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.

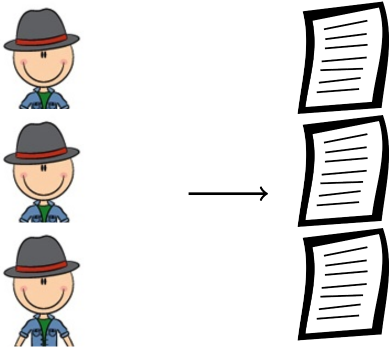
What this means in pictures

Say you have
a lot of people.

What this means in pictures

Say you have
a lot of people.

Each writes
some texts

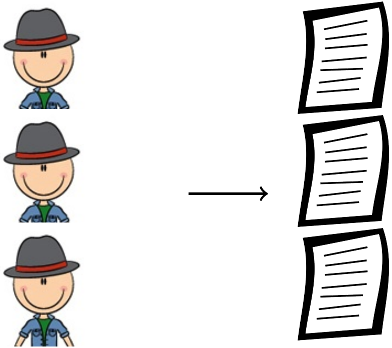


What this means in pictures

Say you have
a lot of people.

Each writes
some texts

that discuss a few
different topics

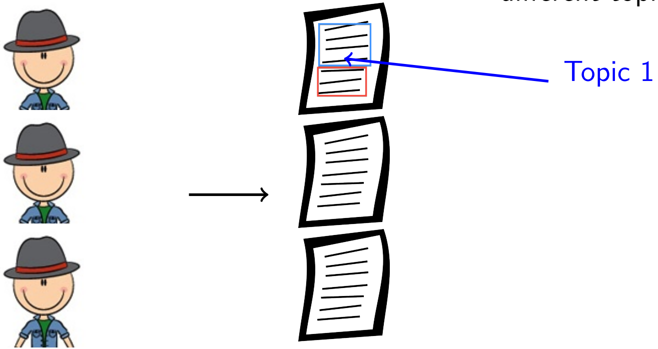


What this means in pictures

Say you have
a lot of people.

Each writes
some texts

that discuss a few
different topics

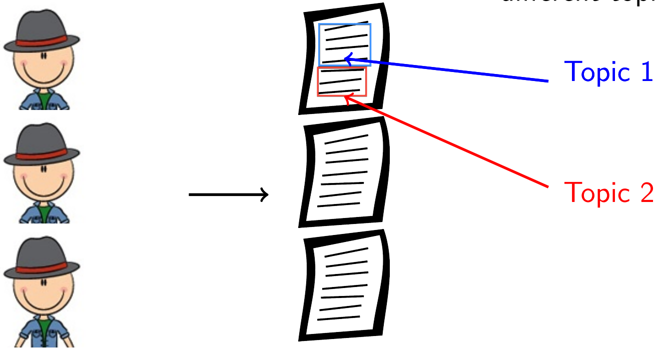


What this means in pictures

Say you have
a lot of people.

Each writes
some texts

that discuss a few
different topics

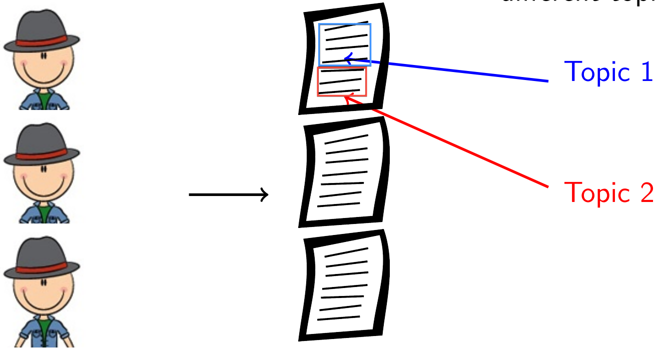


What this means in pictures

Say you have
a lot of people.

Each writes
some texts

that discuss a few
different topics



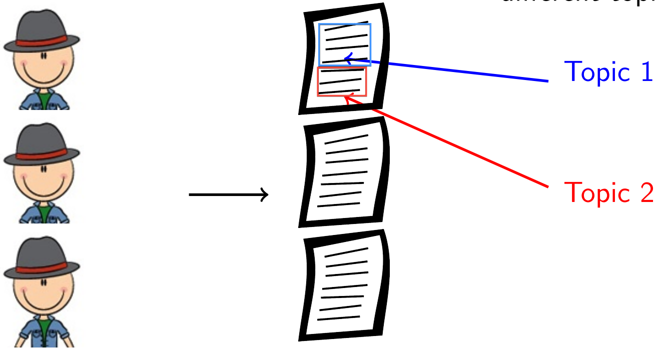
The Latent Dirichlet Allocation estimates:

What this means in pictures

Say you have
a lot of people.

Each writes
some texts

that discuss a few
different topics



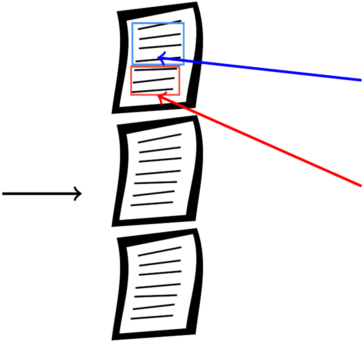
The Latent Dirichlet Allocation estimates:

1. The topics- each is a distribution over words

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

The Latent Dirichlet Allocation estimates:

