog</td>
<td>peer</td>
</tr>
<tr>
<td>search</td>
<td>association</td>
<td>vlsi</td>
<td>ontologies</td>
</tr>
<tr>
<td>evaluation</td>
<td>patterns</td>
<td>neurons</td>
<td>rdf</td>
</tr>
<tr>
<td>user</td>
<td>frequent</td>
<td>gaussian</td>
<td>management</td>
</tr>
<tr>
<td>relevance</td>
<td>streams</td>
<td>network</td>
<td>ontology</td>
</tr>
</tbody>
</table>
Outline

1. Background
   - Text Mining (TM)
   - Statistical Language Models

2. Basic Topic Models
   - Probabilistic Latent Semantic Analysis (PLSA)
   - Latent Dirichlet Allocation (LDA)
   - Applications of Basic Topic Models to Text Mining

3. Advanced Topic Models
   - Capturing Topic Structures
   - Contextualized Topic Models
   - Supervised Topic Models

4. Summary
   - We are here
Summary

• Statistical Topic Models (STMs) are a new family of language models, especially useful for
  – Discovering latent topics in text
  – Analyzing latent structures and patterns of topics
  – Extensible for joint modeling and analysis of text and associated non-textual data

• PLSA & LDA are two basic topic models that tend to function similarly, with LDA better as a generative model

• Many different models have been proposed with probably many more to come

• Many demonstrated applications in multiple domains and many more to come
Summary (cont.)

• However, all topic models suffer from the problem of multiple local maxima
  – Make it hard/impossible to reproduce research results
  – Make it hard/impossible to interpret results in real applications

• Complex models can’t scale up to handle large amounts of text data
  – Collapsed Gibbs sampling is efficient, but only working for conjugate priors
  – Variational EM needs to be derived in a model-specific way
  – Parallel algorithms are promising

• Many challenges remain….
Challenges and Future Directions

• Challenge 1: How can we quantitatively evaluate the benefit of topic models for text mining?
  – Currently, most quantitative evaluation is based on perplexity which doesn’t reflect the actual utility of a topic model for text mining
  – Need to separately evaluate the quality of both topic word distributions and topic coverage
  – Need to consider multiple aspects of a topic (e.g., coherent?, meaningful?) and define appropriate measures
  – Need to compare topic models with alternative approaches to solving the same text mining problem (e.g., traditional IR methods, non-negative matrix factorization)
  – Need to create standard test collections
Challenges and Future Directions (cont.)

- Challenge 2: How can we help users interpret a topic?
  - Most of the time, a topic is manually labeled in a research paper; this is insufficient for real applications
  - Automatic labeling can help, but the utility still needs to be evaluated
  - Need to generate a summary for a topic to enable a user to navigate into text documents to better understand a topic
  - Need to facilitate post-processing of discovered topics (e.g., ranking, comparison)
Challenges and Future Directions (cont.)

• Challenge 3: How can we address the problem of multiple local maxima?
  – All topic models have the problem of multiple local maxima, causing problems with reproducing results
  – Need to compute the variance of a discovered topic
  – Need to define and report the confidence interval for a topic

• Challenge 4: How can we develop efficient estimation/inference algorithms for sophisticated models?
  – How can we leverage a user’s knowledge to speed up inferences for topic models?
  – Need to develop parallel estimation/inference algorithms
Challenges and Future Directions (cont.)

• Challenge 5: How can we incorporate linguistic knowledge into topic models?
  – Most current topic models are purely statistical
  – Some progress has been made to incorporate linguistic knowledge (e.g., [Griffiths et al. 04, Wallach 08])
  – More needs to be done

• Challenge 6: How can we incorporate domain knowledge and preferences from an analyst into a topic model to support complex text mining tasks?
  – Current models are mostly pre-specified with little flexibility for an analyst to “steer” the analysis process
  – Need to develop a general analysis framework to enable an analyst to use multiple topic models together to perform complex text mining tasks
References (incomplete)


[Mei et al. 05] Qiaozhu Mei, ChengXiang Zhai: Discovering evolutionary theme patterns from text: an exploration of temporal text mining. KDD 2005: 198-207

[Mei et al. 06a] Qiaozhu Mei, Chao Liu, Hang Su, ChengXiang Zhai: A probabilistic approach to spatiotemporal theme pattern mining on weblogs. WWW 2006: 533-542
References (incomplete)


