Objectives

- Motivate topic analysis and topic models
- Build the generative story for one such model
- Introduce Latent Dirichlet Allocation (LDA)
- Preview some results!

Motivations

- Find / Discover Topics
- Find a short description of large collections of data
- Find topical trends

What’s a Topic?

- Extracts in topical guides in scripture
- Subject
- Themes, underlying meaning
- Class, genre
- Short summary

Finding Topics

- Unsupervised learning of topics can:
  - Provide “gist” of documents:
    * articles/chapters
    * conversations
    * emails
    * ... any text
  - Reveal semantic associations
- Topics are useful latent structures
A Simple Directed Graphical Model

Advantage to working in the probabilistic modeling framework: we know how to ask and answer queries!

- Inferring latent structure:
  \[ P(\ell | \mathbf{w}) = \frac{P(w|\ell)P(\ell)}{P(w)} \]

- Distribution over words:
  \[ P(w) = \sum \limits_{\ell} P(w, \ell) \]

- Prediction:
  \[ P(w_{k+1} | \ell) = \ldots \]

Generative Process

“Bag of Words” Assumption

- Let’s assume (again) that all the words within a document are exchangeable.

Exchangeability

- A finite set of random variables \( \{x_1, x_2, \ldots, x_N\} \)
  is said to be exchangeable iff the joint distribution is invariant to permutation. Let \( \pi \) be a permutation of the integers from 1 to \( N \). Then:
  \[ p(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(N)}) = p(x_1, x_2, \ldots, x_N) \]

- Equivalent to: conditionally independent and identically distributed (“c.i.i.d.”)

LDA: a Probabilistic Generative Topic Model

- Basic ideas behind LDA:
  - Each document is a probability distribution over topics
  - Each topic is a probability distribution over words

- What’s the primary distinction between this and the Mixture of Multinomials model for clustering?

Desiderata: Prior Distributions

- \textit{desiderata} = Latin for “what we want”

In the LDA model:

- For each document, we need topic mixture proportions (weights)
  - \( T \)-dimensional vectors of non-negative numbers that sum to one.
    - i.e., The mixture proportions constitute a parameter vector for a multinomial or categorical distribution.

- To draw (sample) such a vector, we need a distribution over such vectors.

- The space of all of such vectors has a nice geometric interpretation:
  - \( (T-1) \)-simplex in \( T \) dimensions

- Criteria for selecting our prior:
  - It needs to be defined over a \((T-1)\)-simplex.
  - It should be conjugate to the multinomial and categorical distributions.

- The Dirichlet Distribution: 

\[ \pi \sim \text{Dirichlet}(\alpha) \]
Dirichlet Distribution

- Hyper-parameters $\phi$ determine the form of the distribution:

![Diagram of Dirichlet Distribution]

Generative Process

- For each document $(1 \leq t \leq N)$ do:
  - Draw $\phi_t \sim \text{Dirichlet}(\phi)$ // create a spinner for document $t$
  - Draw $\theta_t \sim \text{Dirichlet}(\theta)$ // create a spinner for topic $t$

LDA Generative Process

- For each topic $t$ $(1 \leq t \leq T)$ do:
  - Draw $M_t \sim \text{Poisson}(\lambda)$ // create a spinner for topic $t$
  - Draw $\alpha_t \sim \text{Dirichlet}(\alpha)$ // create a spinner for document $t$

Graphical Model with Generative Story

For each topic $t$:
- Sample a distribution over words.
- Sample a distribution over topics.
- Choose a document length $M_t$ (tokens).
- For each position $j$ in document $t$:
  - Sample a topic.
  - Sample a word token from that topic.

The LDA Model (without plates)

- From this viewpoint it’s easier to ask influence questions; e.g., Does $\theta_2$ influence $\phi_1$, given $w_j$?

Posterior Inference

- We need to estimate the assignments of topics to words also.
A Play is written to be performed on a stage before a live audience or before motion picture cameras (for later viewing by large audiences). A Play is written because playwrights have something ...

He was listening to music coming from a passing riverboat. The music had already captured his heart as well as his ear. It was jazz. Bix beiderbecke had already had music lessons. He wanted to play the cornet. And he wanted to play jazz ...

∀ , ∀ (1 ≤ j ≤ M): zi,j

DOCUMENT 1: A Play is written to be performed on a stage before a live audience or before motion picture cameras (for later viewing by large audiences). A Play is written because playwrights have something ...

DOCUMENT 2: He was listening to music coming from a passing riverboat. The music had already captured his heart as well as his ear. It was jazz. Bix beiderbecke had already had music lessons. He wanted to play the cornet. And he wanted to play jazz ...

Reward: Some Results

- **Input:**
  - 10,000 text articles from the 20 Newsgroups data set
  - Number of topics: T = 50

- **Output:**
  - the most probable words for a few topics, according to LDA

- **Algorithms:**
  - Exact posterior inference is intractable, so we use approximate inference instead
  - 40 iterations of Variational EM
  - Could use Gibbs sampling

A few topics from 20-Newsgroups

<table>
<thead>
<tr>
<th>political</th>
<th>team</th>
<th>space</th>
<th>drive</th>
<th>god</th>
</tr>
</thead>
<tbody>
<tr>
<td>party</td>
<td>game</td>
<td>nasa</td>
<td>windows</td>
<td>jesus</td>
</tr>
<tr>
<td>business</td>
<td>play</td>
<td>research</td>
<td>card</td>
<td>his</td>
</tr>
<tr>
<td>convention</td>
<td>year</td>
<td>center</td>
<td>dos</td>
<td>bible</td>
</tr>
<tr>
<td>institute</td>
<td>games</td>
<td>Earth</td>
<td>soci</td>
<td>christian</td>
</tr>
<tr>
<td>committee</td>
<td>win</td>
<td>health</td>
<td>dbak</td>
<td>christ</td>
</tr>
<tr>
<td>states</td>
<td>hockey</td>
<td>medical</td>
<td>system</td>
<td>hem</td>
</tr>
<tr>
<td>rights</td>
<td>season</td>
<td>gov</td>
<td>memory</td>
<td>christians</td>
</tr>
</tbody>
</table>

Credit: Jonathan Huang of CMU

Problems vs. Solutions: Important Distinctions

- **MOTELS**
  - Mixture of multinomials
  - LDA

- **Algorithms**
  - Marginalization - Posterior inference
  - MCMC, EM, Gibbs
  - EM

- **Textual Projection**

Next

- Inference!