Stony Brook University

Topic Modeling in Social Science

Yongjun Zhang, Ph.D.
Dept of Sociology and IACS
https://yongjunzhang.com

Today's Agenda

1. Guest Speaker Dr. Charlie Gomez (6:00-6:50PM)
2. Mini-Lecture on Topic Modeling (7:00-7:50 PM)
3. Lab Tutorial on Topic Modeling (8:00-8:50PM)

Guest Speaker: Dr. Charles Gomez

Dr. Gomez is an assistant professor of Sociology at the City University of New York, Queens College who uses global networks as a framework to study how knowledge and culture are shaped across borders over time.

He employs social network analysis, topic models, and traditional social science methods in his research. His work has been featured in *Nature Communications*, *Social Networks*, *Journal of Informetrics*, and *Sociological Science*.

He is also currently the P.I. of a three-year National Science Foundation (NSF) grant (2020-2023) that studies the growing stratification in national influence in global scientific research and its implications on field innovation.

Brief Summary on Last Week – Regular Expression

Regular Expression (RegEx)

"One of the unsung successes in standardization in computer science has been the regular expression (RE), a language for specifying text search strings."

(Daniel Jurafsky & James H. Martin 2019)

Letters inside square brackets []

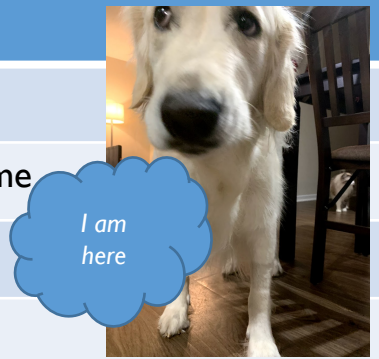
Pattern	Matches
[Bb]enson	Benson, benson
[1234567890]	Any digit

Ranges [A-Z]

Pattern	Matches	
[A-Z]	An upper case letter	Mr Benson is my fur baby.
[a-z]	A lower case letter	my cats are Mr Max and Mr Snow.
[0-9]	A single digit	My lucky number is 4

Negations [^Ss]: Carat means negation only when first in []

Pattern	Matches	
[^A-Z]	Not an upper case letter	Max is a cute boy
[^Ss]	Neither ‘S’ nor ‘s’	Snow is super handsome
[^e^]	Neither e nor ^	Benson is here
a^b	The pattern a carat b	Look up a^b now



RegEx: ? * + .

Pattern	Matches	
colou?r	Optional previous char	color colour
oo*h!	0 or more of previous char	oh! ooh! oooh! ooooh!
o+h!	1 or more of previous char	oh! ooh! oooh! ooooh!
beg.n	Match any char	begin begun begun beg3n

Anchors: ^ $

Pattern	Matches	
^[A-Z]	Starting as an Upper Case	The Begin.
^[^A-Za-z]	Starting not as a letter	2 y/o, Mr Benson
\.$	Ending with period	The end.
.$	Ending with anything	The end? The end!

Find all instances of the word "the' in a text

Pattern	Errors
the	Misses capitalized examples
[Tt]he	Incorrectly returns other or theology
[^a-zA-Z][tT]he[^a-zA-Z]	Bingo!!!
Or \b[Tt]he\b	

Greedy Match or Lazy Match

String ="I am a special member in national NGO located in international cities.

What does **^.*nat** return? What does **^.*?nat** return?

Pattern	Mathes
^.*nat	I am a special member in national NGO located in international cities.
^.*?nat	I am a special member in national NGO located in international cities.

Create a variable Capturing all National Organization for Women using RegEx

String="As the grassroots arm of the women's movement, the National Organization for Women is dedicated to its multi-issue and multi-strategy approach to women's rights, and is the largest organization of feminist grassroots activists in the United States. Natl Org for Women has hundreds of chapters and hundreds of thousands of members and activists in all 50 states and the District of Columbia. Since our founding in 1966, NOW's purpose is to take action through intersectional grassroots activism to promote feminist ideals, lead societal change, eliminate discrimination, and achieve and protect the equal rights of all women and girls in all aspects of social, political, and economic life. Now, we have a lot of members in our Nat Org for Women. NOWHERE TO GO! NOW is the only National Organization designed for Women."

