An Introduction to Topic Modeling

Daniel W. Peterson

Department of Computer Science
University of Colorado at Boulder
daniel.w.peterson@colorado.edu

April 24, 2013
Latent Semantic Analysis

- Documents x Terms matrix: large and sparse
- Use SVD to decompose it into three matrices
- Keep only the “important” dimensions
- Assumptions:
  - Word order doesn’t matter
  - Words are orthogonal dimensions in a high-dimensional space
Documents are generated by a probabilistic process
- Structure based on topics
- Different topics make different words more likely

Assumptions:
- Word order doesn’t matter
- Each word is chosen as the result of exactly one topic
N documents
A document is L words long
Each entry has an assignment to one of K topics
How do we choose a topic?

\[ D \rightarrow k \rightarrow w \]

\[ j=1..L \]

\[ i=1..N \]
Probabilistic Latent Semantic Analysis

- How do we choose a topic? *We sample from a distribution over topics.*

- How do we choose a word?
How do we choose a topic?  
*We sample from a distribution over topics.*

How do we choose a word?  
*We sample from a distribution over words.*
Multinomial Distribution

- Select one of several possible outcomes

Looks like: a $1 \times n$ vector of probabilities

\[
\begin{align*}
  x_1 + x_2 + \cdots + x_n &= 1 \\
  x_i &> 0
\end{align*}
\]

A sample looks like: a number

The outcome of rolling the dice

Probability we get $i$ is given by $x_i$
Multinomial Distribution

- Select one of several possible outcomes
- Outcomes may be equally likely (like dice)
Multinomial Distribution

- Select one of several possible outcomes
- Outcomes may be equally likely (like dice)
- OR: some outcomes may be more likely than others (load the dice)
Multinomial Distribution

- Select one of several possible outcomes
- Outcomes may be equally likely (like dice)
- OR: some outcomes may be more likely than others (load the dice)

Looks like: a $1 \times n$ vector of probabilities

- $[x_1, x_2, \ldots, x_n]$
- $x_1 + x_2 + \ldots + x_n = 1$
- every $x_i > 0$
Multinomial Distribution

- Select one of several possible outcomes
- Outcomes may be equally likely (like dice)
- OR: some outcomes may be more likely than others (load the dice)

Looks like: a $1 \times n$ vector of probabilities
- $[x_1, x_2, \ldots, x_n]$
- $x_1 + x_2 + \ldots + x_n = 1$
- every $x_i > 0$

A sample looks like: a number
- The outcome of rolling the dice
- Probability we get $i$ is given by $x_i$
\( \theta \) is a distribution over topics in a document

- One \( \theta \) for each document
- \( \theta \) is a \( 1 \times K \) vector
- Sum of \( \theta \) is 1
\( \theta \) is a distribution over topics in a document
- One \( \theta \) for each document
- \( \theta \) is a \( 1 \times K \) vector
- Sum of \( \theta \) is 1

\( \phi \) is a distribution over words in a topic
- One \( \phi \) for each topic
- \( \phi \) is a \( 1 \times W \) vector
- Sum of \( \phi \) is 1
Fold $\theta$ into graphical model
- Fold $\theta$ into graphical model

- Where do $\theta$ and $\phi$ come from?
Sample $\theta$ and $\phi$ from an appropriate distribution
Sample $\theta$ and $\phi$ from an appropriate distribution

Dirchlet: a distribution over distributions
Sample $\theta$ and $\phi$ from an appropriate distribution

Dirchlet: a distribution over distributions

Incorporating Dirichlet prior provides smoothing
Dirichlet Distribution

- Takes $n$ parameters $\alpha_1, \alpha_2, \ldots, \alpha_n$
- Distribution over $1 \times n$ vectors with sum of 1
- $\alpha_i$ are called concentration parameters
Dirichlet Distribution with 2 Parameters

