Structural Topic Models

Margaret Roberts

UC San Diego

May 25, 2017

Thanks to Justin Grimmer, Brandon Stewart, and Dustin Tingley from whom many of these slides were derived.
Topic models

- Methods of **unsupervised** text analysis
Topic models

- Methods of *unsupervised* text analysis
- Describe main *themes* of a corpus
Topic models

- Methods of **unsupervised** text analysis
- Describe main **themes** of a corpus
  - Starts with **term document matrix**
Topic models

- Methods of **unsupervised** text analysis
- Describe main **themes** of a corpus
  - Starts with **term document matrix**
  - Specify a **statistical model** for how the text was generated
Topic models

- Methods of unsupervised text analysis
- Describe main themes of a corpus
  - Starts with term document matrix
  - Specify a statistical model for how the text was generated
  - Find most likely topics that generated the text
Topic models

- Methods of unsupervised text analysis
- Describe main themes of a corpus
  - Starts with term document matrix
  - Specify a statistical model for how the text was generated
  - Find most likely topics that generated the text
- Similar to clustering, but with key differences
Topic models

- Methods of **unsupervised** text analysis
- Describe main **themes** of a corpus
  - Starts with **term document matrix**
  - Specify a **statistical model** for how the text was generated
  - Find **most likely** topics that generated the text
- Similar to **clustering**, but with key differences
- **Many** variants of topic models
Topic models

- Methods of **unsupervised** text analysis
- Describe main **themes** of a corpus
  - Starts with **term document matrix**
  - Specify a **statistical model** for how the text was generated
  - Find **most likely** topics that generated the text
- Similar to **clustering**, but with key differences
- **Many** variants of topic models
- Today: Latent Dirichlet Allocation and Structural Topic Model
Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003)

- Idea: don’t restrict topics to a single latent class, model topics as an admixture.
Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003)

- Idea: don’t restrict topics to a single latent class, model topics as an admixture.
- Each document is a mixture over topics. Each topic is a mixture over words.
Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003)

- Idea: don’t restrict topics to a single latent class, model topics as an admixture.
- Each document is a mixture over topics. Each topic is a mixture over words.
- Latent Dirichlet Allocation estimates:
Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003)

- Idea: don’t restrict topics to a single latent class, model topics as an admixture.
- Each document is a mixture over topics. Each topic is a mixture over words.

- Latent Dirichlet Allocation estimates:
  - The distribution over words for each topic.
Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003)

- Idea: don’t restrict topics to a single latent class, model topics as an admixture.
- Each document is a mixture over topics. Each topic is a mixture over words.

- Latent Dirichlet Allocation estimates:
  - The distribution over words for each topic.
  - The proportion of a document in each topic, for each document.
Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003)

- Idea: don’t restrict topics to a single latent class, model topics as an admixture.
- Each document is a mixture over topics. Each topic is a mixture over words.

- Latent Dirichlet Allocation estimates:
  - The distribution over words for each topic.
  - The proportion of a document in each topic, for each document.
Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003)

- Idea: don’t restrict topics to a single latent class, model topics as an admixture.
- Each document is a mixture over topics. Each topic is a mixture over words.

- Latent Dirichlet Allocation estimates:
  - The distribution over words for each topic.
  - The proportion of a document in each topic, for each document.

Maintained assumptions: Bag of words/fix number of topics ex ante.
What this means in pictures

Say you have a lot of people.

---

Roberts (UCSD)
STM
May 25, 2017 4 / 41
What this means in pictures

Say you have a lot of people. Each writes some texts.
What this means in pictures

Say you have a lot of people.

Each writes some texts that discuss a few different topics.

- Congress, nations, power, votes, agreement, bargaining
- Estimator, data, analysis, variance, model, inference
What this means in pictures

Say you have a lot of people.

Each writes some texts that discuss a few different topics.

Topic 1
What this means in pictures

Say you have a lot of people.

Each writes some texts that discuss a few different topics.

Topic 1

Topic 2
What this means in pictures

Say you have a lot of people.

Each writes some texts that discuss a few different topics

The Latent Dirichlet Allocation estimates:

1. The topics - each is a distribution over words
   - congress, nations, power, votes, agreement, bargaining
   - estimator, data, analysis, variance, model, inference

2. The proportion of each document in each topic
   - .7
   - .3

Roberts (UCSD)
What this means in pictures

Say you have a lot of people.

Each writes some texts that discuss a few different topics

The Latent Dirichlet Allocation estimates:

1. The topics—each is a distribution over words

Topic 1

Topic 2
Say you have a lot of people. Each writes some texts that discuss a few different topics.

The Latent Dirichlet Allocation estimates:

1. The topics—each is a distribution over words

   congress, nations, power, votes, agreement, bargaining

   estimator, data, analysis, variance, model, inference
What this means in pictures

Say you have a lot of people.

Each writes some texts that discuss a few different topics

The Latent Dirichlet Allocation estimates:

1. The topics—each is a distribution over words

2. The proportion of each document in each topic

- congress, nations, power, votes, agreement, bargaining
- estimator, data, analysis, variance, model, inference
Cluster 1

Cluster 2

Cluster K

Doc 1

Doc 2

Doc 3

...
Topic and Mixed Membership Models

Clustering
Document $\sim$ One Cluster

Doc 1 $\rightarrow$ Cluster 1
Doc 2 $\rightarrow$ Cluster 2
Doc 3 $\rightarrow$ ...
Doc $N$ $\rightarrow$ Cluster $K$
**Topic and Mixed Membership Models**

**Clustering**
Document $\sim$ One Cluster

Doc 1 → Cluster 1

Doc 2 → Cluster 2

Doc 3

... → Cluster K

Doc N
Topic and Mixed Membership Models

Clustering

Document \(\sim\) One Cluster

\[ \text{Doc 1} \]
\[ \text{Doc 2} \]
\[ \text{Doc 3} \]
\[ \vdots \]
\[ \text{Doc } N \]
Topic and Mixed Membership Models

Clustering
Document $\sim$ One Cluster

Doc 1 $\rightarrow$ Cluster 1

Doc 2

Doc 3

$\vdots$

Doc $N$ $\rightarrow$ Cluster $K$
Topic Models (Mixed Membership)
Document \(\rightsarrow\) Many clusters

Doc 1

Doc 2

Doc 3

\ldots

Doc \(N\)

Cluster 1

Cluster 2

\ldots

Cluster \(K\)
Topic and Mixed Membership Models

**Topic Models (Mixed Membership)**

Document $\sim$ Many clusters

- Doc 1
- Doc 2
- Doc 3
- \vdots
- Doc $N$

$\rightarrow$ Cluster 1

$\rightarrow$ Cluster 2

$\rightarrow$ Cluster $K$
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.
Topic Models

Two primary matrices of interest:

1) Topical Prevalence Matrix ($D \times K$)

$$\theta = \begin{bmatrix}
\text{Topic 1} & \text{Topic 2} & \ldots & \text{TopicK}
\text{Doc 1} & 2 & \ldots & 0.05
\text{Doc 2} & 2 & \ldots & 3 & \ldots & 0
\text{DocD} & 0 & \ldots & 0.05
\end{bmatrix}$$

2) Topical Content Matrix ($V \times K$)

$$\beta = \begin{bmatrix}
\text{Topic 1} & \text{Topic 2} & \ldots & \text{TopicK}
\text{text }'\prime' & 0.02 & \ldots & 0.001
\text{data }'\prime' & 0.001 & \ldots & 0.02
\text{analysis }'\prime' & 0.01 & \ldots & 0.0005
\text{... } & \ldots & \ldots & \ldots & \ldots
\end{bmatrix}$$
Topic Models

Two primary matrices of interest:

1) Topical Prevalence Matrix ($D \times K$)

$$\theta = \begin{bmatrix}
\text{Topic 1} & \text{Topic 2} & \ldots & \text{Topic K} \\
\text{Doc 1} & 2 & 0 & \ldots & 0.5 \\
\text{Doc 2} & 2 & 1 & \ldots & 3 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\text{Doc D} & 0 & 0 & \ldots & 5
\end{bmatrix}$$

2) Topical Content Matrix ($V \times K$)

$$\beta_T = \begin{bmatrix}
\text{Topic 1} & \text{Topic 2} & \ldots & \text{Topic K} \\
\text{“text”} & 0.2 & 0.001 & \ldots & 0.001 \\
\text{“data”} & 0.001 & 0.2 & \ldots & 0.001 \\
\text{“analysis”} & 0.01 & 0.01 & \ldots & 0.0005 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\end{bmatrix}$$
Topic Models

Two primary matrices of interest:

1) Topical Prevalence Matrix \((D \times K)\)
**Topic Models**

Two primary matrices of interest:

