Comparing Text Corpora (2)

Instructor: Björn Ross

08-Nov-2023
LDA Overview
Background: Plate Notation
Background: Plate Notation

Make a basket
Background: Plate Notation

Basketball shooting accuracy ➔ Make a basket
Background: Plate Notation

Basketball shooting accuracy

- Make first basket
- Make second basket
- Make third basket
Background: Plate Notation

Basketball shooting accuracy ➔ Make nth basket
Latent Dirichlet Allocation

- Let’s start with a very simple model
- We will work our way up to the full LDA model
Unigram Model

w is a word
N words in a document

Figure from Blei et al 2003
Unigram Model

w is a word
N words in a document
M documents in a corpus

Figure from Blei et al 2003
Unigram Model

\( w \) is a word

N words in a document

M documents in a corpus

\( w \) is a vector of words (i.e. doc)

\[
p(w) = \prod_{n=1}^{N} p(w_n)
\]

Figure from Blei et al 2003
Probability with a Unigram Model

\[ p(w) = \prod_{n=1}^{N} p(w_n) \]

What is the probability of the example sentence?

“My dog barked at another dog.”

<table>
<thead>
<tr>
<th>word</th>
<th>my</th>
<th>at</th>
<th>dog</th>
<th>another</th>
<th>barked</th>
</tr>
</thead>
<tbody>
<tr>
<td>probability</td>
<td>.10</td>
<td>.10</td>
<td>.05</td>
<td>.04</td>
<td>.03</td>
</tr>
</tbody>
</table>
Probability with a Unigram Model

\[ p(w) = \prod_{n=1}^{N} p(w_n) \]

<table>
<thead>
<tr>
<th>word</th>
<th>my</th>
<th>at</th>
<th>dog</th>
<th>another</th>
<th>barked</th>
</tr>
</thead>
<tbody>
<tr>
<td>probability</td>
<td>.1</td>
<td>.1</td>
<td>.05</td>
<td>.04</td>
<td>.03</td>
</tr>
</tbody>
</table>

Solution:

My dog barked at another dog.

\[ .1 * .05 * .03 * .1 * .04 * .05 = 3e-8 \]
Unigram Model...

• What is the point of making these models more complex?
• Why not just use the basic unigram model for everything?

• Remember:
  • Higher text probability doesn’t imply a better model
  • We want to accurately describe the data
    • → higher probability for real documents, lower probability for noise
Mixture of Unigrams Model

z is the topic of a document

Figure from Blei et al 2003
Mixture of Unigrams Model

\[ p(w) = \sum_z p(z) \prod_{n=1}^{N} p(w_n | z) \]

z is the topic of a document

Figure from Blei et al 2003
Probability with Mixture of Unigrams

\[ p(w) = \sum_z p(z) \prod_{n=1}^{N} p(w_n | z). \]

What is the probability of the sentence?
Ignore stopwords: “my”, “after”, “the”

“My dog chased after the bus.”

<table>
<thead>
<tr>
<th>( w_i )</th>
<th>cat</th>
<th>dog</th>
<th>chased</th>
<th>car</th>
<th>bus</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(w_i</td>
<td>z = \text{pets}) )</td>
<td>.20</td>
<td>.30</td>
<td>.10</td>
<td>.01</td>
</tr>
<tr>
<td>( P(w_i</td>
<td>z = \text{vehicles}) )</td>
<td>.01</td>
<td>.01</td>
<td>.10</td>
<td>.30</td>
</tr>
</tbody>
</table>

\[ p(z = \text{pets}) = 0.6, \]
\[ p(z = \text{vehicles}) = 0.4 \]
## Probability with Mixture of Unigrams

### Solution:

My dog chased after the bus.

\[ .6 \times ( .3 \times .1 ) \times .01 = .00018 \]
\[ .4 \times ( .01 \times .1 ) \times .2 = .00008 \]

Total = .00026
Probabilistic Latent Semantic Indexing

d is a document ID

Figure from Blei et al 2003
Probabilistic Latent Semantic Indexing

d is a document ID

\[ p(d, w_n) = p(d) \sum_z p(w_n | z) p(z | d) \]

Figure from Blei et al 2003
Probability with pLSI

\[ d_1 \text{ “The cat sat down.”} \]

\[
\begin{array}{c|cccccc}
 w_i & \text{cat} & \text{sat} & \text{down} & \text{car} & \text{broke} \\
 p(w_i|z = t_1) & .2 & .1 & .05 & .01 & .1 \\
 p(w_i|z = t_2) & .01 & .05 & .1 & .3 & .1 \\
\end{array}
\]

\[
\begin{array}{c|c}
p(d = d_1) & .01 \\
p(z = t_1|d = d_1) & .6 \\
p(z = t_2|d = d_1) & .4 \\
\end{array}
\]

What is the joint probability of the document and the word “cat”? 