Figure: Image source: Wikipedia
Dirichlet Distribution with 3 Parameters

Figure: Image source: Yee Whye Teh
A Sample from a Dirichlet

- A particular $1 \times n$ vector with sum of 1
A Sample from a Dirichlet

- A particular $1 \times n$ vector with sum of 1
- $[x_1, x_2, \ldots, x_n]$ such that $x_1 + x_2 + \ldots + x_n = 1$
- every $x_i > 0$
A Sample from a Dirichlet

- A particular $1 \times n$ vector with sum of 1
- $[x_1, x_2, \ldots, x_n]$ such that $x_1 + x_2 + \ldots + x_n = 1$
- every $x_i > 0$
- A multinomial distribution
Sample $\theta$ and $\phi$ from a Dirichlet distribution

This is important for when we turn the model around:
Sample $\theta$ and $\phi$ from a Dirichlet distribution

This is important for when we turn the model around:

Dirichlet distribution is conjugate prior of multinomial:

Given a Dirichlet prior, and counts of topic assignments, the posterior is also Dirichlet
Sample $\theta$ and $\phi$ from a Dirichlet distribution.

This is important when we turn the model around:

Dirichlet distribution is conjugate prior of multinomial:

Given a Dirichlet prior, and counts of topic assignments, the posterior is also Dirichlet.

$\beta$ and $\gamma$ are smoothing parameters.
Inference

- Generative model explains how the data was created
Inference

- Generative model explains how the data was created
- Inference: trying to guess model parameters
Gibbs Sampling

- Hard to determine most likely model parameters

Daniel Peterson (University of Colorado)
Gibbs Sampling

- Hard to determine most likely model parameters
- Hard for even relatively likely parameters
Gibbs Sampling

- Hard to determine most likely model parameters
- Hard for even relatively likely parameters
- Can’t sample from overall distribution: sample instead a single variable
Gibbs Sampling

- Hard to determine most likely model parameters
- Hard for even relatively likely parameters
- Can’t sample from overall distribution: sample instead a single variable
- Take a walk through distribution
Gibbs Sampling

- Hard to determine most likely model parameters
- Hard for even relatively likely parameters
- Can’t sample from overall distribution: sample instead a single variable
- Take a walk through distribution
  - One step (parameter) at a time
Gibbs Sampling

- Hard to determine most likely model parameters
- Hard for even relatively likely parameters
- Can’t sample from overall distribution: sample instead a single variable
- Take a walk through distribution
  - One step (parameter) at a time
  - Spend more time walking around more likely areas
Gibbs Sampling

- Hard to determine most likely model parameters
- Hard for even relatively likely parameters
- Can’t sample from overall distribution: sample instead a single variable
- Take a walk through distribution
  - One step (parameter) at a time
  - Spend more time walking around more likely areas
  - We can get to likely areas from anywhere
Gibbs Sampling

- Hard to determine most likely model parameters
- Hard for even relatively likely parameters
- Can’t sample from overall distribution: sample instead a single variable
- Take a walk through distribution
  - One step (parameter) at a time
  - Spend more time walking around more likely areas
  - We can get to likely areas from anywhere
  - It doesn’t matter where we start!
Gibbs Sampling in a Topic Model

- Start with random assignment of topics

![Diagram of Gibbs Sampling in a Topic Model]

\[ \text{Repeat the above many times} \]

Smoothing (\(\beta\) and \(\gamma\)) very important

Daniel Peterson (University of Colorado)
- Start with random assignment of topics
- For each \(<\text{word, document}>\) pair:

\[
\begin{align*}
\text{Sample } \theta & \text{ based on counts and prior} \\
\text{Sample } \phi & \text{ based on counts and prior} \\
\text{Choose } k & \text{ based on } \theta, \phi, \text{ and } w
\end{align*}
\]
Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- For each \(<word, document>\) pair:
  - Sample \(\theta\) based on counts and prior
  - Choose \(k\) based on \(\theta\), \(\phi\), and \(w\)
  - Repeat the above many times