Adapted From https://now.org/about/

National Organization for Women
Nat Org for Women
Natl Org for Women
NOW

RegEx:
Nat.*for Women
Nat.*?for Women
Nat.*?[gn] for Women
NOW
\bNOW\b
Nat.*?[gn] for Women|\bNOW\b

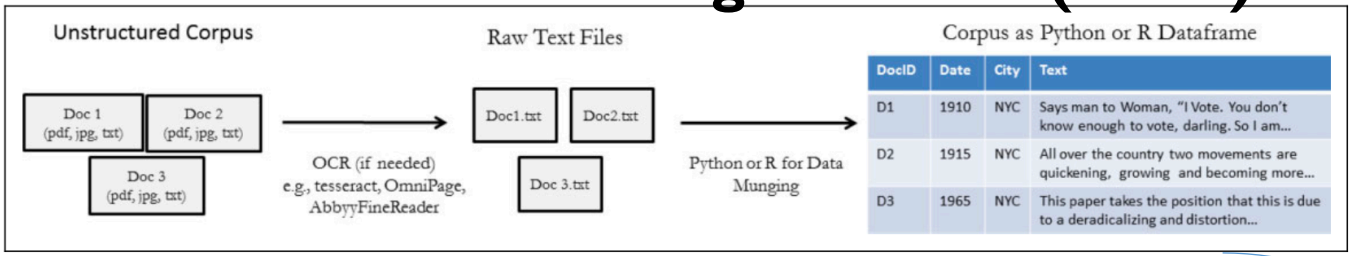
Python RE findall:
re.findall(r"(Nat.*?[gn] for Women)|(\bNOW\b)", string)

Text Normalization (Tidy Text for NLP)

Your text has to be normalized before NLPing it. The normalization process may include:

1. Tokenizing (segmenting) words – Segmenting running text into tokens (e.g., words, punctuations, numbers, etc.)
2. Normalizing word formats – putting words/tokens in a standard format (e.g. u.s.a or us, case folding, lemmatization (determining that two words have the same root), stemming (porter stemmer))
3. Segmenting sentences – for POS tag etc. or use sentence as unit of analysis

Text Transformation--Bag of Words (BoW)



DocID	Date	City	Text
D1	1910	NYC	Says man to Woman, "I Vote. You don't know enough to vote, darling. So I am...
D2	1915	NYC	All over the country two movements are quickening, growing and becoming more...
D3	1965	NYC	This paper takes the position that this is due to a deradicalizing and distortion...

Figure 1. Corpus construction: From text to dataframe. This figure demonstrates a possible path from a collection of texts, saved in separate files, to a digital dataframe suitable for further computer-assisted text analysis techniques. Often historical texts are in the form of pdf or jpg images and thus require an intervening step using optical character recognition software. More contemporary texts are already digitized. Once digitized, the researcher can use Python or R to transform the separate files into one dataframe, with metadata attached to each text (in this example, date of publication and the city in which it was published).

Note: Nelson 2017

Construct a Document-Term Matrix (sparse)

DocID	Text	Term 1 (say)	Term 3 (all)
D1	Says man…	X tf=X/N1	0
D2	All over…	0	Y tf=Y/N2

X, Y are counts of terms → TF-IDF

Longer documents will have higher average count values than shorter documents.
To avoid these potential discrepancies it suffices to divide the number of occurrences of each word in a document by the total number of words in the document: these new features are called tf for Term Frequencies.
Another refinement on top of tf is to downscale weights for words that occur in many documents in the corpus and are therefore less informative than those that occur only in a smaller portion of the corpus.