1) **Topical Prevalence Matrix** ($D \times K$)

\[
\theta = \begin{bmatrix}
\text{Topic}1 & \text{Topic}2 & \ldots & \text{Topic}K \\
\text{Doc}1 & .2 & .1 & \ldots & 0.05 \\
\text{Doc}2 & .2 & .1 & \ldots & .3 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\text{Doc}D & 0 & 0 & \ldots & .5 \\
\end{bmatrix}
\]
Topic Models

Two primary matrices of interest:

1) **Topical Prevalence Matrix** \((D \times K)\)

\[
\theta = \begin{bmatrix}
\text{Doc1} & \text{Topic1} & \text{Topic2} & \ldots & \text{TopicK} \\
.2 & .1 & \ldots & 0.05 \\
\text{Doc2} & .2 & .1 & \ldots & .3 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\text{DocD} & 0 & 0 & \ldots & .5
\end{bmatrix}
\]

2) **Topical Content Matrix** \((V \times K)\)
Topic Models

Two primary matrices of interest:

1) Topical Prevalence Matrix ($D \times K$)

$$X \approx \theta \beta$$

$$\theta = \begin{bmatrix}
\text{Doc1} & \text{Topic1} & \text{Topic2} & \ldots & \text{TopicK} \\
.2 & .1 & \ldots & 0.05 \\
\text{Doc2} & .2 & .1 & \ldots & .3 \\
\vdots & \vdots & \ddots & \vdots \\
\text{DocD} & 0 & 0 & \ldots & .5
\end{bmatrix}$$

2) Topical Content Matrix ($V \times K$)

$$\beta^T = \begin{bmatrix}
\text{“text”} & \text{Topic1} & \text{Topic2} & \ldots & \text{TopicK} \\
.02 & .001 & \ldots & 0.0001 \\
\text{“data”} & .001 & .02 & \ldots & 0.001 \\
\vdots & \vdots & \ddots & \vdots \\
\text{“analysis”} & .01 & .01 & \ldots & 0.0005
\end{bmatrix}$$
Vanilla Latent Dirichlet Allocation $\Rightarrow$ Objective Function

- Consider document $i$, ($i = 1, 2, \ldots, N$).
Vanilla Latent Dirichlet Allocation

⇝ Objective Function

- Consider document $i$, $(i = 1, 2, \ldots, N)$.
- Suppose there are $M_i$ total words and $x_i$ is an $M_i \times 1$ vector, where $x_{im}$ describes the $m^{th}$ word used in the document.

- $\beta_k \sim \text{Dirichlet}(\eta)$
- $\alpha_k \sim \text{Gamma}(\alpha, \beta)$
- $\theta_i | \alpha \sim \text{Dirichlet}(\alpha)$
- $z_{im} | \theta_i \sim \text{Multinomial}(1, \theta_i)$
- $x_{im} | \beta_k, z_{imk} = 1 \sim \text{Multinomial}(1, \beta_k)$

Optimization:
- Variational Approximation
- Find “closest” distribution
- Gibbs sampling
- MCMC algorithm to approximate posterior
Vanilla Latent Dirichlet Allocation \(\rightarrow\) Objective Function

- Consider document \(i\), \((i = 1, 2, \ldots, N)\).
- Suppose there are \(M_i\) total words and \(x_i\) is an \(M_i \times 1\) vector, where \(x_{im}\) describes the \(m^{th}\) word used in the document.
Vanilla Latent Dirichlet Allocation

- Consider document $i$, ($i = 1, 2, \ldots, N$).
- Suppose there are $M_i$ total words and $x_i$ is an $M_i \times 1$ vector, where $x_{im}$ describes the $m^{th}$ word used in the document.

$$\theta_i | \alpha \sim \text{Dirichlet}(\alpha)$$
Consider document \( i \), \((i = 1, 2, \ldots, N)\).

Suppose there are \( M_i \) total words and \( x_i \) is an \( M_i \times 1 \) vector, where \( x_{im} \) describes the \( m^{th} \) word used in the document.

\[
\theta_i | \alpha \sim \text{Dirichlet}(\alpha)
\]

\[
z_{im} | \theta_i \sim \text{Multinomial}(1, \theta_i)
\]
Vanilla Latent Dirichlet Allocation \rightarrow Objective Function

- Consider document $i$, ($i = 1, 2, \ldots, N$).
- Suppose there are $M_i$ total words and $x_i$ is an $M_i \times 1$ vector, where $x_{im}$ describes the $m^{th}$ word used in the document.

\begin{align*}
\theta_i | \alpha & \sim \text{Dirichlet}(\alpha) \\
z_{im} | \theta_i & \sim \text{Multinomial}(1, \theta_i) \\
x_{im} | \beta_k, z_{imk} = 1 & \sim \text{Multinomial}(1, \beta_k)
\end{align*}
Vanilla Latent Dirichlet Allocation \(\Rightarrow\) Objective Function

- Consider document \(i, (i = 1, 2, \ldots, N)\).
- Suppose there are \(M_i\) total words and \(x_i\) is an \(M_i \times 1\) vector, where \(x_{im}\) describes the \(m^{th}\) word used in the document.

\[
\beta_k \sim \text{Dirichlet}(\eta)
\]

\[
\theta_i|\alpha \sim \text{Dirichlet}(\alpha)
\]

\[
z_{im}|\theta_i \sim \text{Multinomial}(1, \theta_i)
\]

\[
x_{im}|\beta_k, z_{imk} = 1 \sim \text{Multinomial}(1, \beta_k)
\]

Optimization:
Vanilla Latent Dirichlet Allocation \(\mapsto\) Objective Function

- Consider document \(i\), \((i = 1, 2, \ldots, N)\).
- Suppose there are \(M_i\) total words and \(\mathbf{x}_i\) is an \(M_i \times 1\) vector, where \(x_{im}\) describes the \(m^{\text{th}}\) word used in the document.

\[
\begin{align*}
\beta_k & \sim \text{Dirichlet}(\eta) \\
\alpha_k & \sim \text{Gamma}(\alpha, \beta) \\
\theta_i|\alpha & \sim \text{Dirichlet}(\alpha) \\
z_{im}|\theta_i & \sim \text{Multinomial}(1, \theta_i) \\
x_{im}|\beta_k, z_{imk} = 1 & \sim \text{Multinomial}(1, \beta_k)
\end{align*}
\]

Optimization:
- Variational Approximation \(\mapsto\) Find “closest” distribution
Vanilla Latent Dirichlet Allocation \(\Rightarrow\) Objective Function

- Consider document \(i\), \((i = 1, 2, \ldots, N)\).
- Suppose there are \(M_i\) total words and \(x_i\) is an \(M_i \times 1\) vector, where \(x_{im}\) describes the \(m^{th}\) word used in the document.

\[
\begin{align*}
\beta_k & \sim \text{Dirichlet}(\eta) \\
\alpha_k & \sim \text{Gamma}(\alpha, \beta) \\
\theta_i|\alpha & \sim \text{Dirichlet}(\alpha) \\
z_{im}|\theta_i & \sim \text{Multinomial}(1, \theta_i) \\
 x_{im}|\beta_k, z_{imk} = 1 & \sim \text{Multinomial}(1, \beta_k)
\end{align*}
\]

Optimization:
- Variational Approximation \(\Rightarrow\) Find “closest” distribution
- Gibbs sampling \(\Rightarrow\) MCMC algorithm to approximate posterior
Example: Japanese Campaign Manifestos (Catalinac 2016)

- Why is Japan revising its constitution?
Example: Japanese Campaign Manifestos (Catalinac 2016)

- Why is Japan revising its constitution?
- **IR** question: why is Japan now willing to engage militaristic foreign action?

