Day 9: Topic Models

Kenneth Benoit

Essex Summer School 2014

July 31, 2014
LDA provides a generative model that describes how the documents in a dataset were created.

Each document is a collection of words, generated according to a multinomial distribution, one for each of \( K \) topics.

So the process is, roughly:

1. Choose a number of topics
2. Choose a distribution of topics, and create a document from this distribution
3. For each topic, generate words according to a distribution specific to that topic

The goal of inference in LDA is to discover the topics from the collection of documents, and to estimate the relationship of words to these topics.
Uses and applications

- Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents.
- Can be used to organize the collection according to the discovered themes.
- Topic modeling algorithms can be applied to massive collections of documents.
- Topic modeling algorithms can be adapted to many kinds of data. Among other applications, they have been used to find patterns in genetic data, images, and social networks.
Consider we have \( D \) (\( I \)) documents consisting of \( V \) (\( J \)) words each.

Assume we know that there are \( K \) topics in our corpus. Each topic describes a *multinomial* distribution over the \( V \) words, where \( \beta_k \) is the multinomial distribution over the words for topic \( k \).
Latent Dirichlet Allocation: Details

- For each document, the LDA generative process is:
  1. randomly choose a distribution over topics (a multinomial of length $K$)
  2. for each word in the document
     2.1 Probabilistically draw one of the $K$ topics from the distribution over topics obtained in (a), say topic $\beta_k$ (each document contains topics in different proportions)
     2.2 Probabilistically draw one of the $V$ words from $\beta_k$ (each individual word in the document is drawn from one of the $K$ topics in proportion to the document’s distribution over topics as determined in previous step)

- The goal of inference in LDA is to discover the topics from the collection of documents, and to estimate the relationship of words to these, *assuming this generative process*
More formal description of LDA

For each document:

1. draw a topic distribution, $\theta_d \sim \text{Dir}(\alpha)$, where $\text{Dir}(\cdot)$ is a draw from a uniform Dirichlet distribution with scaling parameter $\alpha$

2. for each word in the document:
   2.1 Draw a specific topic $z_{d,n} \sim \text{multi}(\theta_d)$ where $\text{multi}(\cdot)$ is a multinomial
   2.2 Draw a word $w_{d,n} \sim \beta_{z_{d,n}}$
Illustration of the LDA generative process