Björn Ross, TTDS 2023/2024
Probability with pLSI

\[ p(d, w_n) = p(d) \sum_z p(w_n \mid z) p(z \mid d) \]

Solution:

The \textbf{cat} sat down.

\[ 0.01 \times (0.2 \times 0.6 + 0.01 \times 0.4) = 0.00124 \]
Latent Dirichlet Allocation

$\theta$ is the distribution over topics in a document

$\alpha$ is the parameter of a Dirichlet distribution giving possible topic distributions within documents

$\beta$ gives word distributions within topics

Figure from Blei et al 2003
Latent Dirichlet Allocation

\[
p(\theta, z, w | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^{N} p(z_n | \theta) p(w_n | z_n, \beta)
\]

Figure from Blei et al 2003
Latent Dirichlet Allocation

\[ p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \]

\[ p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left( \prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta \]

\[ p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left( \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d \]
Model Inference

- Want to learn the model parameters
- Exact inference becomes intractable

Diagram:

- Corpus
- Topic Modeling Method
- Document-Topic Matrix
- Topic-Word Matrix

$k$, $d$, $v$, $k$
Model Inference

• Instead, use an approximate method such as:
  • Gibbs sampling
  • Variational Inference
Gibbs Sampling for LDA

Goal: Learn $\Phi$, $\theta$ given a set of documents $D$

$\Phi = \text{topic-word probabilities}$
$\theta = \text{document-topic probabilities}$

Known:
corpus, $\alpha$, $\beta$ and the probability that a word is from a topic conditional on the assignments of all other words to topics

$$P(z_i = j \mid z_{-i}, w_i, d_i, \cdot) \propto \frac{C_{w_ij}^{WT} + \beta}{\sum_{w=1}^{W} C_{w_j}^{WT} + W \beta} \frac{C_{d_ij}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_it}^{DT} + T \alpha}$$

Note: the $\propto$ symbol means “proportional to”
Gibbs Sampling for LDA

Want to learn $\Phi$, $\theta$ given a set of documents $D$

1. Assign each word a topic randomly
2. Calculate count matrices
3. Repeat until convergence:
   • For every document $d$
     • For every word $i$
       • Decrement count matrices $C^{WT}$ and $C^{DT}$ for current topic assignment
       • Sample a new topic assignment
       • Increment count matrices $C^{WT}$ and $C^{DT}$ for new topic assignment
4. Calculate $\Phi$ and $\theta$
Gibbs Sampling for LDA

d1 Green eggs and ham.
d2 Ham and green peppers.
d3 Ham and cheese.

Green eggs and ham.
Ham and green peppers.
Ham and cheese.

Random initialization.
# Gibbs Sampling for LDA

$$C^{WT}$$

<table>
<thead>
<tr>
<th></th>
<th>green</th>
<th>eggs</th>
<th>and</th>
<th>ham</th>
<th>peppers</th>
<th>cheese</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>t2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Green eggs and ham.
Ham and green peppers.
Ham and cheese.

$$C^{DT}$$

<table>
<thead>
<tr>
<th></th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>t2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
## Gibbs Sampling for LDA

Assume (for the moment) $\alpha = \beta = 0$

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>green</th>
<th>eggs</th>
<th>and</th>
<th>ham</th>
<th>peppers</th>
<th>cheese</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>t2</td>
<td>0.20</td>
<td>0.00</td>
<td>0.40</td>
<td>0.40</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**Green eggs and ham.**
**Ham and green peppers.**
**Ham and cheese.**

<table>
<thead>
<tr>
<th>$\Phi$</th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>0.50</td>
<td>0.50</td>
<td>0.66</td>
</tr>
<tr>
<td>t2</td>
<td>0.50</td>
<td>0.50</td>
<td>0.33</td>
</tr>
</tbody>
</table>
Gibbs Sampling for LDA

<table>
<thead>
<tr>
<th>$C^{WT}$</th>
<th>green</th>
<th>eggs</th>
<th>and</th>
<th>ham</th>
<th>peppers</th>
<th>cheese</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>t2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Green eggs and ham.  
Ham and green peppers.  
Ham and cheese.