Determined (relentless) data collection

Latent Dirichlet Allocation (on Japanese texts)
- Why is Japan revising its constitution?
- **IR question**: why is Japan now willing to engage militaristic foreign action?
- **One explanation**: election reform in 1993, changed electoral incentives
Example: Japanese Campaign Manifestos (Catalinac 2016)

- Why is Japan revising its constitution?
- IR question: why is Japan now willing to engage militaristic foreign action?
- One explanation: election reform in 1993, changed electoral incentives
- To answer well: characterize campaigns across 50 + years
Example: Japanese Campaign Manifestos (Catalinac 2016)

- Why is Japan revising its constitution?
- IR question: why is Japan now willing to engage militaristic foreign action?
- One explanation: election reform in 1993, changed electoral incentives
- To answer well: characterize campaigns across 50 + years
  - That sounds hard
Example: Japanese Campaign Manifestos (Catalinac 2016)

- Why is Japan revising its constitution?
- IR question: why is Japan now willing to engage militaristic foreign action?
- One explanation: election reform in 1993, changed electoral incentives
- To answer well: characterize campaigns across 50 + years
  - That sounds hard
  - That sounds impossible
Example: Japanese Campaign Manifestos (Catalinac 2016)

- Why is Japan revising its constitution?
- IR question: why is Japan now willing to engage militaristic foreign action?
- One explanation: election reform in 1993, changed electoral incentives
- To answer well: characterize campaigns across 50 + years
  - That sounds hard
  - That sounds impossible
- Determined (relentless) data collection
Example: Japanese Campaign Manifestos (Catalinac 2016)

- Why is Japan revising its constitution?
- **IR question:** why is Japan now willing to engage militaristic foreign action?
- **One explanation:** election reform in 1993, changed electoral incentives
- To answer well: characterize campaigns across 50+ years
  - That sounds hard
  - That sounds impossible
- Determined (relentless) data collection
- Latent Dirichlet Allocation (on japanese texts)
Example: Japanese Campaign Manifestos (Catalinac 2016)

- Why is Japan revising its constitution?
- **IR question:** why is Japan now willing to engage militaristic foreign action?
- **One explanation:** election reform in 1993, changed electoral incentives
- To answer well: characterize campaigns across 50 + years
  - That sounds hard
  - That sounds impossible
- Determined (relentless) data collection
- Latent Dirichlet Allocation (on japanese texts)
Example: Japanese Campaign Manifestos (Catalinac 2016)

Japanese Elections:
Example: Japanese Campaign Manifestos (Catalinac 2016)

Japanese Elections:
- Election Administration Commission runs elections → district level
Japanese Elections:
- Election Administration Commission runs elections → district level
- Required to submit manifestos for all candidates to National Diet
新宿・千代田・港区の皆さん!

よりたしかな 未来を！あなたの1票を生かします！

私たちのくちにくちにかけてはならない政治家ね。

国内税制は極めて重要な税制改革を！

相続税導入して定住人口の確保を！

土地有効利用で都心若者若者の住宅を！

全力で取り組んでいます！

Roberts (UCSD)

STM

May 25, 2017 10 / 41
Example: Japanese Campaign Manifestos (Catalinac 2016)

Japanese Elections:
- Election Administration Commission runs elections → district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
Japanese Elections:

- Election Administration Commission runs elections → district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950-2009
  - Available only at district level
Japanese Elections:
- Election Administration Commission runs elections → district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950-2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm
Japanese Elections:
- Election Administration Commission runs elections → district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950-2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm
- Collected from microfilm, hand transcribed (no OCR worked), used a variety of techniques to create a TDM
Japanese Elections:
- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950-2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm
- Collected from microfilm, hand transcribed (no OCR worked), used a variety of techniques to create a TDM
- Harder for Japanese
Example: Japanese Campaign Manifestos (Catalinac 2016)

- Applies Vanilla LDA
- Output: topics (with Japanese characters)
### Example: Japanese Campaign Manifestos (Catalinac 2016)

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>改革</td>
<td>年金</td>
<td>推進</td>
<td>区</td>
<td>政治</td>
<td>改革</td>
</tr>
<tr>
<td>政府</td>
<td>円</td>
<td>政策</td>
<td>政策</td>
<td>国</td>
<td>政策</td>
</tr>
<tr>
<td>民主</td>
<td>郵政</td>
<td>納税</td>
<td>国民</td>
<td>日本</td>
<td>外交</td>
</tr>
<tr>
<td>小選挙区</td>
<td>廃止</td>
<td>国民</td>
<td>企業</td>
<td>日本</td>
<td>社会</td>
</tr>
<tr>
<td>構造</td>
<td>社会</td>
<td>総務</td>
<td>自民党</td>
<td>国民</td>
<td>社会</td>
</tr>
<tr>
<td>政府</td>
<td>実現</td>
<td>全力</td>
<td>国家</td>
<td>国民</td>
<td>安全</td>
</tr>
<tr>
<td>安全</td>
<td>実現</td>
<td>地域</td>
<td>日本</td>
<td>国家</td>
<td>安全</td>
</tr>
<tr>
<td>民主</td>
<td>増税</td>
<td>愛国</td>
<td>企業</td>
<td>日本</td>
<td>拉致</td>
</tr>
<tr>
<td>自民党</td>
<td>削減</td>
<td>経済</td>
<td>企業</td>
<td>日本</td>
<td>経済</td>
</tr>
<tr>
<td>日本</td>
<td>一元化</td>
<td>対策</td>
<td>北朝鮮</td>
<td>日本</td>
<td>経済</td>
</tr>
<tr>
<td>日本</td>
<td>政権</td>
<td>中小</td>
<td>Legalitarian</td>
<td>日本</td>
<td>北朝鮮</td>
</tr>
<tr>
<td>民間</td>
<td>政策</td>
<td>推進</td>
<td>教育</td>
<td>日本</td>
<td>北朝鮮</td>
</tr>
<tr>
<td>年金</td>
<td>流行</td>
<td>機関</td>
<td>教育</td>
<td>日本</td>
<td>教育</td>
</tr>
<tr>
<td>安全</td>
<td>ひと</td>
<td>学校</td>
<td>教育</td>
<td>日本</td>
<td>教育</td>
</tr>
<tr>
<td>進める</td>
<td>サラリーマン</td>
<td>東京</td>
<td>教育</td>
<td>日本</td>
<td>教育</td>
</tr>
<tr>
<td>障害</td>
<td>制度</td>
<td>東京</td>
<td>教育</td>
<td>日本</td>
<td>教育</td>
</tr>
<tr>
<td>地方</td>
<td>連携</td>
<td>公民</td>
<td>政治</td>
<td>日本</td>
<td>政治</td>
</tr>
<tr>
<td>地方</td>
<td>福祉</td>
<td>守る</td>
<td>政治</td>
<td>日本</td>
<td>政治</td>
</tr>
<tr>
<td>保険</td>
<td>斎院</td>
<td>活性</td>
<td>政策</td>
<td>日本</td>
<td>政策</td>
</tr>
<tr>
<td>政策</td>
<td>年金</td>
<td>自民党</td>
<td>多数</td>
<td>日本</td>
<td>政策</td>
</tr>
<tr>
<td>作る</td>
<td>一揆</td>
<td>地方</td>
<td>安心</td>
<td>日本</td>
<td>政策</td>
</tr>
<tr>
<td>効能</td>
<td>郵政</td>
<td>改革</td>
<td>反対</td>
<td>日本</td>
<td>政策</td>
</tr>
<tr>
<td>社会</td>
<td>道路</td>
<td>経済</td>
<td>政策</td>
<td>日本</td>
<td>政策</td>
</tr>
<tr>
<td>国民</td>
<td>項目</td>
<td>支持</td>
<td>政策</td>
<td>日本</td>
<td>政策</td>
</tr>
<tr>
<td>公務員</td>
<td>社会教養</td>
<td>ひと</td>
<td>政策</td>
<td>日本</td>
<td>政策</td>
</tr>
<tr>
<td>支持</td>
<td>月額</td>
<td>場面</td>
<td>政策</td>
<td>日本</td>
<td>政策</td>
</tr>
<tr>
<td>支援</td>
<td>郵政</td>
<td>総務</td>
<td>政策</td>
<td>日本</td>
<td>政策</td>
</tr>
<tr>
<td>安心</td>
<td>支援</td>
<td>締結</td>
<td>政策</td>
<td>日本</td>
<td>政策</td>
</tr>
</tbody>
</table>

Postal privatization | Reducing Wasteful Public Spending | Pork for the District | Policies for the district | Political Reform | National Security Policy

---

Roberts (UCSD)      STM                    May 25, 2017 12 / 41
Example: Japanese Campaign Manifestos (Catalinac 2011)

Change in Mean Proportion of Each Manifesto Devoted to Pork Over Time

Proportions of each Manifesto Devoted to Pork

Election Years


0.1
Example: Japanese Campaign Manifestos (Catalinac 2011)

Change in Mean Proportion of Each Manifesto Devoted to Foreign Policy Over Time
How well does our model perform?

$$\text{Perplexity} = \exp \left( -\log p(x^*_{\text{out}} | \mu, \pi) \right)$$
How well does our model perform? \( \leadsto \) predict new documents?
Measuring Topic Performance: Out of Sample Prediction

How well does our model perform? predict new documents?

Problem
Measuring Topic Performance: Out of Sample Prediction

How well does our model perform? \[ \text{predict new documents?} \]
Problem \[ \text{in sample evaluation leads to overfit.} \]
Measuring Topic Performance: Out of Sample Prediction

How well does our model perform? predict new documents?
Problem in sample evaluation leads to overfit.
Solution evaluate performance on held out data
How well does our model perform? predict new documents?

Problem: in sample evaluation leads to overfit.

Solution: evaluate performance on held out data

For held out document $x^*_\text{out}$
Measuring Topic Performance: Out of Sample Prediction

How well does our model perform? predict new documents?

Problem in sample evaluation leads to overfit.

Solution evaluate performance on held out data

For held out document $x^*_\text{out}$

$$\text{Perplexity} = \exp(-\log p(x^*_\text{out}|\mu, \pi))$$
What’s Prediction Got to Do With It?

