Topic Models

Thomas Bernau 2015
INFORMATION
Example - Wikipedia

Number of articles on en.wikipedia.org

Example - StackOverflow

Cumulative number of questions for tags c# and java over time

http://meta.stackoverflow.com/questions/260570/how-to-compare-tags-by-question-count-over-time
Information

• How to ...
  – ... organize information?
  – ... access information?
  – ... search information?
  – ... understand information?
TOPIC MODELING
Idea

- Uncover hidden (latent) topics in a collection
- Annotate documents according to topics
- Use annotation to ...
  - ... organize
  - ... access
  - ... search
  - ... understand
Recognize this?

<table>
<thead>
<tr>
<th>Tutorials</th>
<th>Web</th>
<th>Bug</th>
<th>Stackoverflow</th>
<th>Commits</th>
<th>Security</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tutorial</td>
<td>Development</td>
<td>Bugs</td>
<td>Prompter</td>
<td>Full</td>
<td>Discussions</td>
</tr>
<tr>
<td>Images</td>
<td>Questions</td>
<td>Assignees</td>
<td>Discussions</td>
<td>Time</td>
<td>Developers</td>
</tr>
<tr>
<td>Links</td>
<td>Related</td>
<td>Reports</td>
<td>Results</td>
<td>Code</td>
<td>Authors</td>
</tr>
<tr>
<td>Code</td>
<td>Use</td>
<td>Results</td>
<td>Code</td>
<td>Context</td>
<td>Effort</td>
</tr>
<tr>
<td>Official</td>
<td>Browser</td>
<td>Tracker</td>
<td>Context</td>
<td>IDE</td>
<td>Github</td>
</tr>
<tr>
<td>Source</td>
<td>CSS</td>
<td>Analysis</td>
<td>IDE</td>
<td>Open</td>
<td>Openstack</td>
</tr>
<tr>
<td>Structure</td>
<td>Html5</td>
<td>Artifacts</td>
<td>Openstack</td>
<td>Comments</td>
<td>Data</td>
</tr>
<tr>
<td>Community</td>
<td>Javascript</td>
<td>Classification</td>
<td>Results</td>
<td>Data</td>
<td>Pull</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Topics</td>
<td>Considered</td>
<td>API</td>
<td>Requests</td>
<td>Sentiments</td>
</tr>
<tr>
<td>Number</td>
<td>Developers</td>
<td>Context</td>
<td>Basic</td>
<td>Source</td>
<td>Analysis</td>
</tr>
<tr>
<td>Topics</td>
<td>Discussions</td>
<td>Data</td>
<td>Considered</td>
<td>Survey</td>
<td>Discussion</td>
</tr>
<tr>
<td>Contain</td>
<td>LDA</td>
<td>Differences</td>
<td>Data</td>
<td>Word</td>
<td>Entities</td>
</tr>
<tr>
<td>Content</td>
<td>Mobile</td>
<td>Discussions</td>
<td>Developers</td>
<td>1400</td>
<td>Github</td>
</tr>
<tr>
<td>Development</td>
<td>Accepted</td>
<td>Extracted</td>
<td>Entropy</td>
<td>Github</td>
<td></td>
</tr>
<tr>
<td><strong>Recognize this?</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>apresnyakova</td>
<td>erwinschens</td>
<td>hartenfels</td>
<td>lukashaertel</td>
<td>mbrack</td>
<td>olgazagovora</td>
</tr>
<tr>
<td>Tutorials</td>
<td>Web</td>
<td>Bug</td>
<td>Stackoverflow</td>
<td>Commits</td>
<td>Security</td>
</tr>
<tr>
<td>Tutorial</td>
<td>Development</td>
<td>Bugs</td>
<td>Prompter</td>
<td>Full</td>
<td>Discussions</td>
</tr>
<tr>
<td>Images</td>
<td>Questions</td>
<td>Assignees</td>
<td>Discussions</td>
<td>Time</td>
<td>Related</td>
</tr>
<tr>
<td>Links</td>
<td>Related</td>
<td>Reports</td>
<td>Results</td>
<td>Developers</td>
<td>Comments</td>
</tr>
<tr>
<td>Code</td>
<td>Use</td>
<td>Results</td>
<td>Code</td>
<td>Authors</td>
<td>Data</td>
</tr>
<tr>
<td>Official</td>
<td>Browser</td>
<td>Tracker</td>
<td>Context</td>
<td>Effort</td>
<td>Results</td>
</tr>
<tr>
<td>Source</td>
<td>CSS</td>
<td>Analysis</td>
<td>IDE</td>
<td>Github</td>
<td>Open</td>
</tr>
<tr>
<td>Structure</td>
<td>Html5</td>
<td>Artifacts</td>
<td>Programmers</td>
<td>Openstack</td>
<td>Pull</td>
</tr>
<tr>
<td>Community</td>
<td>Javascript</td>
<td>Classification</td>
<td>Relevant</td>
<td>Results</td>
<td>Requests</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Topics</td>
<td>Considered</td>
<td>API</td>
<td>Source</td>
<td>Analysis</td>
</tr>
<tr>
<td>Number</td>
<td>Developers</td>
<td>Context</td>
<td>Basic</td>
<td>Survey</td>
<td>Discussion</td>
</tr>
<tr>
<td>Topics</td>
<td>Discussions</td>
<td>Data</td>
<td>Considered</td>
<td>Word</td>
<td>Entities</td>
</tr>
<tr>
<td>Contain</td>
<td>LDA</td>
<td>Differences</td>
<td>Data</td>
<td>1400</td>
<td>Github</td>
</tr>
<tr>
<td>Content</td>
<td>Mobile</td>
<td>Discussions</td>
<td>Developers</td>
<td>2010</td>
<td></td>
</tr>
<tr>
<td>Development</td>
<td>Accepted</td>
<td>Extracted</td>
<td>Entropy</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Topic Modeling

„Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections”

(David Blei, 2015)

https://www.cs.princeton.edu/~blei/topicmodeling.html
Topic Models

• From linear algebra
  – Latent Semantic Indexing
Topic Models

LSI

\[ \text{documents} \times \text{words} = \text{words} \times \text{dim} \]

\[ \Sigma \times \text{dim} \times \text{documents} \]

\[ M \]

\[ U \]

\[ V^* \]
Topic Models

• From linear algebra
  – Latent Semantic Indexing

• Probabilistic methods
  – Probabilistic Latent Semantic Indexing
  – Latent Dirichlet Allocation
Topic Models

\[ M = U \times \Sigma \times V^* \]

**LSI**
- \( M \): \( \text{documents} \times \text{words} \)
- \( U \): \( \text{words} \times \text{dim} \)
- \( \Sigma \): \( \text{dim} \times \text{dim} \)
- \( V^* \): \( \text{dim} \times \text{documents} \)

**LDA**
- \( M \): \( \text{documents} \times \text{words} \)
- \( U \): \( \text{words} \times \text{topics} \)
- \( V^* \): \( \text{topics} \times \text{documents} \)
Generative hierarchical probabilistic model

LATENT DIRICHLET ALLOCATION
Basic Assumption

Data is a set of observations from a generative probabilistic process with hidden variables reflecting the thematic structure
Intuition of LDA

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


http://dl.acm.org/citation.cfm?id=2133826
Intuition of LDA

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.


http://dl.acm.org/citation.cfm?id=2133826
Intuition of LDA
Intuition of LDA

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 analysis to compare known genomes, concluded that today’s organisms can be sustained with just 235 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 73,000 genes in the human genome, notes Stephen Schwartz of the University of North Carolina. “But coming up with a consensus answer may be more than just a matter of numbers, since particularly as more and more genomes are completely mapped and sequenced. It may be a way of organizing any newly sequenced genome,” explains Arathy Madhavan, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing the

http://dl.acm.org/citation.cfm?id=2133826
Plate notation

- Nodes are random variables
- Shading means observation
- Edges are dependencies
- Plates represent replicated structure

Latent Dirichlet Allocation

\[
\alpha \rightarrow \theta_d \rightarrow Z_{d,n} \rightarrow W_{d,n} \rightarrow N \rightarrow D \rightarrow \beta_k \rightarrow K \rightarrow \eta
\]

http://dl.acm.org/citation.cfm?id=2133826
Latent Dirichlet Allocation

\[ \alpha \rightarrow \theta_d \rightarrow Z_{d,n} \rightarrow W_{d,n} \rightarrow \beta_k \rightarrow \eta \]

