Lecture 11: Topic Models

Mirella Lapata

School of Informatics
University of Edinburgh
mlap@inf.ed.ac.uk

February 27, 2015
Slides based on Blei 2011
1 Vector Space Models

2 Topic Models
   - Latent Dirichlet Allocation
   - Inference with Gibbs Sampling
   - Representing Meaning with LDA

3 Discussion

Reading: Griffiths et al. (2007).
Motivation

We’d like to represent the meaning of words and documents, to help us do things like:

- Retrieve documents relevant to a query.
- Disambiguate word senses (for IR, MT, etc.).
- Explain human learning and processing of words (word associations, speed of acquisition, etc).
- Analyze large document collections.
Vector-based Models

Represent the meaning of words in high-dimensional space:

Multiple word meanings are conflated, no probability model.
We want to find themes (or topics) in documents which is useful for e.g., search or browsing.

We don’t want to do supervised topic classification (neither fix topics in advance nor do manual annotation).

Approach must automatically tease out the topics.

Essentially a clustering problem: both words and documents are being clustered.
Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andorsson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completed and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

“Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.”


**SCIENCE • VOL. 272 • 24 MAY 1996**

Simple intuition: documents exhibit multiple topics.
Each **topic** is a distribution over words

Each **document** is a mixture of corpus-wide topics
In reality, we only observe the documents

The other structure are **hidden variables**
Our goal is to infer the hidden variables

i.e., compute their distribution conditioned on the documents

\[ p(\text{topics, proportions, assignments} | \text{documents}) \]
LDA: Key Assumptions

- Documents exhibit multiple topics (but typically not many)
- LDA is a probabilistic model with a corresponding generative process (each document is generated by this process)
- A topic is a distribution over a fixed vocabulary (topics are assumed to be generated first, before the documents)
- Only the number of topics is specified in advance
Generative Process

For \( j = 1 \ldots T \) topics,

Choose \( \phi^{(j)} \sim \text{Dirichlet}(\beta) \).

For \( d = 1 \ldots D \) documents,

Choose \( \theta^{(d)} \sim \text{Dirichlet}(\alpha) \).

For \( i = 1 \ldots N_d \) words in doc \( d \),

Choose \( z_i \sim \text{Multinomial}(\theta^{(d)}) \).

Choose \( w_i \sim \text{Multinomial}(\phi^{(z_i)}) \).

\( \phi^{(j)}_1 \ldots \phi^{(j)}_V : \text{prob. of each wd. in topic } j \)

\( \theta^{(d)}_1 \ldots \theta^{(d)}_T : \text{prob. of each topic in doc. } d \)

\( z_i : \text{topic of word } i \)

\( w_i : \text{identity of word } i \)

Note that, except in the appendices, Griffiths et al. (2007) use \( g \) (for gist) to refer to \( \theta \). They also use the term Discrete rather than Multinomial.
The Graphical Model

- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed; plates \( \approx \) replicated variables.
The Graphical Model

Nodes are random variables; edges indicate dependence.
Shaded nodes are observed; plates $\approx$ replicated variables.
The Graphical Model

\[ P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j) \]
Aside: Multinomial Distribution

For $x_i \in \{0, \ldots, n\}$

$$P(x|\theta) = \frac{n!}{d \prod_{i=1}^{d} x_i!} \prod_{i=1}^{d} \theta_i^{x_i}, \quad n = \sum_{i=1}^{d} x_i, \quad \sum_{i=1}^{d} \theta_i = 1, \quad \theta_i > 0$$

When $n = 1$ the multinomial distribution simplifies to:

$$P(x|\theta) = \prod_{i=1}^{d} \theta_i^{x_i}, \quad \sum_{i=1}^{d} \theta_i = 1, \quad \theta_i > 0$$

- unigram language model, \textit{1-of-V coding} ($d = V$ vocab size)
- $x_i$ indicates word $i$ of the vocabulary observed ($x_i = 1$ if word $i$ is observed and 0 otherwise)
- $\theta_i = P(w_i)$ the probability that word $i$ is seen
The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one. It is **conjugate to the multinomial**. Given a multinomial observation, the posterior distribution of $\theta$ is a Dirichlet. The parameter $\alpha$ smooths the topic distribution in the document. The parameter $\beta$ smooths the word distribution in every topic.
\( \alpha = 1 \)
\[ \alpha = 10 \]