- Prediction $\rightarrow$ One Task
What’s Prediction Got to Do With It?

- Prediction $\rightarrow$ One Task
- Do we care about it?
What’s Prediction Got to Do With It?

- Prediction $\rightarrow$ One Task
- Do we care about it? $\rightarrow$ Social science application where we’re predicting new texts?

Chang et al 2009 (“Reading the Tea Leaves”):
- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

Different strategy:
- measure quality in topics and clusters
- Experiments: measure topic and cluster quality
What’s Prediction Got to Do With It?

- Prediction $\rightarrow$ One Task
- Do we care about it? $\rightarrow$ Social science application where we’re predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 (“Reading the Tea Leaves”):
- Compare perplexity with human based evaluations
- Negative relationship between perplexity and human based evaluations

Different strategy:
- Measure quality in topics and clusters
- Experiments: measure topic and cluster quality
What’s Prediction Got to Do With It?

- Prediction $\Rightarrow$ One Task
- Do we care about it? $\Rightarrow$ Social science application where we’re predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves"):

...
What’s Prediction Got to Do With It?

- Prediction $\Rightarrow$ One Task
- Do we care about it? $\Rightarrow$ Social science application where we’re predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 (“Reading the Tea Leaves”):
- Compare perplexity with *human* based evaluations
What’s Prediction Got to Do With It?

- Prediction ⇾ One Task
- Do we care about it? ⇾ Social science application where we’re predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 (“Reading the Tea Leaves”):
- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations
What’s Prediction Got to Do With It?

- Prediction ⇒ One Task
- Do we care about it? ⇒ Social science application where we’re predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves") :
- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

Different strategy ⇒ measure quality in topics and clusters
What’s Prediction Got to Do With It?

- Prediction $\rightarrow$ One Task
- Do we care about it? $\rightarrow$ Social science application where we’re predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 (“Reading the Tea Leaves”):
- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

Different strategy $\rightarrow$ measure quality in topics and clusters
What’s Prediction Got to Do With It?

- Prediction $\Rightarrow$ One Task
- Do we care about it? $\Rightarrow$ Social science application where we’re predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 (“Reading the Tea Leaves”):
- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

Different strategy $\Rightarrow$ measure quality in topics and clusters
- Experiments: measure topic and cluster quality
Measuring Cohesiveness and Exclusivity

- Consider the output of a topic model
- We might select 5 top words for each topic
  - Topic 1
    - bill, congressman, earmarks, following, house
  - Topic 2
    - immigration, reform, security, border, worker
  - Topic 3
    - earmark, egregious, pork, fiscal, today

- An ideal topic?
  ⇝ will see these words co-occur in documents

- Define $v_k = (v_1^k, v_2^k, \ldots, v_L^k)$ be the top words for a topic
  - For example $v_3 = (earmark, egregious, pork, fiscal, today)$
Measuring Cohesiveness and Exclusivity

- Consider the output of a topic model

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>bill</td>
<td>immigration</td>
<td>earmark</td>
</tr>
<tr>
<td>congressman</td>
<td>reform</td>
<td>egregious</td>
</tr>
<tr>
<td>earmarks</td>
<td>security</td>
<td>pork</td>
</tr>
<tr>
<td>following</td>
<td>border</td>
<td>fiscal</td>
</tr>
<tr>
<td>house</td>
<td>worker</td>
<td>today</td>
</tr>
</tbody>
</table>

- An ideal topic?
  - will see these words co-occur in documents

Define $v_k = (v_1^k, v_2^k, ..., v_L^k)$ be the top words for a topic
Measuring Cohesiveness and Exclusivity

- Consider the output of a topic model
- We might select 5 top words for each topic
Measuring Cohesiveness and Exclusivity

- Consider the output of a topic model.
- We might select 5 top words for each topic.

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>bill</th>
<th>congressman</th>
<th>earmarks</th>
<th>following</th>
<th>house</th>
</tr>
</thead>
</table>

An ideal topic might see these words co-occur in documents.

Define $v_k = (v_1^k, v_2^k, ..., v_L^k)$ be the top words for a topic. For example, $v_3 = (earmark, egregious, pork, fiscal, today)$. 
Measuring Cohesiveness and Exclusivity

- Consider the output of a topic model
- We might select 5 top words for each topic

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>bill</th>
<th>congressman</th>
<th>earmarks</th>
<th>following</th>
<th>house</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 2</td>
<td>immigration</td>
<td>reform</td>
<td>security</td>
<td>border</td>
<td>worker</td>
</tr>
</tbody>
</table>
Measuring Cohesiveness and Exclusivity

- Consider the output of a topic model
- We might select 5 top words for each topic

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>bill</th>
<th>congressman</th>
<th>earmarks</th>
<th>following</th>
<th>house</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 2</td>
<td>immigration</td>
<td>reform</td>
<td>security</td>
<td>border</td>
<td>worker</td>
</tr>
<tr>
<td>Topic 3</td>
<td>earmark</td>
<td>egregious</td>
<td>pork</td>
<td>fiscal</td>
<td>today</td>
</tr>
</tbody>
</table>
Measuring Cohesiveness and Exclusivity

- Consider the output of a topic model.
- We might select 5 top words for each topic.

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>bill</th>
<th>congressman</th>
<th>earmarks</th>
<th>following</th>
<th>house</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 2</td>
<td>immigration</td>
<td>reform</td>
<td>security</td>
<td>border</td>
<td>worker</td>
</tr>
<tr>
<td>Topic 3</td>
<td>earmark</td>
<td>egregious</td>
<td>pork</td>
<td>fiscal</td>
<td>today</td>
</tr>
</tbody>
</table>

- An ideal topic? will see these words co-occur in documents.
Measuring Cohesiveness and Exclusivity

- Consider the output of a topic model
- We might select 5 top words for each topic

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>bill</th>
<th>congressman</th>
<th>earmarks</th>
<th>following</th>
<th>house</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 2</td>
<td>immigration</td>
<td>reform</td>
<td>security</td>
<td>border</td>
<td>worker</td>
</tr>
<tr>
<td>Topic 3</td>
<td>earmark</td>
<td>egregious</td>
<td>pork</td>
<td>fiscal</td>
<td>today</td>
</tr>
</tbody>
</table>

- An ideal topic will see these words co-occur in documents
- Define $v_k = (v_{1k}, v_{2k}, \ldots, v_{Lk})$ be the top words for a topic
Measuring Cohesiveness and Exclusivity

- Consider the output of a topic model
- We might select 5 top words for each topic

<table>
<thead>
<tr>
<th>Topic</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>bill, congressman, earmarks, following, house</td>
</tr>
<tr>
<td>Topic 2</td>
<td>immigration, reform, security, border, worker</td>
</tr>
<tr>
<td>Topic 3</td>
<td>earmark, egregious, pork, fiscal, today</td>
</tr>
</tbody>
</table>

- An ideal topic? \( \Rightarrow \) will see these words co-occur in documents
- Define \( \mathbf{v}_k = (v_{1k}, v_{2k}, \ldots, v_{Lk}) \) be the top words for a topic
- For example \( \mathbf{v}_3 = (\text{earmark, egregious, pork, fiscal, today}) \)
Measuring Cohesiveness and Exclusivity

Define the function $D$ as a function that counts the number of times its argument occurs:

$$D(\text{earmark, egregious}) = \text{No. times earmark and egregious co-occur}$$

$$D(\text{egregious}) = \text{Number of times Egregious occurs}$$

Define cohesiveness for topic $k$ as

$$C_{\text{Cohesive}}^k = \left( \sum_{l=2}^{L} \sum_{m=1}^{L-1} \log \left( \frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})} \right) \right)$$

Define overall cohesiveness as:

$$C_{\text{Cohesive}} = \left( \frac{1}{K} \sum_{k=1}^{K} C_{\text{Cohesive}}^k \right)$$
Measuring Cohesiveness and Exclusivity

Define the function $D$ as a function that counts the number of times its argument occurs:

$$D(\text{earmark, egregious}) = \text{No. times earmark and egregious co-occur}$$
Measuring Cohesiveness and Exclusivity

Define the function $D$ as a function that counts the number of times its argument occurs:

\[ D(\text{earmark, egregious}) = \text{No. times earmark and egregious co-occur} \]
\[ D(\text{egregious}) = \text{Number of times Egregious occurs} \]
Measuring Cohesiveness and Exclusivity

Define the function $D$ as a function that counts the number of times its argument occurs:

$$D(\text{earmark, egregious}) = \text{No. times earmark and egregious co-occur}$$

$$D(\text{egregious}) = \text{Number of times Egregious occurs}$$

Define cohesiveness for topic $k$ as

$$\text{Cohesive}_k = \sum_{l=2}^{L} \sum_{m=1}^{M} \log \left( \frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})} \right)$$

