Latent Dirichlet Allocation

*It Sounds Complicated!*

Cameron Carpenter
UMBC CMSC 676
April 28, 2015
Topic Modeling
Topic Modeling
Latent Semantic Indexing

Find a linear subspace of features that captures most of the variance

...But how do you evaluate it?
pLSI (aspect model)

*Every document is a mixture of topic proportions*
Key Idea

Given a generative model of text, one can fit the model to data using statistical methods.
This Isn’t a Grant Proposal

Let me try that again...
If you make a few assumptions about how the documents were created, you can use math to guess what the topics might be.
1. Make a list of all relevant topics
2. Until you have enough documents:
   a. Make a new empty document
   b. Pick a short list of topics for the document to be about*
   c. Until the document has enough words in it:
      i. Pick a topic from the short list
      ii. Pick a word from that topic**
      iii. Write that word down

This is LDA (really!)
## Topics

*probability distributions over words*

<table>
<thead>
<tr>
<th>Animals</th>
<th>Food</th>
<th>Music</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pig</td>
<td>Bread</td>
<td>Note</td>
</tr>
<tr>
<td>0.00053</td>
<td>0.00159</td>
<td>0.00243</td>
</tr>
<tr>
<td>Lemur</td>
<td>Sushi</td>
<td>Bass</td>
</tr>
<tr>
<td>0.00013</td>
<td>0.00029</td>
<td>0.00198</td>
</tr>
<tr>
<td>Bass</td>
<td>Table</td>
<td>Jam</td>
</tr>
<tr>
<td>0.00009</td>
<td>0.00143</td>
<td>0.00089</td>
</tr>
<tr>
<td>Cat</td>
<td>Jam</td>
<td>Violin</td>
</tr>
<tr>
<td>0.00102</td>
<td>0.00073</td>
<td>0.00252</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
An Example

- Document 1
  - 60% Food, 40% Music
  - note bread sushi table bass note jam jam table bread
- Document 2
  - 20% Animals, 80% Food
  - bread table table bread pig sushi cat bread sushi table
You don’t know...
○ ...how much of each topic a document contains
○ ...which topic a word in a document belongs to
○ ...what words get what weight in each topic

All of these attributes are hidden, or latent
The Posterior

Given that you only know the words in the documents, how can you figure out the rest?

Sampling Methods

Variational Methods