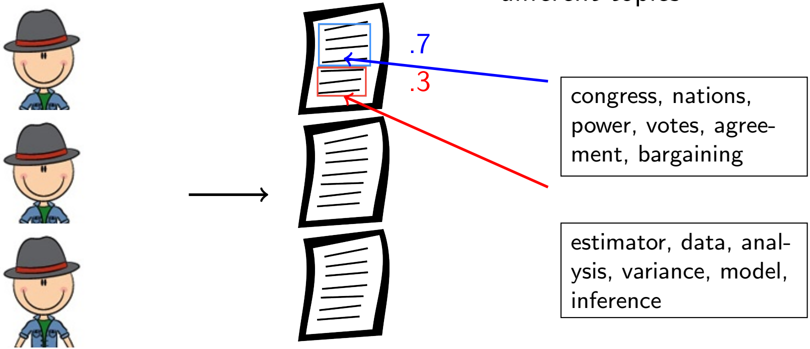
1. The topics- each is a distribution over words

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

Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1	
	Cluster 1
Doc 2	
	Cluster 2
Doc 3	
	$\vdots$
$\vdots$	
	Cluster K
Doc N	

Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1

Doc 2

Doc 3

$\vdots$

Doc N

Cluster 1

Cluster 2

$\vdots$

Cluster K

Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1

Doc 2

Doc 3

$\vdots$

Doc N

Cluster 1

Cluster 2

$\vdots$

Cluster K

Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1

Cluster 1

Doc 2

Cluster 2

Doc 3

$\vdots$

Cluster K

Doc N

Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1

Doc 2

Doc 3

$\vdots$

Doc N

Cluster 1

Cluster 2

$\vdots$

Cluster K

Topic and Mixed Membership Models

Topic Models (Mixed Membership)

Document $\rightsquigarrow$ Many clusters

Doc 1	
	Cluster 1
Doc 2	
	Cluster 2
Doc 3	
	$\vdots$
$\vdots$	
	Cluster K
Doc N	

Topic and Mixed Membership Models

Topic Models (Mixed Membership)

Document $\rightsquigarrow$ Many clusters

Doc 1 → Cluster 1, Cluster 2, Cluster K

Doc 2

Doc 3

$\vdots$

Doc N

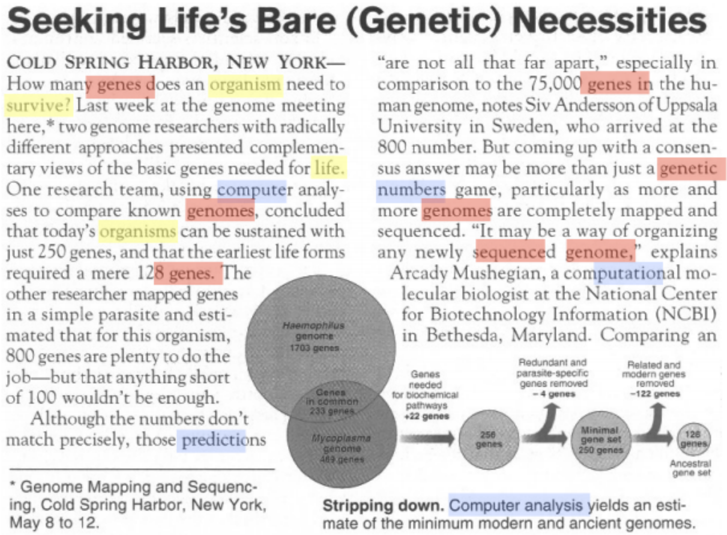
Cluster 1

Cluster 2

$\vdots$

Cluster K

A Statistical Highlighter (With Many Colors)



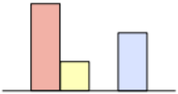
Seeking Life's Bare (Genetic) Necessities

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.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.



Topic Models

Topic Models

Two primary matrices of interest:

Topic Models

Two primary matrices of interest:

1) Topical Prevalence Matrix (DxK)

Topic Models

Two primary matrices of interest:

1) Topical Prevalence Matrix (DxK)

$$\theta = \left[\begin{array}{c|cccc} & Topic1 & Topic2 & \dots & TopicK \\ \hline Doc1 & .2 & .1 & \dots & 0.05 \\ Doc2 & .2 & .1 & \dots & .3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ DocD & 0 & 0 & \dots & .5 \end{array}\right]$$

Topic Models

Two primary matrices of interest:

1) Topical Prevalence Matrix (DxK)

$$\theta = \left[\begin{array}{c|cccc} & Topic1 & Topic2 & \dots & TopicK \\ \hline Doc1 & .2 & .1 & \dots & 0.05 \\ Doc2 & .2 & .1 & \dots & .3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ DocD & 0 & 0 & \dots & .5 \end{array}\right]$$

2) Topical Content Matrix (VxK)

Topic Models

Two primary matrices of interest: $X \approx \theta\beta$

1) Topical Prevalence Matrix (DxK)

$$\boldsymbol{\theta} = \left[\begin{array}{c|cccc} & Topic1 & Topic2 & \dots & TopicK \\ \hline Doc1 & .2 & .1 & \dots & 0.05 \\ Doc2 & .2 & .1 & \dots & .3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ DocD & 0 & 0 & \dots & .5 \end{array}\right]$$

2) Topical Content Matrix (VxK)

$$\boldsymbol{\beta^T} = \left[\begin{array}{c|cccc} & Topic1 & Topic2 & \dots & TopicK \\ \hline \text{``}text\text{''} & .02 & .001 & \dots & 0.001 \\ \text{``}data\text{''} & .001 & .02 & \dots & 0.001 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{``}analysis\text{''} & .01 & .01 & \dots & 0.0005 \end{array}\right]$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

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

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

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

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

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

$$\boldsymbol{\theta}_i | \boldsymbol{\alpha} \quad \sim \quad \text{Dirichlet}(\boldsymbol{\alpha})$$

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

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

$$\begin{aligned} \boldsymbol{\theta}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{z}_{im} | \boldsymbol{\theta}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}_i) \end{aligned}$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