Define overall cohesiveness as:

$$\text{Cohesive} = \frac{1}{K} \sum_{k=1}^{K} \text{Cohesive}_k$$
Measuring Cohesiveness and Exclusivity

Define the function $D$ as a function that counts the number of times its argument occurs:

\[ D(\text{earmark, egregious}) = \text{No. times earmark and egregious co-occur} \]
\[ D(\text{egregious}) = \text{Number of times Egregious occurs} \]

Define cohesiveness for topic $k$ as

\[
\text{Cohesive}_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log \left( \frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})} \right)
\]

Define overall cohesiveness as:

\[
\text{Cohesive} = \frac{\sum_{k=1}^{K} \text{Cohesive}_k}{K}
\]
Measuring Cohesiveness and Exclusivity

Define the function $D$ as a function that counts the number of times its argument occurs:

$$D(\text{earmark, egregious}) = \text{No. times earmark and egregious co-occur}$$

$$D(\text{egregious}) = \text{Number of times Egregious occurs}$$

Define cohesiveness for topic $k$ as

$$\text{Cohesive}_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log \left( \frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})} \right)$$

Define overall cohesiveness as:

$$\text{Cohesive} = \frac{\sum_{k=1}^{K} \text{Cohesive}_k}{K}$$
Measuring Cohesiveness and Exclusivity

Define the function $D$ as a function that counts the number of times its argument occurs:

$$D(\text{earmark, egregious}) = \text{No. times earmark and egregious co-occur}$$

$$D(\text{egregious}) = \text{Number of times Egregious occurs}$$

Define cohesiveness for topic $k$ as

$$\text{Cohesive}_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log \left( \frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})} \right)$$

Define overall cohesiveness as:

$$\text{Cohesive} = \left( \sum_{k=1}^{K} \text{Cohesive}_k \right) / K$$
Measuring Cohesiveness and Exclusivity

Define the function $D$ as a function that counts the number of times its argument occurs:

\[ D(\text{earmark, egregious}) = \text{No. times earmark and egregious co-occur} \]
\[ D(\text{egregious}) = \text{Number of times Egregious occurs} \]

Define cohesiveness for topic $k$ as

\[
\text{Cohesive}_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log \left( \frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})} \right)
\]

Define overall cohesiveness as:

\[
\text{Cohesive} = \left( \sum_{k=1}^{K} \text{Cohesive}_k \right) / K
\]

\[
= \left( \sum_{k=1}^{K} \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log \left( \frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})} \right) \right) / K
\]
Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive
Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive \(\rightarrow\) few replicates of each topic
Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive to few replicates of each topic.

\[
\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}
\]
Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive\(\iff\) few replicates of each topic

\[
\text{Exclusivity}(k, \nu) = \frac{\mu_{k, \nu}}{\sum_{l=1}^{K} \mu_{l, \nu}}
\]

Suppose again we pick \(L\) top words. Measure Exclusivity for a topic as for a topic as:
Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive\[\implies\] few replicates of each topic

\[
\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}
\]

Suppose again we pick \( L \) top words. Measure Exclusivity for a topic as for a topic as:

\[
\text{Exclusivity}_k = \sum_{j: v_j \in v_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}}
\]
Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive \( \rightarrow \) few replicates of each topic

\[
\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}
\]

Suppose again we pick \( L \) top words. Measure Exclusivity for a topic as for a topic as:

\[
\text{Exclusivity}_k = \sum_{j: v_j \in v_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}}
\]

\[
\text{Exclusivity} = \left( \sum_{k=1}^{K} \text{Exclusivity}_k \right) / K
\]
Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive \( \approx \) few replicates of each topic.

\[
\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}
\]

Suppose again we pick \( L \) top words. Measure Exclusivity for a topic as for a topic as:

\[
\text{Exclusivity}_k = \sum_{j: v_j \in v_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}}
\]

\[
\text{Exclusivity} = \left( \sum_{k=1}^{K} \text{Exclusivity}_k \right) / K
\]

\[
= \left( \sum_{k=1}^{K} \sum_{j: v_j \in v_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}} \right) / K
\]
How do we Choose $K$?

Generate many candidate models

1) Assess Cohesiveness/Exclusivity, select models on frontier
2) Use experiments
3) Read
4) Final decision $\rightarrow$ combination
Examples of Topic Models

- How do senators present their work to the public? What explains variation in representational style? (Grimmer 2013)

- Does electoral reform alter the content of Japanese Party manifestos? (Catalinac 2016)

- How do Muslim clerics supporting violent Jihad differ from those who do not in choice of fatwa topics? (Nielsen 2013)

- Do presidential candidates move to the center after the convention? (Gross et al 2013)
Examples of Topic Models

- How do senators present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Does electoral reform alter the content of Japanese Party manifestos? (Catalinac 2016)
Examples of Topic Models

- How do senators present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Does electoral reform alter the content of Japanese Party manifestos? (Catalinac 2016)
- How do Muslim clerics supporting violent Jihad differ from those who do not in choice of fatwa topics? (Nielsen 2013)
Examples of Topic Models

- How do senators present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Does electoral reform alter the content of Japanese Party manifestos? (Catalinac 2016)
- How do Muslim clerics supporting violent Jihad differ from those who do not in choice of fatwa topics? (Nielsen 2013)
- Do presidential candidates move to the center after the convention? (Gross et al 2013)
Examples of Topic Models

- How do senators present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Does electoral reform alter the content of Japanese Party manifestos? (Catalinac 2016)
- How do Muslim clerics supporting violent Jihad differ from those who do not in choice of fatwa topics? (Nielsen 2013)
- Do presidential candidates move to the center after the convention? (Gross et al 2013)
Elements of a Common Structure

- Measuring variation of topics with some observed covariates
Elements of a Common Structure

- Measuring variation of topics with some observed covariates
- Interest in aggregate trends (e.g. proportion of total press release from a given center about appropriations)
Elements of a Common Structure

- Measuring variation of topics with some observed covariates
- Interest in aggregate trends (e.g. proportion of total press release from a given center about appropriations)
- We want to tell a story not just about what, but *how* and *why*
In Practice

- Run standard LDA model and estimate covariate effects after the fact

First we assume exchangeability then we show it doesn’t hold! Designing custom models would be better but too much for practitioners. Practitioners see hundreds of options—but hard to find one that fits individual cases.
In Practice

- Run standard LDA model and estimate covariate effects after the fact
- First we assume exchangeability then we show it doesn’t hold!
In Practice

- Run standard LDA model and estimate covariate effects after the fact
- First we assume exchangeability then we show it doesn’t hold!
- Designing custom models would be better but too much for practitioners
In Practice

- Run standard LDA model and estimate covariate effects after the fact
- First we assume exchangeability then we show it doesn’t hold!
- Designing custom models would be better but too much for practitioners
- Practitioners see hundreds of options- but hard to find one that fits individual cases.
In Practice

- Run standard LDA model and estimate covariate effects after the fact
- First we assume exchangeability then we show it doesn’t hold!
- Designing custom models would be better but too much for practitioners
- Practitioners see hundreds of options- but hard to find one that fits individual cases.
Goal of Structural Topic Model (Roberts, Stewart, Tingley et al (2014))

Provide a basic framework for applied users to incorporate observed data which is
Goal of Structural Topic Model (Roberts, Stewart, Tingley et al (2014))

Provide a basic framework for applied users to incorporate observed data which is

- Easy to use (R package)
Goal of Structural Topic Model (Roberts, Stewart, Tingley et al (2014))

Provide a basic framework for applied users to incorporate observed data which is

- Easy to use (R package)
- Flexible
Goal of Structural Topic Model (Roberts, Stewart, Tingley et al (2014))

Provide a basic framework for applied users to incorporate observed data which is

- Easy to use (R package)
- Flexible
- Integrated with support tools (visualization/uncertainty calculation/model selection)
- See structuraltopicmodel.com
Leveraging Information Within and About Texts

Previous methods leverage the information within documents, primarily analyzing unstructured text, using words within a document to infer its subject. But, we also have information about documents captured by metadata: data about data, such as author, source, date, audience. This is important because speech is deeply contextual, for example, who says it, where, when, to whom. We want to avoid throwing away valuable information we have.