<table>
<thead>
<tr>
<th>$C^{DT}$</th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>t2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
## Gibbs Sampling for LDA

### $C^{WT}$

<table>
<thead>
<tr>
<th></th>
<th>green</th>
<th>eggs</th>
<th>and</th>
<th>ham</th>
<th>peppers</th>
<th>cheese</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>t2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- Green eggs and ham.
- Ham and green peppers.
- Ham and cheese.

### $C^{DT}$

<table>
<thead>
<tr>
<th></th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>t2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Gibbs Sampling for LDA

Assume (for the moment) $\alpha = \beta = 0$

$$\begin{align*}
\frac{C_{w,j}^{WT} + \beta}{\sum_{w=1}^{W} C_{w,j}^{WT} + W \beta} & \quad \frac{C_{d,t}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d,t}^{DT} + T \alpha}
\end{align*}$$

<table>
<thead>
<tr>
<th>$C^{WT}$</th>
<th>green</th>
<th>eggs</th>
<th>and</th>
<th>ham</th>
<th>peppers</th>
<th>cheese</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>t2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Green eggs and ham. Ham and green peppers. Ham and cheese.

<table>
<thead>
<tr>
<th>$C^{DT}$</th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>t2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
# Gibbs Sampling for LDA

**Green eggs and ham.**
Ham and green peppers.
Ham and cheese.

<table>
<thead>
<tr>
<th>$C^{WT}$</th>
<th>green</th>
<th>eggs</th>
<th>and</th>
<th>ham</th>
<th>peppers</th>
<th>cheese</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>t2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$C^{DT}$</th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>t2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Gibbs Sampling for LDA

Green \text{green eggs} and ham.
Ham and green peppers.
Ham and cheese.

\begin{array}{|c|c|c|c|c|c|}
\hline
C^W & \text{green} & \text{eggs} & \text{and} & \text{ham} & \text{peppers} \\
\hline
\text{t1} & 0 & 1 & 1 & 1 & 1 \\
\hline
\text{t2} & 2 & 0 & 2 & 2 & 0 \\
\hline
\end{array}

\begin{array}{|c|c|c|}
\hline
C^D & \text{d1} & \text{d2} & \text{d3} \\
\hline
\text{t1} & 1 & 2 & 2 \\
\hline
\text{t2} & 3 & 2 & 1 \\
\hline
\end{array}
Gibbs Sampling for LDA

Assume (for the moment) $\alpha = \beta = 0$

\[
\begin{array}{cccccc}
C^W_T & \text{green} & \text{eggs} & \text{and} & \text{ham} & \text{peppers} & \text{cheese} \\
t1 & 0 & 0 & 1 & 1 & 1 & 1 \\
t2 & 2 & 0 & 2 & 2 & 0 & 0 \\
\end{array}
\]

Green \text{eggs} and ham.
Ham and green peppers.
Ham and cheese.

\[
\begin{array}{ccc}
C^D_T & d1 & d2 & d3 \\
t1 & 0 & 2 & 2 \\
t2 & 3 & 2 & 1 \\
\end{array}
\]
Gibbs Sampling for LDA

\[
\frac{C_{w_i,j}^{WT} + \beta}{\sum_{w=1}^{W} C_{w_i,j}^{WT} + W \beta} \quad \frac{C_{d_i,t}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_i,t}^{DT} + T \alpha}
\]

<table>
<thead>
<tr>
<th>(C^{WT} + \alpha)</th>
<th>green</th>
<th>eggs</th>
<th>and</th>
<th>ham</th>
<th>peppers</th>
<th>cheese</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>0.01</td>
<td>0.01</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>t2</td>
<td>2.01</td>
<td>0.01</td>
<td>2.01</td>
<td>2.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Green **eggs** and ham.
Ham and **green** peppers.
Ham and cheese.

<table>
<thead>
<tr>
<th>(C^{DT} + \beta)</th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>0.01</td>
<td>2.01</td>
<td>2.01</td>
</tr>
<tr>
<td>t2</td>
<td>3.01</td>
<td>2.01</td>
<td>1.01</td>
</tr>
</tbody>
</table>
Gibbs Sampling for LDA

• Repeat until convergence

• Probabilistic algorithm – results depend on random initialisation and random samples!
Topic Modeling Examples
## What do students look for in a professor?