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

$$
\begin{aligned}
\boldsymbol{\theta}_i | \boldsymbol{\alpha} & \sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{z}_{im} | \boldsymbol{\theta}_i & \sim \text{Multinomial}(1, \boldsymbol{\theta}_i) \\
x_{im} | \boldsymbol{\beta}_k, z_{imk} = 1 & \sim \text{Multinomial}(1, \boldsymbol{\beta}_k)
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

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

$$\boldsymbol{\beta}_k \sim \text{Dirichlet}(\boldsymbol{\eta})$$

$$\begin{aligned} \boldsymbol{\theta}_i|\boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{z}_{im}|\boldsymbol{\theta}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}_i) \\ x_{im}|\boldsymbol{\beta}_k, z_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\beta}_k) \end{aligned}$$

Optimization:

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

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

$$
\begin{aligned}
\boldsymbol{\beta}_k &\sim \text{Dirichlet}(\boldsymbol{\eta}) \\
\alpha_k &\sim \text{Gamma}(\alpha, \beta) \\
\boldsymbol{\theta}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{z}_{im} | \boldsymbol{\theta}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}_i) \\
x_{im} | \boldsymbol{\beta}_k, z_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\beta}_k)
\end{aligned}
$$

Optimization:

- Variational Approximation $\rightsquigarrow$ Find "closest" distribution

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

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

$$
\begin{aligned}
\boldsymbol{\beta}_k &\sim \text{Dirichlet}(\boldsymbol{\eta}) \\
\alpha_k &\sim \text{Gamma}(\alpha, \beta) \\
\boldsymbol{\theta}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{z}_{im} | \boldsymbol{\theta}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}_i) \\
x_{im} | \boldsymbol{\beta}_k, z_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\beta}_k)
\end{aligned}
$$

Optimization:

- Variational Approximation $\rightsquigarrow$ Find "closest" distribution
- Gibbs sampling $\rightsquigarrow$ 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?

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

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 $\rightarrow$ district level

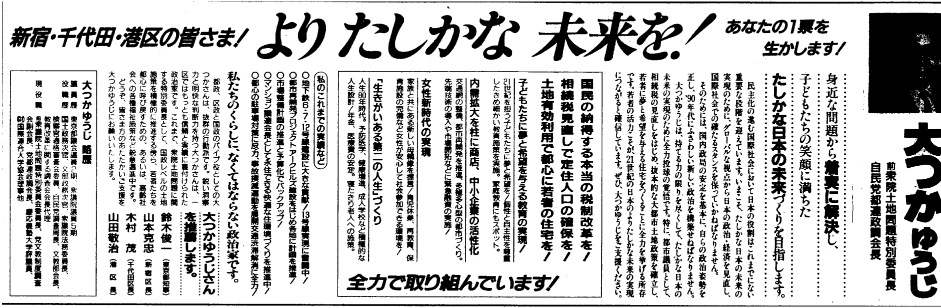
Example: Japanese Campaign Manifestos (Catalinac 2016)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet

Example: Japanese Campaign Manifestos (Catalinac 2016)

Typical Manifesto:



新宿・千代田・港区の皆さま！　よりたしかな未来を！　あなたの1票を生かします！

大つかゆうじ

前衆院土地問題特別委員長
自民党都連政調会長

身近な問題から着実に解決し、
子どもたちの笑顔に満ちた
たしかな日本の未来づくりを目指します。

民主化の進む国際社会において日本の役割はこれまでにない重要な段階を迎えています。いまこそ、たしかな日本の未来の実現のために、グローバルな視点から日本の政治・経済を見直し、国際社会での確たるスタンスを保っていかねばなりません。

そのためには、国内政局の安定を基本に、自らの政治姿勢を正し、'90年代にふさわしい新しい政治を構築せねばなりません。

大つかゆうじは、持てる力の限りを尽くして、たしかな日本の未来の実現のために全力投球の覚悟です。特に、都市議員として、相続税の見直しをはじめ、抜本的な大都市土地政策を確立し、若者に夢と希望を与える住宅をつくることに全力を挙げる所存です。若者の活力こそが、21世紀の日本のよりたしかな未来の実現につながると確信しています。ぜひ、大つかゆうじをご支援ください。

国民の納得する本当の税制改革を！

相続税見直しで定住人口の確保を！

土地有効利用で都心に若者の住宅を！

子どもたちに夢と希望を与える教育の実現！

21世紀を担う子どもたちに夢と希望を／個性と自主性を尊重したきめ細かい教育施策を実施。家庭教育にもスポット。

内需拡大を柱に商店、中小企業の活性化

交通網の整備、都市再開発を推進、多極多心型の都市づくり。先端技術の導入や市場開拓などに緊急融資の実施／

女性新時代の実現

家族と共にある新しい母親像を提言／育児休業、再教育、保育施設の完備など女性が安心して社会参加できる環境を／

「生きがいある第二の人生」づくり

人生80年時代。予防医学、健康増進、成人学校など積極的な人生設計／年金、医療費の安定。寝たきり老人への施策。

全力で取り組んでいます！

私のこれまでの実績など

○地下鉄6・7・12号線建設に大きな貢献／13号線実現に奮闘中／
○都市再開発プロジェクトアークヒルズ建設をはじめ各地に計画を推進／
○市場整備計画の推進と予算化でバックアップ／
○マンション議連会長として永住できる快適な住環境づくりを推進中／
○都心の駐車場対策に尽力、事故撲滅運動を展開交通渋滞解消に全力／

私たちのくらしに、なくてはならない政治家です。

都政、区政と国政のパイプ役としての大つかさんは、抜群の行動派です。鋭い洞察力と明快な判断力を持つ大つかさんは、一区ではもっとも信頼と実績に裏付けられた政治家です。これまで、衆院土地問題に関する特別委員長として、国政レベルの土地施策を積極的に推進する傍ら、若者たちを都心に呼び戻すための、あるいは、高齢社会への各種福祉施策など鋭意邁進中です。