Structural Topic Model (STM) is a general method for modeling documents with context, enabling comparison of document sets. Two uses of metadata: topic prevalence and topical content.
Previous methods leverage the information within documents.
Leveraging Information Within and About Texts

- Previous methods leverage the information within documents
  - methods developed in computer science and statistics
  - primarily analyzing unstructured text
  - use words within document to infer its subject

But, we also have information about documents
- captured by metadata: data about data
  - e.g. author, source, date, audience
  - important because speech is deeply contextual
  - e.g. who says it, where, when, to whom

we want to avoid throwing away valuable information we have

Structural Topic Model (STM)
- general method for modeling documents with context
  - modeling context in document sets with enable comparison
  - two uses of metadata: topic prevalence and topical content
Leveraging Information Within and About Texts

- Previous methods leverage the information within documents
  - methods developed in computer science and statistics
  - primarily analyzing unstructured text
Leveraging Information Within and About Texts

- Previous methods leverage the information *within* documents
  - methods developed in computer science and statistics
  - primarily analyzing unstructured text
  - use words *within* document to infer its subject

But, we also have information about documents
- captured by metadata: data about data
  - e.g. author, source, date, audience
  - important because speech is deeply contextual
    - e.g. who says it, where, when, to whom

Structural Topic Model (STM)
- general method for modeling documents with context
  - modeling context in document sets with enable comparison
  - two uses of metadata: topic prevalence and topical content
Leveraging Information Within and About Texts

- Previous methods leverage the information within documents
  - methods developed in computer science and statistics
  - primarily analyzing unstructured text
  - use words within document to infer its subject
- But, we also have information about documents
  - captured by metadata: data about data
  - e.g. author, source, date, audience
  - important because speech is deeply contextual
    - e.g. who says it, where, when, to whom
  - we want to avoid throwing away valuable information we have

Structural Topic Model (STM)
- general method for modeling documents with context
- modeling context in document sets with enable comparison
- two uses of metadata: topic prevalence and topical content
Leveraging Information **Within** and **About** Texts

- Previous methods leverage the information **within** documents
  - methods developed in computer science and statistics
  - primarily analyzing unstructured text
  - use words **within** document to infer its subject
- But, we also have information **about** documents
  - captured by metadata: data about data

---

Roberts (UCSD)  May 25, 2017  24 / 41
Leveraging Information Within and About Texts

- Previous methods leverage the information within documents
  - methods developed in computer science and statistics
  - primarily analyzing unstructured text
  - use words within document to infer its subject

- But, we also have information about documents
  - captured by metadata: data about data
  - e.g. author, source, date, audience
Leveraging Information Within and About Texts

- Previous methods leverage the information within documents
  - methods developed in computer science and statistics
  - primarily analyzing unstructured text
  - use words within document to infer its subject
- But, we also have information about documents
  - captured by metadata: data about data
  - e.g. author, source, date, audience
  - important because speech is deeply contextual
Leveraging Information Within and About Texts

- Previous methods leverage the information within documents
  - methods developed in computer science and statistics
  - primarily analyzing unstructured text
  - use words within document to infer its subject

- But, we also have information about documents
  - captured by metadata: data about data
  - e.g. author, source, date, audience
  - important because speech is deeply contextual
  - e.g. who says it, where, when, to whom

Structural Topic Model (STM)
- general method for modeling documents with context
- modeling context in document sets with enable comparison
- two uses of metadata: topic prevalence and topical content
Leveraging Information Within and About Texts

- Previous methods leverage the information within documents
  - methods developed in computer science and statistics
  - primarily analyzing unstructured text
  - use words within document to infer its subject
- But, we also have information about documents
  - captured by metadata: data about data
  - e.g. author, source, date, audience
  - important because speech is deeply contextual
  - e.g. who says it, where, when, to whom
  - we want to avoid throwing away valuable information we have
Leveraging Information Within and About Texts

- Previous methods leverage the information within documents
  - methods developed in computer science and statistics
  - primarily analyzing unstructured text
  - use words within document to infer its subject

- But, we also have information about documents
  - captured by metadata: data about data
  - e.g. author, source, date, audience
  - important because speech is deeply contextual
  - e.g. who says it, where, when, to whom
  - we want to avoid throwing away valuable information we have

- Structural Topic Model (STM)
Leveraging Information Within and About Texts

- Previous methods leverage the information within documents
  - methods developed in computer science and statistics
  - primarily analyzing unstructured text
  - use words within document to infer its subject

- But, we also have information about documents
  - captured by metadata: data about data
  - e.g. author, source, date, audience
  - important because speech is deeply contextual
  - e.g. who says it, where, when, to whom
  - we want to avoid throwing away valuable information we have

- Structural Topic Model (STM)
  - general method for modeling documents with context
Leveraging Information **Within** and **About** Texts

- Previous methods leverage the information **within** documents
  - methods developed in computer science and statistics
  - primarily analyzing unstructured text
  - use words **within** document to infer its subject

- But, we also have information **about** documents
  - captured by **metadata**: data about data
  - e.g. author, source, date, audience
  - important because speech is deeply **contextual**
  - e.g. who says it, where, when, to whom
  - we want to avoid throwing away valuable information we have

- **Structural Topic Model (STM)**
  - general method for modeling documents with context
  - modeling context in document sets with enable **comparison**
Leveraging Information Within and About Texts

- Previous methods leverage the information within documents
  - methods developed in computer science and statistics
  - primarily analyzing unstructured text
  - use words within document to infer its subject

- But, we also have information about documents
  - captured by metadata: data about data
  - e.g. author, source, date, audience
  - important because speech is deeply contextual
  - e.g. who says it, where, when, to whom
  - we want to avoid throwing away valuable information we have

- Structural Topic Model (STM)
  - general method for modeling documents with context
  - modeling context in document sets with enable comparison
  - two uses of metadata: topic prevalence and topical content
STM = LDA + Contextual Information

STM provides two ways to include contextual information:

▶ Topic prevalence can vary by metadata
    ⋆ e.g. Democrats talk more about education than Republicans

▶ Topic content can vary by metadata
    ⋆ e.g. Democrats are less likely to use the word “life” when talking about abortion than Republicans

Including context improves the model:

▶ more accurate estimation
▶ better qualitative interpretability
STM = LDA + Contextual Information

STM provides two ways to include contextual information:

- Topic prevalence can vary by metadata, e.g., Democrats talk more about education than Republicans.
- Topic content can vary by metadata, e.g., Democrats are less likely to use the word "life" when talking about abortion than Republicans.

Including context improves the model:

- More accurate estimation.
- Better qualitative interpretability.
STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - e.g. Democrats talk more about education than Republicans
    - e.g. Democrats are less likely to use the word "life" when talking about abortion than Republicans

Including context improves the model:
  - more accurate estimation
  - better qualitative interpretability
STM = LDA + Contextual Information

STM provides two ways to include contextual information:

- **Topic prevalence** can vary by metadata
  - e.g. Democrats talk more about education than Republicans

- **Topic content** can vary by metadata
  - e.g. Democrats are less likely to use the word “life” when talking about abortion than Republicans
STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic *prevalence* can vary by metadata
    - e.g. Democrats talk more about education than Republicans
  - Topic *content* can vary by metadata
STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic **prevalence** can vary by metadata
    - e.g. Democrats talk more about education than Republicans
  - Topic **content** can vary by metadata
    - e.g. Democrats are less likely to use the word “life” when talking about abortion than Republicans
STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - e.g. Democrats talk more about education than Republicans
  - Topic content can vary by metadata
    - e.g. Democrats are less likely to use the word “life” when talking about abortion than Republicans

- Including context improves the model:
STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - e.g. Democrats talk more about education than Republicans
  - Topic content can vary by metadata
    - e.g. Democrats are less likely to use the word “life” when talking about abortion than Republicans

- Including context improves the model:
  - more accurate estimation
STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - e.g. Democrats talk more about education than Republicans
  - Topic content can vary by metadata
    - e.g. Democrats are less likely to use the word “life” when talking about abortion than Republicans

- Including context improves the model:
  - more accurate estimation
  - better qualitative interpretability
STM: What this means in pictures

Say you have a lot of people.

Each writes some text that discuss a few different topics:

Politics
- congress, nations, power, votes, agreement, bargaining

Statistics
- estimator, data, analysis, variance, model, inference

The STM Allows for:

Roberts (UCSD)
STM: What this means in pictures

Say you have a lot of people.

Each writes some text that discusses a few different topics.

The STM Allows for:

1. The words in each topic to vary by gender.
STM: What this means in pictures

Say you have a lot of people. Each writes some text that discuss a few different topics.

The STM Allows for:
1. The words in each topic to vary by gender

Politics
- congress, nations, power, votes, agreement, bargaining

Statistics
- estimator, data, analysis, variance, model, inference
Say you have a lot of people. Each writes some text that discuss a few different topics.

The STM Allows for:
1. The words in each topic to vary by gender
2. The topic proportions to vary by group

Politics
- congress, nations, power, votes, agreement, bargaining

Statistics
- estimator, data, analysis, variance, model, inference
STM: What this means in pictures

Say you have a lot of people. Each writes some text that discuss a few different topics.