![Graph showing item-value distribution for different categories with \( \alpha = 10 \).]
$\alpha = 100$
$\alpha = 1$
\[ \alpha = 0.1 \]
$\alpha = 0.01$
$\alpha = 0.001$
Inference with Gibbs Sampling

- An iterative process.
- Start with random topic assignments for each word.
- In each iteration, *for each word* in the data:
  - Assume you know (from the prev. iteration) the topics of all other words. (pretend they are correct)
  - Determine the probabilities of each topic-assignment given the rest of the data.
  - Choose the most probable assignment.
- Iterate until convergence.
Gibbs Sampling

- Collection of documents is a set of word indices $w_i$ and document indices $d_i$, for each word token $i$.
- Consider each token and estimate:
  $$P(z_i = j | z_{-i}, w_i, d_i, \cdot) = \frac{C_{wij} + \beta}{\sum_{w=1}^{W} C_{wj}^{WT} + W\beta} \frac{C_{dij} + \alpha}{\sum_{t=1}^{T} C_{d_{-i}t}^{DT} + T\alpha}$$
  
- From this conditional distribution, a topic is sampled and stored as the new topic assignment for this word token.

$z_i = j$: topic assignment of token $i$ to topic $j$
$z_{-i}$: the topic assignments of all other word tokens
$\cdot$: all other known information ($w_{-i}$, $d_{-i}$, $\alpha$, $\beta$).
$C^{WT}$: matrix of counts with dimensions $W \times T$
$C^{DT}$: matrix of counts with dimensions $D \times T$
Posterior Estimates of $\beta$ and $\theta$

$$\beta_{ij} = \frac{C_{ij}^{WT} + \beta}{C_{wij}^{WT} + \beta} \quad \theta_{dj} = \frac{C_{dij}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_{it}}^{DT} + T\alpha}$$

Using the count matrices as before, where $\beta_{ij}$ is the probability of word type $i$ for topic $j$, and $\theta_{dj}$ is the proportion of topic $j$ in document $d$. 
Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions “are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an


Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Mirella Lapata

Lecture 11: Topic Models
## Example Inference

<table>
<thead>
<tr>
<th>human</th>
<th>evolution</th>
<th>disease</th>
<th>computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>genome</td>
<td>evolutionary</td>
<td>host</td>
<td>models</td>
</tr>
<tr>
<td>dna</td>
<td>species</td>
<td>bacteria</td>
<td>information</td>
</tr>
<tr>
<td>genetic</td>
<td>organisms</td>
<td>diseases</td>
<td>data</td>
</tr>
<tr>
<td>genes</td>
<td>life</td>
<td>resistance</td>
<td>computers</td>
</tr>
<tr>
<td>sequence</td>
<td>origin</td>
<td>bacterial</td>
<td>system</td>
</tr>
<tr>
<td>gene</td>
<td>biology</td>
<td>new</td>
<td>network</td>
</tr>
<tr>
<td>molecular</td>
<td>groups</td>
<td>strains</td>
<td>model</td>
</tr>
<tr>
<td>sequencing</td>
<td>phylogenetic</td>
<td>control</td>
<td>parallel</td>
</tr>
<tr>
<td>map</td>
<td>living</td>
<td>infectious</td>
<td>methods</td>
</tr>
<tr>
<td>information</td>
<td>diversity</td>
<td>malaria</td>
<td>networks</td>
</tr>
<tr>
<td>genetics</td>
<td>group</td>
<td>parasite</td>
<td>software</td>
</tr>
<tr>
<td>mapping</td>
<td>new</td>
<td>parasites</td>
<td>new</td>
</tr>
<tr>
<td>project</td>
<td>two</td>
<td>united</td>
<td></td>
</tr>
</tbody>
</table>

*Note: The table lists examples of terms that may be relevant in the context of topic modeling and inference.*
Example Inference

Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (1) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural populations can show persistent oscillatory dynamics and chaos, the latter characterized by extreme sensitivity to initial conditions. If such chaotic dynamics were common in nature, then this would have important ramifications for the management and conservation of natural resources. On page 389 of this issue, Costantino et al. (2) provide the most convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, Tribolium castaneum (see figure).