どうぞ、皆さま方のあたたかいご支援を大つかゆうじにお願いいたします。

大つかゆうじさんを推薦します。

鈴木俊一（東京都知事）
山本克忠（新宿区長）
木村　茂（千代田区長）
山田敬治（港区長）

大つかゆうじ　略歴

議員歴　東京都議会議員2期、衆議院議員5期
役職歴　国土政務次官、文部政務次官、衆議院法務委員長、検察官適格審査会委員、自民党調査局長、文教部会長、教育改革に関する調査会会長代理
現役職　前衆議院土地問題特別委員会委員長、党文教制度調査会副会長、党都連政調会長、慶応義塾大学評議員、(財)国際連合大学協会理事他

Example: Japanese Campaign Manifestos (Catalinac 2016)

Japanese Elections:

- Election Administration Commission runs elections $\rightarrow$ district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009

Example: Japanese Campaign Manifestos (Catalinac 2016)

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

Example: Japanese Campaign Manifestos (Catalinac 2016)

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

Example: Japanese Campaign Manifestos (Catalinac 2016)

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

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
  - 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

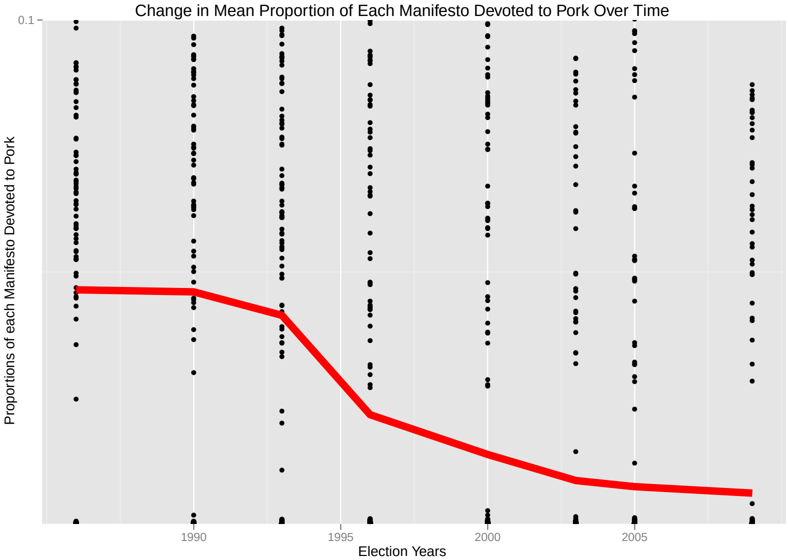
Example: Japanese Campaign Manifestos (Catalinac 2016)

- Applies Vanilla LDA
- Output: topics (with Japanese characters)

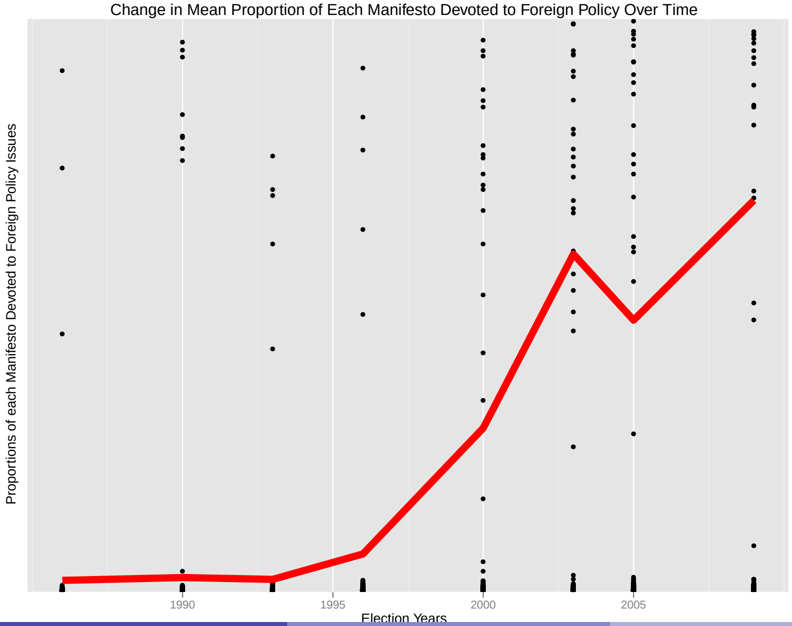
Example: Japanese Campaign Manifestos (Catalinac 2016)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
改革	年金	推進	区	政治	日本
郵政	円	整備	政策	改革	国
民営	廃止	図る	地域	国民	外交
小泉	改革	つとめる	まち	企業	国家
構造	兆	社会	鹿児島	自民党	社会
政府	実現	対策	全力	日本	国民
官	無駄	振興	選挙	共産党	保障
推進	日本	充実	国政	献金	安全
民	増税	促進	作り	金権	地域
自民党	削減	安定	横浜	党	拉致
日本	一元化	確立	対策	選挙	経済
制度	政権	企業	中小	禁止	守る
民間	子供	実現	発電	憲法	問題
年金	地域	中小	推進	腐敗	北朝鮮
実現	ひと	育成	エネルギー	団体	教育
進める	サラリーマン	制度	企業	区	責任
断行	制度	政治	声	ソ連	力
地方	議員	地域	実現	守る	創る
止める	金	福祉	活性	平和	安心
保障	民主党	事業	自民党	円	目指す
財政	年間	改革	地方	反対	誇り
作る	一掃	確保	尽くす	真	憲法
賛成	郵政	強化	商店	是正	可能
社会	道路	教育	いかす	一掃	道
国民	交代	施設	全国	悪政	未来
公務員	社会保険庁	生活	政党	抜本	ひと
力	月額	支援	ひと	定数	再生
経済	手当	環境	支援	政党	将来
国	談合	発展	経済	金丸	解決
安心	支援	施策	福祉	改悪	基本
Postal privatization	Reducing Wasteful Public Spending	Pork for the District	Policies for the district	Political Reform	National Security Policy

Example: Japanese Campaign Manifestos (Catalinac 2011)



Example: Japanese Campaign Manifestos (Catalinac 2011)



Measuring Topic Performance: Out of Sample Prediction

How well does our model perform?