Group A

Group B

Group A

The STM Allows for:

1. The words in each topic to vary by gender
2. The topic proportions to vary by group
Mixed-Membership Topic Models

More formal terminology:
Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: $K$
Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: \( K \)
- Observed data for standard topic models
Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: \( K \)
- Observed data for standard topic models
  - Each document \((i \in 1 \ldots D)\) is a collection of \( M_i \) tokens
Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: $K$
- Observed data for standard topic models
  - Each document ($i \in 1 \ldots D$) is a collection of $M_i$ tokens
- Additional data for STM
Mixed-Membership Topic Models

More formal terminology:
- User specifies the number of topics: $K$
- Observed data for standard topic models
  - Each document ($i \in 1 \ldots D$) is a collection of $M_i$ tokens
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $X$
Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: $K$
- Observed data for standard topic models
  - Each document $(i \in 1 \ldots D)$ is a collection of $M_i$ tokens
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $X$
  - Topical content groups: $D$ length vector $Y$
Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: $K$
- Observed data for standard topic models
  - Each document ($i \in 1 \ldots D$) is a collection of $M_i$ tokens
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $X$
  - Topical content groups: $D$ length vector $Y$
- Latent variables
Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: $K$
- Observed data for standard topic models
  - Each document ($i \in 1 \ldots D$) is a collection of $M_i$ tokens
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $X$
  - Topical content groups: $D$ length vector $Y$
- Latent variables
  - $D \times K$ matrix $\theta$: proportion of document on each topic.
Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: $K$
- Observed data for standard topic models
  - Each document ($i \in 1 \ldots D$) is a collection of $M_i$ tokens
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $X$
  - Topical content groups: $D$ length vector $Y$
- Latent variables
  - $D \times K$ matrix $\theta$: proportion of document on each topic.
  - $K \times V$ matrix $\beta$: probability of drawing a word conditional on topic.
Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: $K$
- Observed data for standard topic models
  - Each document ($i \in 1\ldots D$) is a collection of $M_i$ tokens
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $X$
  - Topical content groups: $D$ length vector $Y$
- Latent variables
  - $D \times K$ matrix $\theta$: proportion of document on each topic.
  - $K \times V$ matrix $\beta$: probability of drawing a word conditional on topic.
The Structural Topic Model

- $\theta$, $D \times K$ document-topic matrix

- $\beta$, $K \times V$ topic-word matrix

- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{i,m}$ from $\text{Multinomial}(\theta_i)$
  - Draw observed word $w_{i,m}$ from $\text{Multinomial}(\beta_{k=z,})$
The Structural Topic Model

- \( \theta, D \times K \) document-topic matrix \( \leftarrow \) logistic normal glm with covariates

- \( \beta, K \times V \) topic-word matrix

Each token has a topic drawn from the document mixture
  - Draw token topic \( z_{i,m} \) from Multinomial(\( \theta_i \))
  - Draw observed word \( w_{i,m} \) from Multinomial(\( \beta_{k=z} \))
The Structural Topic Model

- $\theta$, $D \times K$ document-topic matrix $\leftarrow$ logistic normal glm with covariates
  - Covariate-specific prior with global topic covariance
  - $\theta_{i,:} \sim \text{LogisticNormal}(X_i \gamma, \Sigma)$
- $\beta$, $K \times V$ topic-word matrix

Each token has a topic drawn from the document mixture
- Draw token topic $z_{i,m}$ from Multinomial($\theta_i$)
- Draw observed word $w_{i,m}$ from Multinomial($\beta_{k=z,:}$)
The Structural Topic Model

- $\theta, D \times K$ document-topic matrix $\leftarrow$ logistic normal glm with covariates
  - Covariate-specific prior with global topic covariance
  - $\theta_{i,.} \sim \text{LogisticNormal}(X_i \gamma, \Sigma)$
- $\beta, K \times V$ topic-word matrix $\leftarrow$ multinomial logit with covariates

Each token has a topic drawn from the document mixture
- Draw token topic $z_{i,m}$ from $\text{Multinomial}(\theta_i)$
- Draw observed word $w_{i,m}$ from $\text{Multinomial}(\beta_{k=z,})$
The Structural Topic Model

- $\theta$, $D \times K$ document-topic matrix $\leftarrow$ logistic normal glm with covariates
  - Covariate-specific prior with global topic covariance
  - $\theta_{i, \cdot} \sim \text{LogisticNormal}(X_i \gamma, \Sigma)$

- $\beta$, $K \times V$ topic-word matrix $\leftarrow$ multinomial logit with covariates
  - Each topic is now a covariate-specific deviation from a baseline distribution.
  - $\vec{\beta}_{k, \cdot} \propto \exp(m + \kappa^{(\text{topic})} + \kappa^{(\text{cov})} + \kappa^{(\text{int})})$
  - Three parts: topic, covariate, topic-covariate interaction

- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{i,m}$ from Multinomial($\theta_i$)
  - Draw observed word $w_{i,m}$ from Multinomial($\beta_{k=z,}$)
When you think about immigration, what makes you worried?...

When you think about immigration, what do you think of?...

Original analysis:

Human coders using pre-established coding categories (Fear, Anger, Enthusiasm)

Treatment had impact on Fear and Anger.
Treatment/Control:
Albertson and Gadarian: Anxiety and Immigration

Treatment/Control:

- “... When you think about immigration, what makes you worried?...”
Albertson and Gadarian: Anxiety and Immigration

Treatment/Control:

- “... When you think about immigration, what makes you worried?...”
- “... When you think about immigration, what do you think of?...”
Treatment/Control:

- “... When you think about immigration, what makes you worried?...”
- “... When you think about immigration, what do you think of?...”
Albertson and Gadarian: Anxiety and Immigration

Treatment/Control:

- “... When you think about immigration, what makes you worried?...”
- “... When you think about immigration, what do you think of?...”

Original analysis:
Albertson and Gadarian: Anxiety and Immigration

Treatment/Control:

- “... When you think about immigration, what makes you worried?...”
- “... When you think about immigration, what do you think of?...”

Original analysis:

- Human coders using pre-established coding categories (Fear, Anger, Enthusiasm)
Treatment/Control:

- “... When you think about immigration, what makes you worried?...”
- “... When you think about immigration, what do you think of?...”

Original analysis:

- Human coders using pre-established coding categories (Fear, Anger, Enthusiasm)
- Treatment had impact on Fear and Anger.
Topics

- **Topic 1**

  - "problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals."
  - "crime lost jobs benefits paid to illegals health care and food....we cannot feed the world when we have americans starving, etc"

- **Topic 2**

  - "i worry about the republican party doing something very stupid. this country was built on immigration, to deny anyone access to citizenship is unconstitutional. what happened to give me your poor, sick, and tired?"
  - "border control, certain illegal immigrants tolerated, and others immediately deported."
Topics

- Topic 1

• "problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals."
• "crime lost jobs benefits paid to illegals health care and food....we cannot feed the world when we have americans starving, etc"

- Topic 2

• "i worry about the republican party doing something very stupid. this country was built on immigration, to deny anyone access to citizenship is unconstitutional. what happened to give me your poor, sick, and tired?"
• "border control, certain illegal immigrants tolerated, and others immediately deported."
Topics

- Topic 1
  - illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag

- Topic 2
  - immigr, illeg, legal, border, need, worri, mexico, think, countri, law, mexican, make, america, worker

"problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals."

"crime lost jobs benefits paid to illegals health care and food....we cannot feed the world when we have americans starving, etc"

"i worry about the republican party doing something very stupid. this country was built on immigration, to deny anyone access to citizenship is unconstitutional. what happened to give me your poor, sick, and tired?"

"border control, certain illegal immigrants tolerated, and others immediately deported."
Topics

- Topic 1
  - illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag
  - “problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals.”