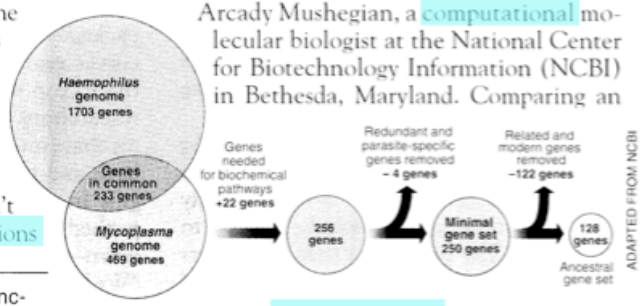
Topic Modeling

"Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents."
(Blei 2012:77)

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.

SCIENCE • VOL. 272 • 24 MAY 1996

Documents exhibit multiple topics.

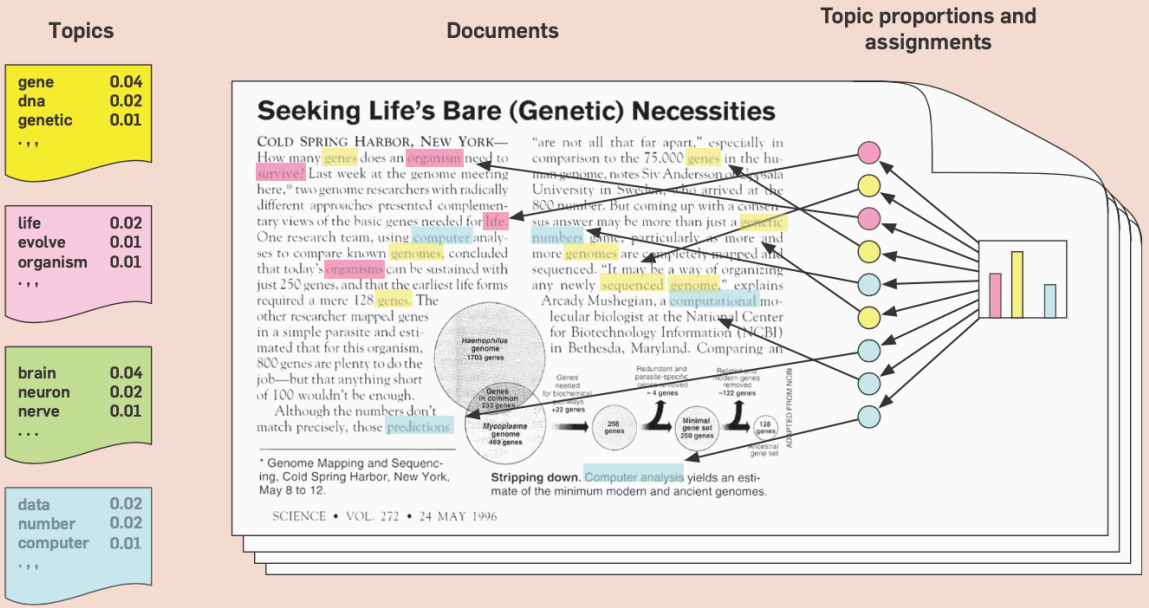
Topic Modeling – A non-tech Intro to LDA

1. Each document (text) within a corpus is viewed as a *bag-of-words* produced according to a mixture of themes that the author of the text intended to discuss.
2. Each theme (or topic) is a distribution over all observed words in the corpus, such that words that are strongly associated with the document's dominant topics have a higher chance of being selected and placed in the document bag.
3. The objective of topic modeling is to find the parameters of the LDA process that has likely generated the corpus.