Measuring Topic Performance: Out of Sample Prediction

How well does our model perform? $\rightsquigarrow$ predict new documents?

Measuring Topic Performance: Out of Sample Prediction

How well does our model perform? $\rightsquigarrow$ predict new documents?
Problem

Measuring Topic Performance: Out of Sample Prediction

How well does our model perform?$\rightsquigarrow$ predict new documents?
Problem$\rightsquigarrow$ in sample evaluation leads to overfit.

Measuring Topic Performance: Out of Sample Prediction

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

Measuring Topic Performance: Out of Sample Prediction

How well does our model perform? $\rightsquigarrow$ predict new documents?
Problem $\rightsquigarrow$ in sample evaluation leads to overfit.
Solution $\rightsquigarrow$ evaluate performance on held out data
For held out document $\boldsymbol{x}^*_{\text{out}}$

Measuring Topic Performance: Out of Sample Prediction

How well does our model perform? $\rightsquigarrow$ predict new documents?
Problem $\rightsquigarrow$ in sample evaluation leads to overfit.
Solution $\rightsquigarrow$ evaluate performance on held out data
For held out document $\boldsymbol{x}^{*}_{\text{out}}$

$$\text{Perplexity} \quad = \quad \exp\left(-\log p(\boldsymbol{x}^{*}_{\text{out}} | \boldsymbol{\mu}, \boldsymbol{\pi})\right)$$

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task

What's Prediction Got to Do With It?

- Prediction$\rightsquigarrow$ One Task
- Do we care about it?

What's Prediction Got to Do With It?

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

What's Prediction Got to Do With It?

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

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ 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 $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ 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 $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ 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 $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ 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 $\rightsquigarrow$ measure quality in topics and clusters

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ 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 $\rightsquigarrow$ measure quality in topics and clusters

- Statistics: measure cohesiveness and exclusivity (Roberts, et al 2014)

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ 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 $\rightsquigarrow$ measure quality in topics and clusters

- Statistics: measure cohesiveness and exclusivity (Roberts, et al 2014)
- Experiments: measure topic and cluster quality

Measuring Cohesiveness and Exclusivity

Measuring Cohesiveness and Exclusivity

- Consider the output of a topic model)

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

Topic 1	bill	congressman	earmarks	following	house

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

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

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? $\rightsquigarrow$ 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

Topic 1	bill	congressman	earmarks	following	house
Topic 2	immigration	reform	security	border	worker
Topic 3	earmark	egregious	pork	fiscal	today

- An ideal topic? $\rightsquigarrow$ will see these words co-occur in documents
- Define $\boldsymbol{v}_k = (v_{1k}, v_{2k}, \ldots, v_{Lk})$ 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

Topic 1	bill	congressman	earmarks	following	house
Topic 2	immigration	reform	security	border	worker
Topic 3	earmark	egregious	pork	fiscal	today

- An ideal topic? $\rightsquigarrow$ will see these words co-occur in documents
- Define $\boldsymbol{v}_k = (v_{1k}, v_{2k}, \ldots, v_{Lk})$ be the top words for a topic
- For example $\boldsymbol{v}_3 =$ (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:

Measuring Cohesiveness and Exclusivity

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

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

Measuring Cohesiveness and Exclusivity

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

$$\begin{aligned} D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\ D(\text{egregious}) &= \text{Number of times Egregious occurs} \end{aligned}$$

Measuring Cohesiveness and Exclusivity

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

$$
\begin{aligned}
D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\
D(\text{egregious}) &= \text{Number of times Egregious occurs}
\end{aligned}
$$

Define cohesiveness for topic k as

Measuring Cohesiveness and Exclusivity

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

$$\begin{aligned} D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\ D(\text{egregious}) &= \text{Number of times Egregious occurs} \end{aligned}$$

Define cohesiveness for topic k as

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

Measuring Cohesiveness and Exclusivity

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

$$\begin{aligned} D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\ D(\text{egregious}) &= \text{Number of times Egregious occurs} \end{aligned}$$

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:

Measuring Cohesiveness and Exclusivity

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

$$\begin{aligned} D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\ D(\text{egregious}) &= \text{Number of times Egregious occurs} \end{aligned}$$

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:

$$\text{Cohesive} = (\sum_{k=1}^{K} \text{Cohesive}_k)/K$$

Measuring Cohesiveness and Exclusivity

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

$$\begin{aligned} D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\ D(\text{egregious}) &= \text{Number of times Egregious occurs} \end{aligned}$$

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:

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

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive $\rightsquigarrow$ few replicates of each topic

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive $\rightsquigarrow$ 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 $\rightsquigarrow$ 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:

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive $\rightsquigarrow$ 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:

$$\text{Exclusivity}_k = \sum_{j:v_j \in \boldsymbol{v}_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}}$$

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive$\rightsquigarrow$ 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:

$$\text{Exclusivity}_k = \sum_{j:v_j \in \boldsymbol{v}_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}}$$

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

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive $\rightsquigarrow$ 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:

$$\text{Exclusivity}_k = \sum_{j: v_j \in \boldsymbol{v}_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}}$$