- Topic 2
  - immigr, illeg, legal, border, need, worri, mexico, think, countri, law, mexican, make, america, worker
  - “I worry about the Republican party doing something very stupid. This country was built on immigration, to deny anyone access to citizenship is unconstitutional. What happened to give me your poor, sick, and tired?”
  - “border control, certain illegal immigrants tolerated, and others immediately deported.”
Topic 1

- illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag
- “problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals.”
- “crime lost jobs benefits paid to illegals health care and food....we cannot feed the world when we have americans starving, etc”
Topics

Topic 1

- illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag
- “problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals.”
- “crime lost jobs benefits paid to illegals health care and food....we cannot feed the world when we have americans starving, etc”
Topics

- **Topic 1**
  - illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag
  - “problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals.”
  - “crime lost jobs benefits paid to illegals health care and food....we cannot feed the world when we have americans starving, etc”

- **Topic 2**
Topics

- **Topic 1**
  - illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag
  - “problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals.”
  - “crime lost jobs benefits paid to illegals health care and food....we cannot feed the world when we have americans starving, etc”

- **Topic 2**
  - immigr, illeg, legal, border, need, worri, mexico, think, countri, law, mexican, make, america, worker
Topics

• Topic 1
  ▶ illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag
  ▶ “problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals.”
  ▶ “crime lost jobs benefits paid to illegals health care and food….we cannot feed the world when we have americans starving, etc”

• Topic 2
  ▶ immigr, illeg, legal, border, need, worri, mexico, think, countri, law, mexican, make, america, worker
  ▶ “i worry about the republican party doing something very stupid. this country was built on immigration, to deny anyone access to citizenship is unconstitutional. what happened to give me your poor, sick, and tired?”
**Topics**

- **Topic 1**
  - illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag
  - “problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of education and the quality of care in hospitals.”
  - “crime lost jobs benefits paid to illegals health care and food....we cannot feed the world when we have americans starving, etc”

- **Topic 2**
  - immigr, illeg, legal, border, need, worri, mexico, think, countri, law, mexican, make, america, worker
  - “i worry about the republican party doing something very stupid. this country was built on immigration, to deny anyone access to citizenship is unconstitutional. what happened to give me your poor, sick, and tired?”
  - “border control, certain illegal immigrants tolerated, and others immediately deported.”
Effects on Topic 1

Figure: Topic 1.
Different Newspapers, Different Perspectives
(Roberts, Stewart, Airoldi 2017)
Different Newspapers, Different Perspectives

![Graph showing mean topic proportion in corpus for different newspapers: XIN, BBC, JEN, AP, AFP.](image)
Fatwas (Lucas et al 2015 and Nielsen 2014)

fatwas: Islamic legal rulings on any virtually any aspect of human behavior, ranging from sex and dietary restrictions to violent Jihad.

We combine expert assessments 33 clerics (20 Jihadists and 13 non-Jihadists) with their Fatwas, giving us 11,045 texts.

Estimate STM with Jihadi vs. non-Jihadi classification as a topic prevalence parameter.
<table>
<thead>
<tr>
<th>Topic</th>
<th>Terms</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fighting</td>
<td>Muslim, Jihad, Islam, fight, Jihad fighters, pathway, allmighty, that</td>
<td>FreEx: jihad, fighting, jihadist fighters, pulpil, approves of us, annotated, to fight, vicinity</td>
</tr>
<tr>
<td>Social theory</td>
<td>person, life, soul/self, knowledge/science, society, work, image, material/physical</td>
<td>FreEx: imagine, morals, develop, society, product, necessarily, environment, traditions, activity</td>
</tr>
<tr>
<td>Politics</td>
<td>Arab, Jews, country, Islam, A.D., year, West, Muslim</td>
<td>FreEx: capitol, Asia, Iran, South, Washington, A.D., Russia, Turkey</td>
</tr>
<tr>
<td>The Prophet</td>
<td>said, prayers (be upon him), peace (be upon him), allmighty, messenger, glory, prophet, that</td>
<td>FreEx: allmighty, allmighty, glory, bless you, magic, punishment, hypocrisy, sins</td>
</tr>
<tr>
<td>Prayer</td>
<td>prayer, pray, son, prophet, sheikh, mosque, fatwah, group</td>
<td>FreEx: prostration, prostrated, Abd al-Aziz, supplicant, Baz, prayer space, omission, prostration</td>
</tr>
<tr>
<td>Ramadan</td>
<td>day, fasting, Ashura, Ramadan, sheikh, group, fatwas, Uthaymeen</td>
<td>FreEx: wash, one who fasts, fasting, fasting, to break fast, Ramadan, travel, dirty</td>
</tr>
<tr>
<td>Family and Women</td>
<td>woman, O, man, girl, one, says, men, people</td>
<td>FreEx: veil, youth, (sheikh) Tamim, Azzam, tanks, finery, wear, (typo)</td>
</tr>
<tr>
<td>Money, Pilgrimage, and Marriage</td>
<td>tithing, money, pilgrimage, permitted, religion, marriage, believe/verify, divorce</td>
<td>FreEx: tithing, divorce, banks, divorce, card, banks, to perform pilgrimage, poor</td>
</tr>
<tr>
<td>Islam and Modernity</td>
<td>Islam, land, mankind, people, religion, life, other, God</td>
<td>FreEx: Europe, civilization, European, mankind, church, goods, generations, their lives</td>
</tr>
<tr>
<td>Hadith</td>
<td>Saying, hadith, said, prayers (be upon him), peace (be upon him), Muslim, legally, not</td>
<td>FreEx: to forbid, analogy, permission, general, evidence, forbid, text, absolutely</td>
</tr>
<tr>
<td>Excommunication</td>
<td>Apostasy, said, allmighty, polytheism, Islam, Apostate, saying, people</td>
<td>FreEx: excommunicate, apostate, apostasy, sponsorship, idolatry, excommunication, idols, to make permissible</td>
</tr>
<tr>
<td>Salafism</td>
<td>Sunna, sheikh, son, people, book, knowledge, Salafi, Muhammad</td>
<td>FreEx: heterodoxy, innovator, Sufi, Salafi, to draw near to, distinguish, (the) saved (group), to undertake</td>
</tr>
<tr>
<td>Shari'a and Law</td>
<td>Islam, wisdom, right, people, thing, legally, Shari'a, religion</td>
<td>FreEx: Shari'a, to legislate, to send down, to judge, judgment, justice, parliament, court</td>
</tr>
</tbody>
</table>
100 Topics Occurring in "Normal" Fatwas (Jihad Score < 0 )

**Favorite Jihadi Topics**
- Shariah
- The Prophet
- Ibn Taymiyya
- Ablutions
- Money
- Prayer
- Permissibility
- Heaven and Hell
- Hajj
- Duty

**Evenly Split Topics**
- Sin
- Sheikh Uthaymeen
- God's Oneness
- Quran
- Knowledge
- Apostasy
- Quran
- Ulama
- Heaven and Earth
- Knowledge

**Favorite Non-Jihadi Topics**
- Hadeeth
- Dating
- Zakat
- Surahs and Verses
- Hadeeth
- Ramadan Fasting
- Hadeeth
- Divorce, Marriage, Sex
- Fatwa Greeting Formula

**Figure:** Estimated topic proportions by fighting the west and excommunication topics, separated out by jihadist versus jihadist coding.
Figure: The network of correlated topics for a 15-topic Structural Topic Model with Jihadi/not-Jihadi as the predictor of topics in Arab Muslim cleric writings.
Ingest
Process
Estimate
textProcessor
readCorpus
{ prepDocuments
plotRemoved{
stm
{
Evaluate Understand ... orr
labelTopics
cloud
plotQuote
plot.estimateEffect
plot.STM
plot.topicCorr
Extensions
stmBrowser
stmCorrViz
...
STM Package in R

1. Many functions for reading in texts and manipulating the corpus.
stm Package in R

1. Many functions for reading in texts and manipulating the corpus.
2. Simple GLM style syntax for the model using formulas

```r
mod.out <- stm(documents,vocab, K=10,
               prevalence= ~treatment,
               content= ~gender,
               data=metadata)
```

Simple syntax for including smooth functional forms for continuous variables via `s()`. Wrappers to automate model selection.

Available at structuraltopicmodel.com – example data/code: https://goo.gl/j6T42I
stm Package in R

1. Many functions for reading in texts and manipulating the corpus.

2. Simple GLM style syntax for the model using formulas

```r
mod.out <- stm(documents,vocab, K=10,
prevalence= ~treatment,
content= ~gender,
data=metadata)
```

3. Simple syntax for including smooth functional forms for continuous variables via `s()`.
stm Package in R

1. Many functions for reading in texts and manipulating the corpus.
2. Simple GLM style syntax for the model using formulas
   
   ```r
   mod.out <- stm(documents, vocab, K=10,
                   prevalence= ~treatment,
                   content= ~gender,
                   data=metadata)
   ```

3. Simple syntax for including smooth functional forms for continuous variables via `s()`.
4. Wrappers to automate model selection.
Many functions for reading in texts and manipulating the corpus.

Simple GLM style syntax for the model using formulas

```
mod.out <- stm(documents,vocab, K=10,
               prevalence= ~treatment,
               content= ~gender,
               data=metadata)
```

Simple syntax for including smooth functional forms for continuous variables via `s()`.

Wrappers to automate model selection.

Available at structuraltopicmodel.com – example data/code: https://goo.gl/j6T42I
Lots of quantities of interest

1. Label topics (4 styles of most informative words) (summary, labelTopics)
2. Plot predicted topic/covariate relationships and CI’s with uncertainty (plot)
3. Documents highly associated with particular topics (findThoughts)
Lots of quantities of interest

1. Label topics (4 styles of most informative words) (*summary*, *labelTopics*)
2. Plot predicted topic/covariate relationships and CI’s with uncertainty (*plot*)
3. Documents highly associated with particular topics (*findThoughts*)
New Functionality: stmBrowser

http://pages.ucsd.edu/~meroberts/stm-online-example/index.html