**Figure 2.** Illustration of the generative process and the problem of statistical inference underlying topic models

(from Steyvers and Griffiths 2007)
Topics example

<table>
<thead>
<tr>
<th>word</th>
<th>prob.</th>
<th>word</th>
<th>prob.</th>
<th>word</th>
<th>prob.</th>
<th>word</th>
<th>prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRUGS</td>
<td>.069</td>
<td>RED</td>
<td>.202</td>
<td>MIND</td>
<td>.081</td>
<td>DOCTOR</td>
<td>.074</td>
</tr>
<tr>
<td>DRUG</td>
<td>.060</td>
<td>BLUE</td>
<td>.099</td>
<td>THOUGHT</td>
<td>.066</td>
<td>DR.</td>
<td>.063</td>
</tr>
<tr>
<td>MEDICINE</td>
<td>.027</td>
<td>GREEN</td>
<td>.096</td>
<td>REMEMBER</td>
<td>.064</td>
<td>PATIENT</td>
<td>.061</td>
</tr>
<tr>
<td>EFFECTS</td>
<td>.026</td>
<td>YELLOW</td>
<td>.073</td>
<td>MEMORY</td>
<td>.037</td>
<td>HOSPITAL</td>
<td>.049</td>
</tr>
<tr>
<td>BODY</td>
<td>.023</td>
<td>WHITE</td>
<td>.048</td>
<td>THINKING</td>
<td>.030</td>
<td>CARE</td>
<td>.046</td>
</tr>
<tr>
<td>MEDICINES</td>
<td>.019</td>
<td>COLOR</td>
<td>.048</td>
<td>PROFESSOR</td>
<td>.028</td>
<td>MEDICAL</td>
<td>.042</td>
</tr>
<tr>
<td>PAIN</td>
<td>.016</td>
<td>BRIGHT</td>
<td>.030</td>
<td>FELT</td>
<td>.025</td>
<td>NURSE</td>
<td>.031</td>
</tr>
<tr>
<td>PERSON</td>
<td>.016</td>
<td>COLORS</td>
<td>.029</td>
<td>REMEMBERED</td>
<td>.022</td>
<td>PATIENTS</td>
<td>.029</td>
</tr>
<tr>
<td>MARIJUANA</td>
<td>.014</td>
<td>ORANGE</td>
<td>.027</td>
<td>THOUGHTS</td>
<td>.020</td>
<td>DOCTORS</td>
<td>.028</td>
</tr>
<tr>
<td>LABEL</td>
<td>.012</td>
<td>BROWN</td>
<td>.027</td>
<td>FORGOTTEN</td>
<td>.020</td>
<td>HEALTH</td>
<td>.025</td>
</tr>
<tr>
<td>ALCOHOL</td>
<td>.012</td>
<td>PINK</td>
<td>.017</td>
<td>MOMENT</td>
<td>.020</td>
<td>MEDICINE</td>
<td>.017</td>
</tr>
<tr>
<td>DANGEROUS</td>
<td>.011</td>
<td>LOOK</td>
<td>.017</td>
<td>THINK</td>
<td>.019</td>
<td>NURSING</td>
<td>.017</td>
</tr>
<tr>
<td>ABUSE</td>
<td>.009</td>
<td>BLACK</td>
<td>.016</td>
<td>THING</td>
<td>.016</td>
<td>DENTAL</td>
<td>.015</td>
</tr>
<tr>
<td>EFFECT</td>
<td>.009</td>
<td>PURPLE</td>
<td>.015</td>
<td>WONDER</td>
<td>.014</td>
<td>NURSES</td>
<td>.013</td>
</tr>
<tr>
<td>KNOWN</td>
<td>.008</td>
<td>CROSS</td>
<td>.011</td>
<td>FORGET</td>
<td>.012</td>
<td>PHYSICIAN</td>
<td>.012</td>
</tr>
<tr>
<td>PILLS</td>
<td>.008</td>
<td>COLORED</td>
<td>.009</td>
<td>RECALL</td>
<td>.012</td>
<td>HOSPITALS</td>
<td>.011</td>
</tr>
</tbody>
</table>

**Figure 1.** An illustration of four (out of 300) topics extracted from the TASA corpus.

(from Steyvers and Griffiths 2007)

Often $K$ is quite large!
Probabilistic Topic Models

- Consider \( P(z) \) as the distribution over topics \( z \) in a particular document and \( P(w|z) \) as the probability distribution over words \( w \) given topic \( z \).
- Each word \( w_i \) in a document (where the index refers to the \( i \)th word token) is generated by first sampling a topic from the topic distribution, then choosing a word from the topic-word distribution.
- \( P(z_i = j) \) is the probability that the \( k \)th topic was sampled for the \( i \)th word token.
- \( P(w_i|z_i = k) \) is the probability of word \( w_i \) under topic \( k \).
- \( K \) is the number of topics.
- Then

\[
P(w_i) = \sum_{k=1}^{K} P(w_i|z_i = k) P(z_i = k)
\]
Elements of the model

- Let $\phi^{(k)} = P(w|z = k)$ refer to the multinomial distribution over words for topic $k$
- $\theta^{(d)} = P(z)$ refers to the multinomial distribution over topics for document $d$
- Text collection consists of $D$ documents and each document $d$ consists of $N_d$ word tokens, with $N = \sum N_d$ (the total tokens in the document collection)
- $\phi$ which words are important for which topic
- $\theta$ which topics are important for a particular document
- $z_i$ is the assignment of word tokens to topics
Graphical model for LDA using plate notation

Interpreting a plate model:
- shaded variables are observed, unshaded are latent
- the hyperparameters $\alpha$ and $\beta$ are treated as constants in the model (these are parameters of the Dirichlet distribution)
- arrows indicate conditional dependencies between variables
- plates (boxes) are repetitions of sampling steps, with the lower right corner variable indicating the number of samples
The Dirichlet is the conjugate prior distribution for the multinomial, and is used in the Bayesian inference required to estimate these parameters.

Estimation is performed using (collapsed) Gibbs sampling and/or variational Expectation-Maximization.

(for us) Implemented in the lda library and can be used with quanteda dfm objects.
Challenges in applying LDA

- How many topics ($K$)?
- How to interpret (label) the topics?
- Should we expect all topics to make sense?
- Some models are complicated and expensive to estimate