Smoothing (\(\beta\) and \(\gamma\)) is very important.
Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- For each \(<\text{word, document}>\) pair:
  - Sample \(\theta\) based on counts and prior
  - Sample \(\phi\) based on counts and prior

\[
\begin{align*}
\beta & \\
\theta & \\
k & \\
w & \\
j=1..L & \\
i=1..N & \\
\phi & \rightarrow Y
\end{align*}
\]
Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- For each \(<word, document>\) pair:
  - Sample \(\theta\) based on counts and prior
  - Sample \(\phi\) based on counts and prior
  - Choose \(k\) based on \(\theta\), \(\phi\), and \(w\)

\[\beta\]
\[\theta\]
\[k\]
\[w\]
\[j=1..L\]
\[i=1..N\]
\[\phi\]
\[j=1..L\]
\[Y\]
Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- For each \(<\text{word}, \text{document}>\) pair:
  - Sample \(\theta\) based on counts and prior
  - Sample \(\phi\) based on counts and prior
  - Choose \(k\) based on \(\theta\), \(\phi\), and \(w\)
- Repeat the above many times

\[ \begin{array}{c}
\beta \\
\theta \\
k \\
w \\
\phi \\
y \\
\end{array} \]
Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- For each `<word, document>` pair:
  - Sample $\theta$ based on counts and prior
  - Sample $\phi$ based on counts and prior
  - Choose $k$ based on $\theta$, $\phi$, and $w$
- Repeat the above many times
- Smoothing ($\beta$ and $\gamma$) very important
Bayes Rule

\[ P(k|\beta, X) \propto P(k|\beta)P(X|k) \]

Sampling from a conditional distribution can be broken down into sampling based on the parent nodes (prior, \( \beta \)) and the children (likelihood, \( X \))
• Start with random assignment of topics
- Start with random assignment of topics
- Repeat many times:

- Sample all $\theta$ and $\phi$ from counts and prior
- Choose $k$ for a number of $\langle \text{word}, \text{document} \rangle$ pairs
- More sampling, less counting
Blocked Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- Repeat many times:
  - Sample all $\theta$ and $\phi$ from counts and prior

\[
\begin{align*}
\beta & \rightarrow \theta \\
\theta & \rightarrow k \\
k & \rightarrow w \\
w & \rightarrow \phi \\
\phi & \rightarrow \gamma
\end{align*}
\]
Start with random assignment of topics

Repeat many times:
- Sample all $\theta$ and $\phi$ from counts and prior
- Choose $k$ for a number of $\langle \text{word, document} \rangle$ pairs
Blocked Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- Repeat many times:
  - Sample all $\theta$ and $\phi$ from counts and prior
  - Choose $k$ for a number of $<\text{word, document}>$ pairs
- More sampling, less counting
Integrate out $\theta$ and $\phi$
Integrate out $\theta$ and $\phi$

Start with random assignment of topics
Integrate out $\theta$ and $\phi$
Start with random assignment of topics
For each $\langle \text{word}, \text{document} \rangle$ pair:

\[
P(z_i = k | z_{-i}, w) \propto n(w_i) - i, k + \gamma n(\cdot) - i, k + W\gamma n(d_i) - i, k + \beta n(\cdot) + K\beta
\]
Integrate out $\theta$ and $\phi$

Start with random assignment of topics

For each $<\text{word}, \text{document}>$ pair:
- Sample $k$ directly from counts

\[
P(z_i = k | z_{-i}, w) \propto \frac{n_{i,k}^{(w_i)} + \gamma}{n^{(\cdot)}_{-i,k} + W \gamma} \frac{n_{i,k}^{(d_i)} + \beta}{n^{(\cdot)}_{-i} + K \beta}
\]
Integrate out $\theta$ and $\phi$
Start with random assignment of topics
For each $<\text{word}, \text{document}>$ pair:
  - Sample $k$ directly from counts
Repeat many times