Topics

http://dl.acm.org/citation.cfm?id=2133826
Latent Dirichlet Allocation

α \rightarrow \theta_d \rightarrow Z_{d,n} \rightarrow W_{d,n} \rightarrow N \rightarrow D \rightarrow \beta_k \rightarrow \eta

Per-document topic distributions

Topics

http://dl.acm.org/citation.cfm?id=2133826
Latent Dirichlet Allocation

Topics
Per-word topic assignment
Per-document topic distributions

http://dl.acm.org/citation.cfm?id=2133826
Latent Dirichlet Allocation

http://dl.acm.org/citation.cfm?id=2133826
Latent Dirichlet Allocation

\[ \alpha \xrightarrow{\theta_d} Z_{d,n} \xrightarrow{W_{d,n}} N \xrightarrow{D} \beta_k \xrightarrow{\eta} \]

Topics

Observed word

Per-word topic assignment

Per-document topic distributions

Dirichlet parameter

Topic hyperparameter

http://dl.acm.org/citation.cfm?id=2133826
The Dirichlet Distribution

• Family of continuous multivariate probability distributions

\[ p(\theta | \hat{\alpha}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1} \]

• our \( \theta \) is a K-dimensional Dirichlet
• our \( \beta \) is a V-dimensional Dirichlet
The Dirichlet Distribution

The Dirichlet Distribution

Dirichlet in R

http://cran.r-project.org/web/packages/MCMCpack/MCMCpack.pdf
Latent Dirichlet Allocation

\[ \alpha \rightarrow \theta_d \rightarrow Z_{d,n} \rightarrow W_{d,n} \rightarrow \beta_k \rightarrow \eta \]

\[ N_D \]

Topics

Observed word

Per-word topic assignment

Per-document topic distributions

Dirichlet parameter

Topic hyperparameter

http://dl.acm.org/citation.cfm?id=2133826
Collapsed Gibbs sampling (idea)

1. Start from random topic assignments
2. Change topic assignment $Z_{d,n}$ of a word $w_{d,n}$
3. Repeat 2. for all topic assignments
4. Iterate over 3. until "stable" assignments

(a Markov-Chain-Monte-Carlo method)
Collapsed Gibbs sampling (idea)

• What we want:

\[ p(z_{d,n} | z_{-d,n}, w_{d,n}, \alpha, \eta, \beta_{1:K}, \theta_d) \]

• Collapsed Gibbs sampling:
  – Integrate out \( \beta \) and \( \Theta \) (Rao-Blackwellize)

\[ p(z_{d,n} | z_{-d,n}, w_{d,n}, \alpha, \eta) \]
Collapsed Gibbs sampling (idea)

• What we get:

\[ p(z_{d,n} \mid z_{-d,n}, w_{d,n}, \alpha, \eta) = \frac{p(z_{d,n}, z_{-d,n} \mid w_{d,n}, \alpha, \eta)}{p(z_{-d,n} \mid w_{d,n}, \alpha, \eta)} \]
Collapsed Gibbs sampling (idea)

• Leads to:

\[
p(z_{d,n} | z_{-d,n}, w_{d,n}, \alpha, \eta)
= \frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \cdot \frac{v_{k,w_{d,n}} + \eta_{w_{d,n}}}{\sum_i^N v_{k,i} + \eta_i}
\]

where \( n_{d,i} \) is the number of words taking topic i in document d and \( v_{k,w_{d,n}} \) is the number of times \( w_{d,n} \) is used in topic k
### Example

**Sample document**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>mining</td>
<td>software</td>
<td>repositories</td>
<td>saves</td>
<td>lives</td>
</tr>
<tr>
<td>W</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\[
= \frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \eta_{w_{d,n}}}{\sum_i^N v_{k,i} + \eta_i}
\]

Sample document

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>mining</td>
<td>software</td>
<td>repositories</td>
<td>saves</td>
<td>lives</td>
</tr>
</tbody>
</table>