It has proven extremely difficult to demonstrate complex dynamics in populations in the field. By its very nature, a chaotically fluctuating population will superficially resemble a stable or cyclic population buffered by the normal random perturbations experienced by all species. Given a long enough time series, diagnostic tools from nonlinear mathematics can be used to identify the telltale signatures of chaos. In phase space, chaotic trajectories come to lie on "strange attractors," curious geometric objects with fractal structure and hence noninteger dimension. As they move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Lyapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Lyapunov exponents from time series data, and some candidate chaotic population have been identified (some insects, rodents, and most convincingly, human childhood diseases), but the statistical difficulties preclude any broad generalization (3).

An alternative approach is to parameterize population models with data from natural populations and then compare their predictions with the dynamics in the field. This technique has been gaining popularity in recent years, helped by statistical advances in parameter estimation. Good ex-

The authors are in the Department of Biology, Imperial College at Silwood Park, Ascot, Berks, SL5 7PZ, UK. E-mail: m.hassell@ic.ac.uk

Cannibalism and chaos. The flour beetle, Tribolium castaneum, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.
## Example Inference

<table>
<thead>
<tr>
<th>problem</th>
<th>model</th>
<th>selection</th>
<th>species</th>
</tr>
</thead>
<tbody>
<tr>
<td>problems</td>
<td>rate</td>
<td>male</td>
<td>forest</td>
</tr>
<tr>
<td>mathematical</td>
<td>constant</td>
<td>males</td>
<td>ecology</td>
</tr>
<tr>
<td>number</td>
<td>distribution</td>
<td>females</td>
<td>fish</td>
</tr>
<tr>
<td>new</td>
<td>time</td>
<td>sex</td>
<td>ecological</td>
</tr>
<tr>
<td>mathematics</td>
<td>number</td>
<td>species</td>
<td>conservation</td>
</tr>
<tr>
<td>university</td>
<td>size</td>
<td>female</td>
<td>diversity</td>
</tr>
<tr>
<td>two</td>
<td>values</td>
<td>evolution</td>
<td>population</td>
</tr>
<tr>
<td>first</td>
<td>value</td>
<td>populations</td>
<td>natural</td>
</tr>
<tr>
<td>numbers</td>
<td>average</td>
<td>population</td>
<td>ecosystems</td>
</tr>
<tr>
<td>work</td>
<td>rates</td>
<td>sexual</td>
<td>populations</td>
</tr>
<tr>
<td>time</td>
<td>data</td>
<td>behavior</td>
<td>endangered</td>
</tr>
<tr>
<td>mathematicians</td>
<td>density</td>
<td>evolutionary</td>
<td>tropical</td>
</tr>
<tr>
<td>chaos</td>
<td>measured</td>
<td>genetic</td>
<td>forests</td>
</tr>
<tr>
<td>chaotic</td>
<td>models</td>
<td>reproductive</td>
<td>ecosystem</td>
</tr>
</tbody>
</table>
What does this have to do with semantics?

<table>
<thead>
<tr>
<th></th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic n</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>0.39</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>difficulty</td>
<td>0.03</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>produce</td>
<td>0.06</td>
<td>0.17</td>
<td></td>
</tr>
</tbody>
</table>

Topics are the dimensions of the space (500, 1000)

Vector components: probability of word given topic

Topics correspond to coarse-grained sense distinctions

Cosine similarity can be used (probabilistic alternatives)
Griffiths et al. present a range of evaluations and modeling tasks comparing LDA and LSA.

- Synonym tests from TOEFL, free association, ambiguity, semantic priming, reading time, free recall.
Griffiths et al. present a range of evaluations and modeling tasks comparing LDA and LSA.

Synonym tests from TOEFL, free association, ambiguity, semantic priming, reading time, free recall

Free-association: produce the first word that comes to mind in response to a cue word.
Griffiths et al. present a range of evaluations and modeling tasks comparing LDA and LSA.

- Synonym tests from TOEFL, **free association**, ambiguity, semantic priming, reading time, free recall
- **Free-association**: produce the first word that comes to mind in response to a cue word.

```
SPOTS
```

```
rice
```

```
separate
```

```
dog
```

```
chinese
```

```
divide
```

```
dirty
```

```
wedding
```

```
divorce
```

```
dirt
```

```
food
```

```
part
```

```
stripes
```

```
white
```

```
split
```

```
dark
```

```china
```

```
remove
```
Griffiths et al. present a range of evaluations and modeling tasks comparing LDA and LSA.