<table>
<thead>
<tr>
<th>Topic</th>
<th>Sample words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approachability</td>
<td>prof, fair, clear, helpful, teaching, approachable, nice, organized, extremely, friendly, super, amazing</td>
</tr>
<tr>
<td>Clarity</td>
<td>understand, hard, homework, office, material, clear, helpful, problems, explains, accent, questions, extremely</td>
</tr>
<tr>
<td>Course Logistics</td>
<td>book, study, boring, extra, nice, credit, lot, hard, attendance, make, fine, attention, pay, mandatory</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>teaching, passionate, awesome, enthusiastic, professors, loves, cares, wonderful, fantastic, passion</td>
</tr>
<tr>
<td>Expectations</td>
<td>hard, work, time, lot, comments, tough, expects, worst, stuff, avoid, horrible, classes</td>
</tr>
<tr>
<td>Helpfulness</td>
<td>helpful, nice, recommend, cares, super, understanding, kind, extremely, effort, sweet, friendly, approachable</td>
</tr>
<tr>
<td>Humor</td>
<td>guy, funny, fun, awesome, cool, entertaining, humor, hilarious, jokes, stories, love, hot, enjoyable</td>
</tr>
<tr>
<td>Interestingness</td>
<td>interesting, material, recommend, lecturer, engaging, classes, knowledgeable, enjoyed, loved, topics</td>
</tr>
<tr>
<td>Readings/Discussions</td>
<td>readings, papers, writing, ta, interesting, discussions, grader, essays, boring, books, participation</td>
</tr>
<tr>
<td>Study Material</td>
<td>exams, notes, questions, material, textbook, hard, slides, study, answer, clear, tricky, attend, long, understand</td>
</tr>
</tbody>
</table>

Azab, Mihalcea, and Abernathy, 2016
### What do students look for in a professor?

<table>
<thead>
<tr>
<th></th>
<th>Sociology</th>
<th>Computer Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enthusiasm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interestingness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Course Logistics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clarity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Helpfulness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study Material</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Readings/Discussions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approachability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expectations</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Azab, Mihalcea, and Abernathy, 2016
What do students look for in a professor?

Azab, Mihalcea, and Abernathy, 2016
How do personal attributes relate to values?

<table>
<thead>
<tr>
<th>Theme</th>
<th>Example Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respect others</td>
<td>people, respect, care, human, treat</td>
</tr>
<tr>
<td>Religion</td>
<td>god, heart, belief, religion, right</td>
</tr>
<tr>
<td>Family</td>
<td>family, parent, child, husband, mother</td>
</tr>
<tr>
<td>Hard Work</td>
<td>hard, work, better, honest, best</td>
</tr>
<tr>
<td>Time &amp; Money</td>
<td>money, work, time, day, year</td>
</tr>
<tr>
<td>Problem solving</td>
<td>consider, decision, situation, problem</td>
</tr>
<tr>
<td>Relationships</td>
<td>family, friend, relationship, love</td>
</tr>
<tr>
<td>Optimism</td>
<td>enjoy, happy, positive, future, grow</td>
</tr>
<tr>
<td>Honesty</td>
<td>honest, truth, lie, trust, true</td>
</tr>
<tr>
<td>Rule following</td>
<td>moral, rule, principle, follow</td>
</tr>
<tr>
<td>Societal</td>
<td>society, person, feel, thought, quality</td>
</tr>
<tr>
<td>Personal Growth</td>
<td>personal, grow, best, decision, mind</td>
</tr>
<tr>
<td>Achievement</td>
<td>heart, achieve, complete, goal</td>
</tr>
<tr>
<td>Principles</td>
<td>important, guide, principle, central</td>
</tr>
<tr>
<td>Experiences</td>
<td>look, see, experience, choose, feel</td>
</tr>
</tbody>
</table>

Wilson, Mihalcea, Boyd, and Pennebaker 2016
Annotation + Classification
**Annotation + Classification**

- **Method 1: Traditional Supervised Learning**
  - Annotate representative samples
  - Train a classifier
  - Apply to rest of data

- **Method 2: Transfer Learning**
  - Find another large, but similar dataset
  - Train a classifier on that dataset
  - *Optionally: fine-tune classifier to your smaller dataset*
  - Apply to rest of your data
After Classification

• Which features are most relevant for each class?
• What are common words/topics for each class?
• How do predicted classes relate to other variables?

• More about text classification coming up next week!
Wrap-up

- Content analysis background
- Word-level differences
- Dictionaries and Lexica
- Topic modeling
- Annotation + classification
Readings

• Manning: IR book section 13.5
• “Probabilistic Topic Models” by David Blei
• “Latent Dirichlet Allocation” by David Blei, Andrew Y. Ng, and Michael I. Jordan
• “Probabilistic Topic Models” by Mark Steyvers and Tom Griffiths

To watch:
• Guest lecture (2017) by David Blei at University of Edinburgh School of Informatics