Mohr and Bogdanov 2013

Stony Brook University

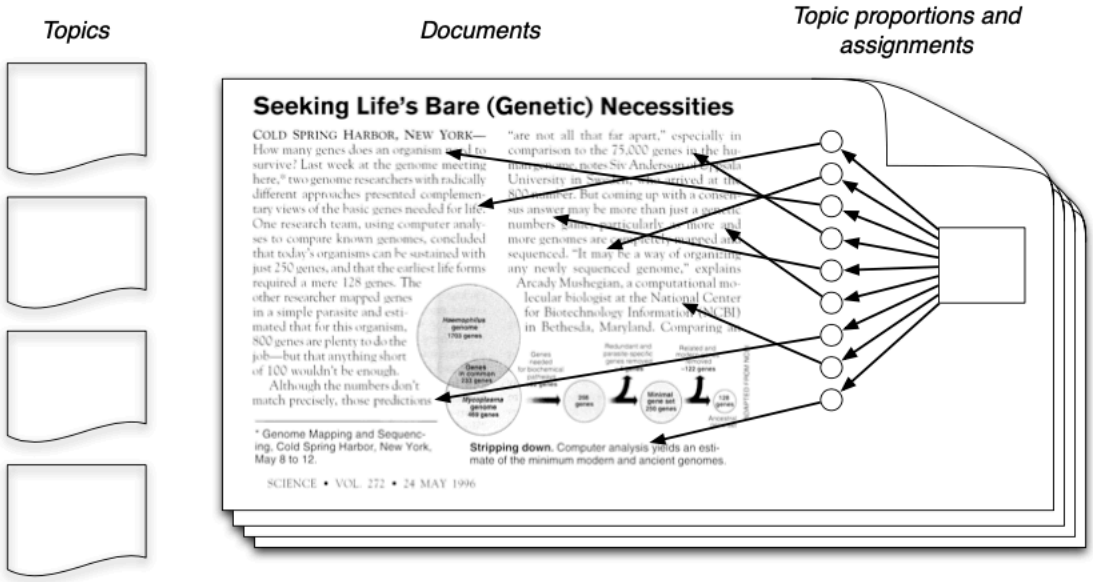
Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



We formally define a *topic* to be a distribution over a fixed vocabulary. For example, the *genetics* topic has words about genetics with high probability and the *evolutionary biology* topic has words about evolutionary biology with high probability. We assume that these topics are specified before any data has been generated.[a] Now for each document in the collection, we generate the words in a two-stage process.

- Randomly choose a distribution over topics.
- For each word in the document
  a. Randomly choose a topic from the distribution over topics in step #1.
  b. Randomly choose a word from the corresponding distribution over the vocabulary.

This statistical model reflects the intuition that documents exhibit multiple topics. Each document exhibits the topics in different proportion (step #1); each word in each document is drawn from one of the topics (step #2b), where the selected topic is chosen from the per-document distribution over topics (step #2a).[b]



Latent Dirichlet Allocation

Topic Modeling – A Tech Intro to Latent Direchlet Allocation

The Formal Definition of LDA (Blei 2012)

We can describe LDA more formally with the following notation. The topics are $\beta_{1:K}$, where each β_k is a distribution over the vocabulary (the distributions over words at left in Figure 1). The topic proportions for the dth document are θ_d, where $\theta_{d,k}$ is the topic proportion for topic k in document d (the cartoon histogram in Figure 1). The topic assignments for the dth document are z_d, where $z_{d,n}$ is the topic assignment for the nth word in document d (the colored coin in Figure 1). Finally, the observed words for document d are w_d, where $w_{d,n}$ is the nth word in document d, which is an element from the fixed vocabulary.

With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables,

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right). \quad (1)$$

Notice that this distribution specifies a number of dependencies. For example, the topic assignment $z_{d,n}$ depends on the per-document topic proportions θ_d. As another example, the observed word $w_{d,n}$ depends on the topic assignment $z_{d,n}$ and *all* of the topics $\beta_{1:K}$. (Operationally, that term is defined by looking up as to which topic $z_{d,n}$ refers to and looking up the probability of the word $w_{d,n}$ within that topic.)

problem, computing the conditional distribution of the topic structure given the observed documents. (As we mentioned, this is called the *posterior*.) Using our notation, the posterior is