Total counts from corpus (v)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>mining</td>
<td>46</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>lives</td>
<td>3</td>
<td>75</td>
<td>0</td>
</tr>
<tr>
<td>repositories</td>
<td>0</td>
<td>63</td>
<td>4</td>
</tr>
<tr>
<td>saves</td>
<td>41</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>software</td>
<td>1</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\[
\begin{align*}
\text{Total counts from corpus (v)} & \\
\begin{array}{|c|c|c|c|}
\hline
 & 1 & 2 & 3 \\
\hline
\text{mining} & 46 & 3 & 0 \\
\text{lives} & 3 & 75 & 0 \\
\text{repositories} & 0 & 63 & 4 \\
\text{saves} & 41 & 0 & 0 \\
\text{software} & 1 & 13 & \textbf{7} \\
\ldots & & & \\
\hline
\end{array}
\end{align*}
\]
\[
\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \quad \frac{v_{k,w_{d,n}} + \eta_{w_{d,n}}}{\sum_i^N v_{k,i} + \eta_i}
\]

<table>
<thead>
<tr>
<th>Sample document</th>
<th>1</th>
<th>?</th>
<th>2</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>mining</td>
<td>mining</td>
<td>software</td>
<td>repositories</td>
<td>saves</td>
<td>lives</td>
</tr>
</tbody>
</table>
| Total counts from corpus (v) | 1 | 2 | 3 | 41 | 0 | 0 | 6 | 1 | 13 | 6 |...

|   | mining | lives | repositories | saves | software | ...
|---|--------|-------|-------------|-------|----------|
|   | 46     | 3     | 0           | 0     | 1        | 13      | 6

|   | mining | lives | repositories | saves | software | ...
|---|--------|-------|-------------|-------|----------|
| v | 3 | 75 | 63 | 4 | 13 |...
\[
    \text{Total counts from corpus (v)}
\]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>mining</td>
<td>46</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>lives</td>
<td>3</td>
<td>75</td>
<td>0</td>
</tr>
<tr>
<td>repos</td>
<td>0</td>
<td>63</td>
<td>4</td>
</tr>
<tr>
<td>saves</td>
<td>41</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>software</td>
<td>1</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
    \text{Sample document}
\]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>?</th>
<th>2</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>z</td>
<td>mining</td>
<td>software</td>
<td>repositories</td>
<td>saves</td>
<td>lives</td>
</tr>
<tr>
<td>w</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
    = \frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \quad \frac{v_{k,w_{d,n}} + \eta_{w_{d,n}}}{\sum_i^N v_{k,i} + \eta_i}
\]
\[
\text{Total counts from corpus (v)}
\]

\[
= \frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \quad \frac{v_{k,w_d,n} + \eta_{w_d,n}}{\sum_i^N v_{k,i} + \eta_i}
\]

<table>
<thead>
<tr>
<th>Z \</th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>mining</td>
<td>software</td>
<td>repositories</td>
<td>saves</td>
<td>lives</td>
</tr>
</tbody>
</table>

Sample document

<table>
<thead>
<tr>
<th>mining</th>
<th>software</th>
<th>repositories</th>
<th>saves</th>
<th>lives</th>
</tr>
</thead>
<tbody>
<tr>
<td>46</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>63</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>14</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\text{n}_{d,1} &= 2, \\
\text{n}_{d,2} &= 2, \\
\text{n}_{d,3} &= 0
\end{align*}
\]
\[
\frac{n_{d,k} + \alpha_k}{\sum_{i}^{K} n_{d,i} + \alpha_i} \quad \frac{v_{k,w_{d,n}} + \eta_{w_{d,n}}}{\sum_{i}^{N} v_{k,i} + \eta_i}
\]

- Where do \(\alpha\) and \(\eta\) come from?
  - (Hyper-)parameter for the model
  - Usually sampled (e.g. slice sampling) simultaneously
LDA in R

http://cran.r-project.org/web/packages/lda/lda.pdf
LDA - Strengths and Weaknesses

• Strengths
  – Realistically models multiple topics per document
  – Easily generalizable to new documents

• Weaknesses
  – Bag-of-words assumption has its limits
  – All words are forced into topics
EXTENSIONS OF LDA
Correlated Topic Model

• Model correlation between topic occurrences with a logistic normal distribution

Topical N-grams

- Model order of words using e.g. bigrams ...

(a) Bigram topic model

Topical N-grams

• ... or n-grams

Author-Topic Models

- Incorporate authorship as an influence for topics assignments

http://dl.acm.org/citation.cfm?id=1036902
• Incorporate authorship as an influence for topics assignments
Questions?