Synonym tests from TOEFL, **free association**, ambiguity, semantic priming, reading time, free recall

**Free-association**: produce the first word that comes to mind in response to a cue word.

<table>
<thead>
<tr>
<th>SPOTS</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DOG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIRTY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIRT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STRIPES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DARK</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Griffiths et al. present a range of evaluations and modeling tasks comparing LDA and LSA.

Synonym tests from TOEFL, **free association**, ambiguity, semantic priming, reading time, free recall

**Free-association**: produce the first word that comes to mind in response to a cue word.

<table>
<thead>
<tr>
<th>SPOTS</th>
<th>RICE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DOG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIRTY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIRT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STRIPES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DARK</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Griffiths et al. present a range of evaluations and modeling tasks comparing LDA and LSA.

Synonym tests from TOEFL, **free association**, ambiguity, semantic priming, reading time, free recall

**Free-association**: produce the first word that comes to mind in response to a cue word.

<table>
<thead>
<tr>
<th>SPOTS</th>
<th>RICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOG</td>
<td>CHINESE</td>
</tr>
<tr>
<td>DIRTY</td>
<td>WEDDING</td>
</tr>
<tr>
<td>DIRT</td>
<td>FOOD</td>
</tr>
<tr>
<td>STRIPES</td>
<td>WHITE</td>
</tr>
<tr>
<td>DARK</td>
<td>CHINA</td>
</tr>
</tbody>
</table>
Griffiths et al. present a range of evaluations and modeling tasks comparing LDA and LSA.

Synonym tests from TOEFL, **free association**, ambiguity, semantic priming, reading time, free recall

**Free-association**: produce the first word that comes to mind in response to a cue word.

<table>
<thead>
<tr>
<th>SPOTS</th>
<th>RICE</th>
<th>SEPARATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOG</td>
<td>CHINESE</td>
<td></td>
</tr>
<tr>
<td>DIRTY</td>
<td>WEDDING</td>
<td></td>
</tr>
<tr>
<td>DIRT</td>
<td>FOOD</td>
<td></td>
</tr>
<tr>
<td>STRIPES</td>
<td>WHITE</td>
<td></td>
</tr>
<tr>
<td>DARK</td>
<td>CHINA</td>
<td></td>
</tr>
</tbody>
</table>
Griffiths et al. present a range of evaluations and modeling tasks comparing LDA and LSA.

Synonym tests from TOEFL, free association, ambiguity, semantic priming, reading time, free recall

Free-association: produce the first word that comes to mind in response to a cue word.

<table>
<thead>
<tr>
<th>SPOTS</th>
<th>RICE</th>
<th>SEPARATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOG</td>
<td>CHINESE</td>
<td>DIVIDE</td>
</tr>
<tr>
<td>DIRTY</td>
<td>WEDDING</td>
<td>DIVORCE</td>
</tr>
<tr>
<td>DIRT</td>
<td>FOOD</td>
<td>PART</td>
</tr>
<tr>
<td>STRIPES</td>
<td>WHITE</td>
<td>SPLIT</td>
</tr>
<tr>
<td>DARK</td>
<td>CHINA</td>
<td>REMOVE</td>
</tr>
</tbody>
</table>
Computing Word Similarity

- Build LSA and LDA representations from corpus of educational materials.
- Compute associates predicted by each model.
- LSA: top associate of $w_1$ is word with closest cosine (or dot product) similarity.
- LDA: top associate of $w_1$ is word with highest $P(w_2|w_1)$.

\[
P(w_2|w_1) = \sum_z P(w_2|z)P(z|w_1)
\]
If we take the top $n$ associates returned by the model (set size), what is the probability that this set contains the true first associate?
LSA is useful in practice: IR, unsupervised morphology, essay scoring, text coherence, language modeling, summarization.

LDA is useful for publishing machine learning papers: Hierarchical LDA, nested LDA, inference for LDA, author-topic model, topics over time, topics + syntax, etc.

So far, limited work showing these models useful in NLP.

LDA for discourse segmentation, extensions for text classification, WSD, tracking social phenomena.

In comparisons, LSA often performs as well or better.
Discussion

Mirella Lapata  Lecture 11: Topic Models  36