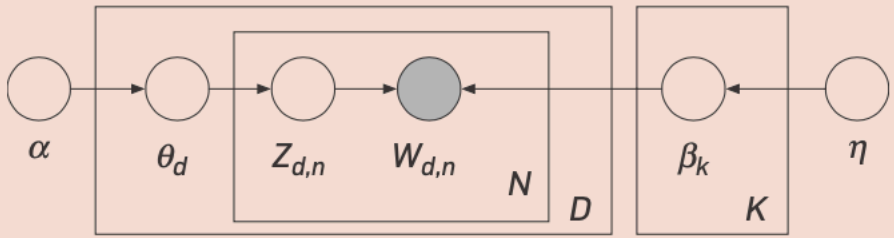
$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} \mid w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}. \quad (2)$$

The numerator is the joint distribution of all the random variables, which can be easily computed for any setting of the hidden variables. The denominator is the *marginal probability* of the observations, which is the probability of seeing the observed corpus under any topic model. In theory, it can be computed by summing the joint distribution over every possible instantiation of the hidden topic structure.

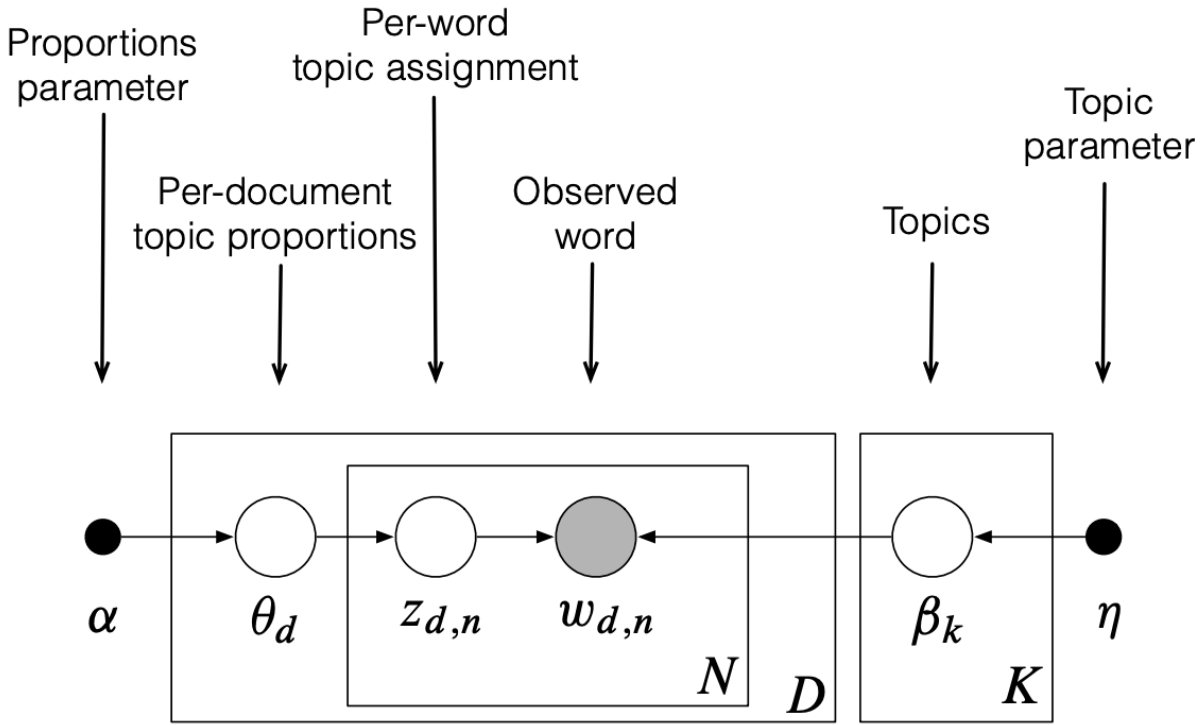
Topic Modeling – A Tech Intro to Latent Direchlet Allocation

The graphical Model of LDA

Figure 4. The graphical model for latent Dirichlet allocation. Each node is a random variable and is labeled according to its role in the generative process (see Figure 1). The hidden nodes—the topic proportions, assignments, and topics—are unshaded. The observed nodes—the words of the documents—are shaded. The rectangles are "plate" notation, which denotes replication. The *N* plate denotes the collection words within documents; the *D* plate denotes the collection of documents within the collection.