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

$$= \left(\sum_{k=1}^{K} \sum_{j: v_j \in \boldsymbol{v}_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}} \right) / K$$

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 $\rightsquigarrow$ combination

Examples of Topic Models

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

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

Leveraging Information Within and About Texts

- 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

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

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

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

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

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 = LDA + Contextual Information

- STM provides two ways to include contextual information

STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence 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

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

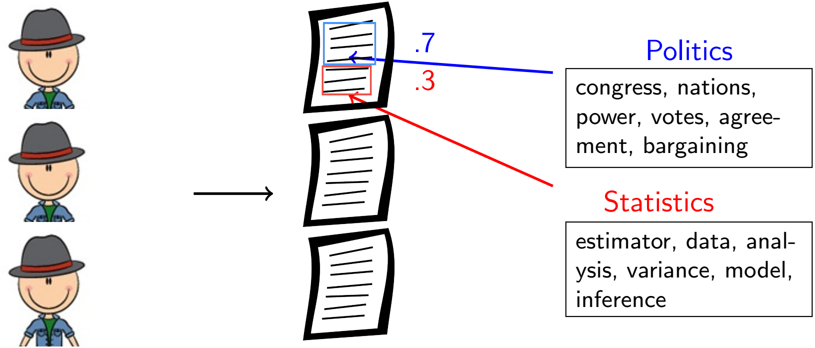
- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - e.g. Democrats talk more about education than Republicans
  - Topic content can vary by metadata
    - e.g. Democrats are less likely to use the word “life” when talking about abortion than Republicans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



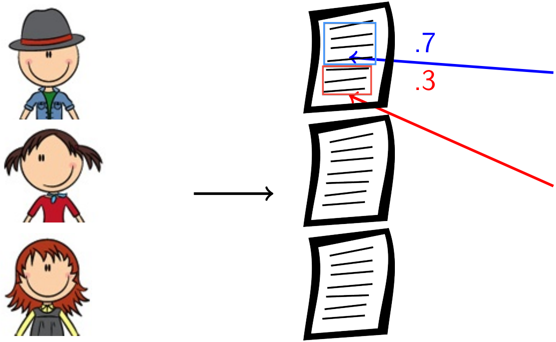
The STM Allows for:

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:

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

.7

.3

that discuss a few
different topics

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

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.

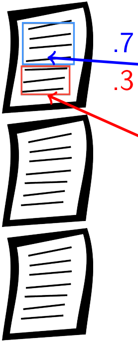
Each writes
some text

that discuss a few
different topics

Group A

Group B

Group A



Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The STM Allows for:

1. The words in each topic to vary by gender
2. The topic proportions to vary by group

STM: What this means in pictures

Say you have a lot of people.

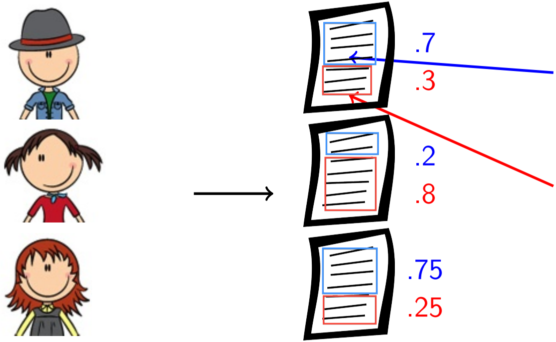
Each writes some text

that discuss a few different topics

Group A

Group B

Group A



Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

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 $\mathbf{X}$

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 $\mathbf{X}$
  - Topical content groups: D length vector $\mathbf{Y}$

Mixed-Membership Topic Models

More formal terminology:

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

Mixed-Membership Topic Models

More formal terminology:

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

The Structural Topic Model

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

- β, $K \times V$ topic-word matrix

- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{i,m}$ from Multinomial(θ_i)
  - Draw observed word $w_{i,m}$ from Multinomial($\beta_{k=z_i}$)

The Structural Topic Model

- θ, $D \times K$ document-topic matrix $\Leftarrow$ logistic normal glm with covariates

- β, $K \times V$ topic-word matrix

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

The Structural Topic Model

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

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

The Structural Topic Model

- θ, $D \times K$ document-topic matrix $\Leftarrow$ logistic normal glm with covariates
  - Covariate-specific prior with global topic covariance
  - $\theta_{i,\cdot} \sim \text{LogisticNormal}(X_i\gamma, \Sigma)$
- β, $K \times V$ 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 Multinomial(θ_i)
  - Draw observed word $w_{i,m}$ from Multinomial($\beta_{k=z,}$)

The Structural Topic Model

- θ, $D \times K$ document-topic matrix $\Leftarrow$ logistic normal glm with covariates
  - Covariate-specific prior with global topic covariance
  - $\theta_{i,\cdot} \sim \text{LogisticNormal}(X_i\gamma, \Sigma)$
- β, $K \times V$ topic-word matrix $\Leftarrow$ multinomial logit with covariates
  - Each topic is now a covariate-specific deviation from a baseline distribution.
  - $\vec{\beta}_{k,\cdot} \propto \exp\left(m + \kappa^{(\text{topic})} + \kappa^{(\text{cov})} + \kappa^{(\text{int})}\right)$
  - Thee parts: topic, covariate, topic-covariate interaction
- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{i,m}$ from Multinomial(θ_i)
  - Draw observed word $w_{i,m}$ from Multinomial($\beta_{k=z,}$)

Albertson and Gadarian: Anxiety and Immigration

Albertson and Gadarian: Anxiety and Immigration

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?..."

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?..."

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)

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 had impact on Fear and Anger.

Topics

- Topic 1

Topics

- Topic 1

Topics

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

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."

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”

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

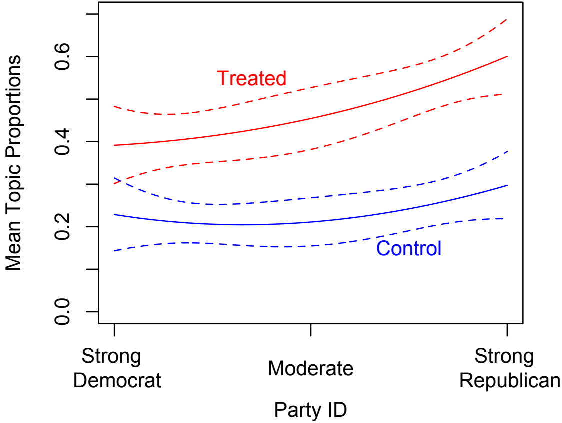
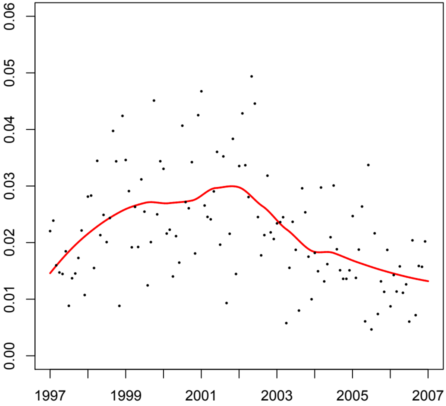


Figure: Topic 1.

Different Newspapers, Different Perspectives (Roberts, Stewart, Airoldi 2017)



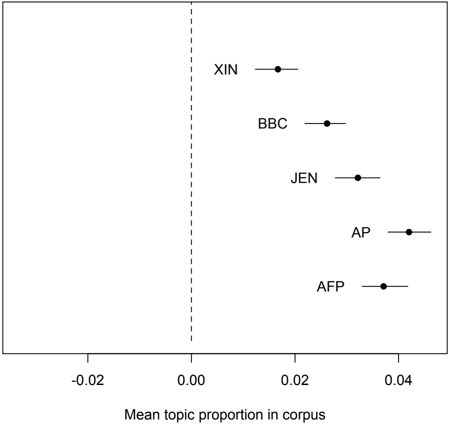
Associated Press

polic protest gong arrest
falun detain author releas
follow prison inform offic
wang crackdown activist
movement sentenc center squar
china demonstr zhang
tiananmen dissid investig

Xinhua

polic illeg smuggl public
gong investig arrest offic
crimin cult falun immigr
suspect case custom organ
author depart order china
properti told accord sentenc
terrorist

Different Newspapers, Different Perspectives



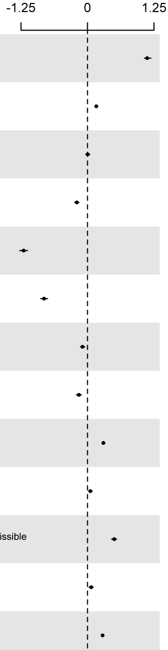
Fatwas (Lucas et al 2015 and Nielsen 2014)

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.

1.	**Fighting**	F: Muslim, Jihad, Islam, fight, Jihadi fighters, pathway, almighty, that FREX: jihad, fighting, jihadist fighters, pulpit, approves of us, annotated, to fight, vicinity F: مسلم, جهاد, اسلام, قتل, مجاهد, سبيل, تعال, ذين FREX: جهاد, قتال, مجاهد, منبر, يوافقنا, مذيل, يقاتل, بجوار
2.	**Social theory**	F: person, life, soul/self, knowledge/science, society, work, image, material/physical FREX: imagine, morals, develop, society, product,necessarily, environment, traditions, activity F: انس, حيا, نفس, علم, مجتمع, عمل, صور, ماد FREX: تصور, اخلاق, تطور, مجتمع, انتاج, حتم, بيئ, تقاليد
3.	**Politics**	F: Arab, Jews, country, Islam, A.D., year, West, Muslim FREX: capitol, Asia, Iran, South, Washington, A.D., Russia, Turkey F: عرب, يهود, دول, اسلام, م, سن, غرب, مسلم FREX: عاصمت, اسيا, اير, جنوب, اشنطن, م, روسيا, تركيا
4.	**The Prophet**	F: said, prayers (be upon him), peace (be upon him), almighty, messenger, glory, prophet, that FREX: almighty, almighty, glory, bless you, magic, punishment, hypocrisy, sins F: قال, صل, سلم, تعال, رسول, سبح, نب, ذين FREX: وجل, عز, سبح, تبارك, سحر, عذاب, رياء, ذنوب
5.	**Prayer**	F: prayer, pray, son, prophet, sheikh, mosque, fatwas, group FREX: prostration, prostrated, Abd al-Aziz, supplicant, Baz, prayer space, omission, prostration F: صلا, صل, سلم, بن, نب, شيخ, مسجد, فتاو FREX: ركع, ركعت, عبدالعزيز, ماموم, باز, مصل, سهو, ركوع
6.	**Ramadan**	F: day, fasting, Ashura, Ramadan, sheikh, group, fatwas, Uthaymeen FREX: wash, one who fasts, fasting, fasting, to break fast, Ramadan, travel, dirty F: يوم, صيام, عشر, رمض, شيخ, مجموع, فتاو, عثيم FREX: غسل, صائم, صيام, صوم, يفطر, رمض, مسافر, نجاس
7.	**Family and Women**	F: woman, O, man, girl, one, says, men, people FREX: veil, youth, (sheikh) Tamim, Azzam, tanks, finery, wear, r(typo) F: مرا, يا, رجل, نساء, احد, يقول, رجال, ناس FREX: حجاب, شاب, تميم, عزام, دباب, تبرج, لباس, ر
8.	**Money, Pilgrimage, and Marriage**	F: tithing, money, pilgrimage, permitted, religion, marriage, believe/ratify, divorce FREX: tithing, divorce, banks, divorce, card, banks, to perform pilgrimage, poor F: زكا, مال, حج, يجوز, دين, زوج, صدق, طلاق FREX: زكا, طلاق, بنك, طلق, بطاق, بنوك, يحج, فقراء
9.	**Islam and Modernity**	F: Islam, land, mankind, people, religion, life, other, God FREX: Europe, civilization, European, mankind, church, goods, generations, their lives F: اسلام, ارض, بشر, ناس, دين, حيا, اخر, ال FREX: اوربا, حضار, اورب, بشر, كنيس, متاع, اجيال, حياتهم
10.	**Hadith**	F: Saying, hadith, said, prayers (be upon him), peace (be upon him), Muslim, legally, not FREX: to forbid, analogy, permission, general, evidence, forbid, text, absolutely F: قول, حديث, قال, صل, سلم, مسلم, شرع, ليس FREX: تحريم, قياس, جواز, عموم, ادل, منع, نص, مطلقا
11.	**Excommunication**	F: Apostasy, said, almighty, polytheism, Islam, Apostate, saying, people FREX: excommunicate, apostate, apostasy, sponsorship, idolatry, excommunication, idols, to make permissible F: كفر, قال, تعال, شرك, اسلام, كافر, قول, اهل FREX: يكفر, كفرا, كفر, موالا, شرك, تكفير, اصنام, استحلال
12.	**Salafism**	F: Sunna, sheikh, son, people, book, knowledge, Salafi, Muhammad FREX: heterodoxy, innovator, Sufi, Salafi, to draw near to, distinguish, (the) saved (group), to undertake F: سن, شيخ, بن, اهل, كتاب, علم, سلف, محمد FREX: ابتداع, مبتدع, صوف, سلف, جانبوا, فتميزوا, ناج, استقاموا
13.	**Shari'a and Law**	F: Islam, wisdom, right, people, thing, legally, Shari'a, religion FREX: Shari'a, to legislate, to send down, to judge, judgment, justice, parliament, court F: اسلام, حكم, حق, ناس, امر, شرع, شريع, دين FREX: شريع, تشريع, انزل, تحاكم, تحكيم, عدل, برلم, محاكم



Jihad

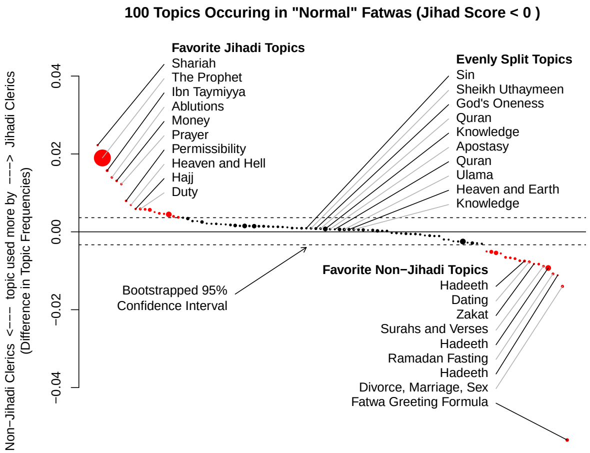


Figure: Estimated topic proportions by fighting the west and excommunication topics, separated out by jihadist versus jihadist coding.

Jihad

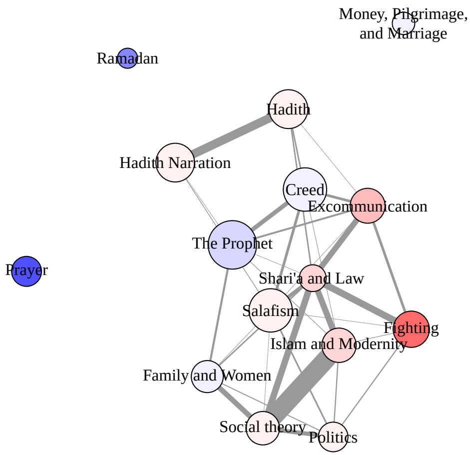
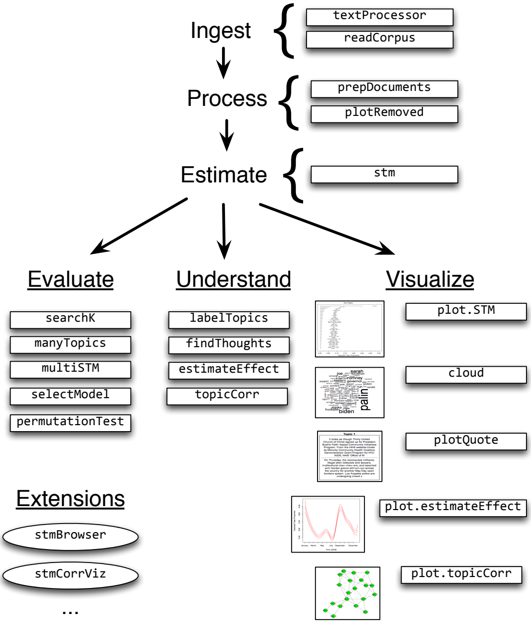


Figure: The network of correlated topics for a 15-topic Structural Topic Model with Jihadi/not-Jihadi as the predictor of topics in Arab Muslim cleric writings.

stm Packag


Ingest
textProcessor
readCorpus
Process
prepDocuments
plotRemoved
Estimate
stm
Evaluate
searchK
manyTopics
multiSTM
selectModel
permutationTest
Understand
labelTopics
findThoughts
estimateEffect
topicCorr
Visualize
plot.STM
cloud
plotQuote
plot.estimateEffect
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

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

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

```
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

```
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.

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

```
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.

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