We can describe LDA more formally with the following notation. The topics are $\beta_{1:K}$, where each β_k is a distribution over the vocabulary (the distributions over words at left in Figure 1). The topic proportions for the dth document are θ_d, where $\theta_{d,k}$ is the topic proportion for topic k in document d (the cartoon histogram in Figure 1). The topic assignments for the dth document are z_d, where $z_{d,n}$ is the topic assignment for the nth word in document d (the colored coin in Figure 1). Finally, the observed words for document d are w_d, where $w_{d,n}$ is the nth word in document d, which is an element from the fixed vocabulary.



LDA as a graphical model

Stony Brook University

Topics found in 1.8M articles from the New York Times

Some Problems

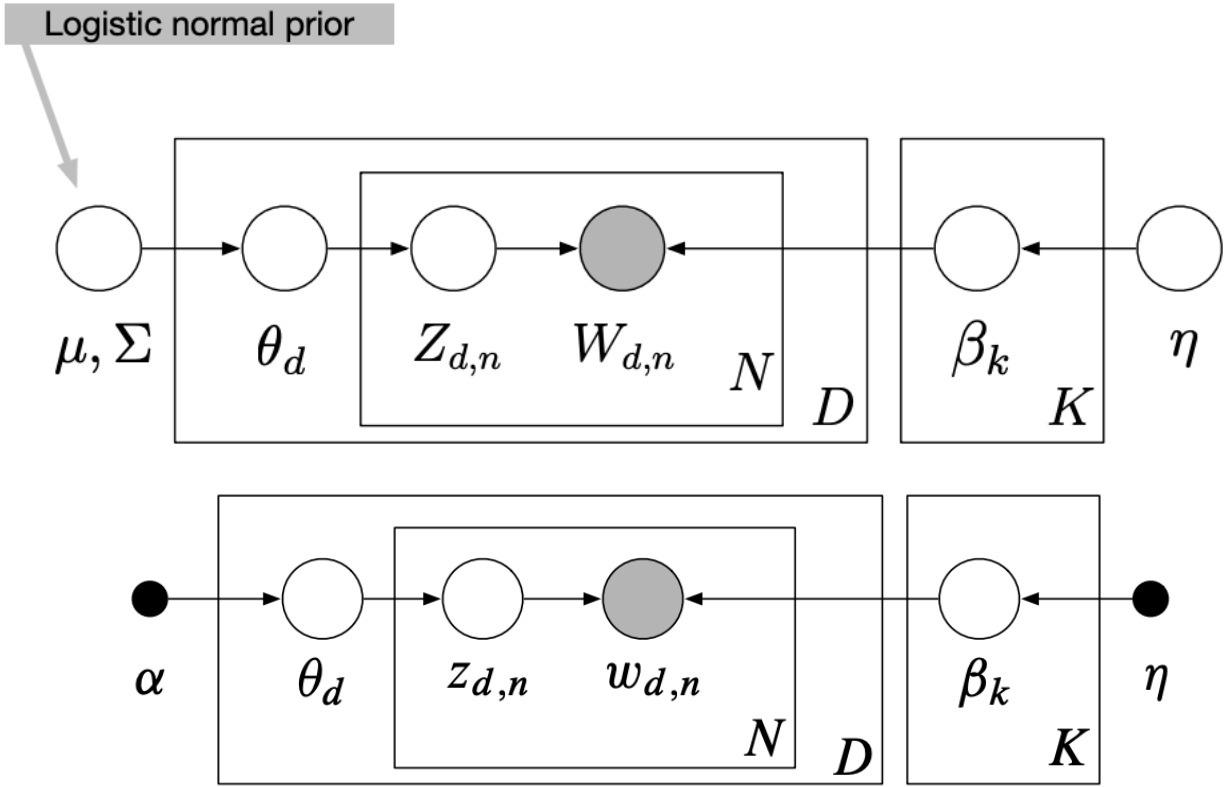
The order of words in a document does not matter

The order of documents in a collection does not matter

The topics are not correlated

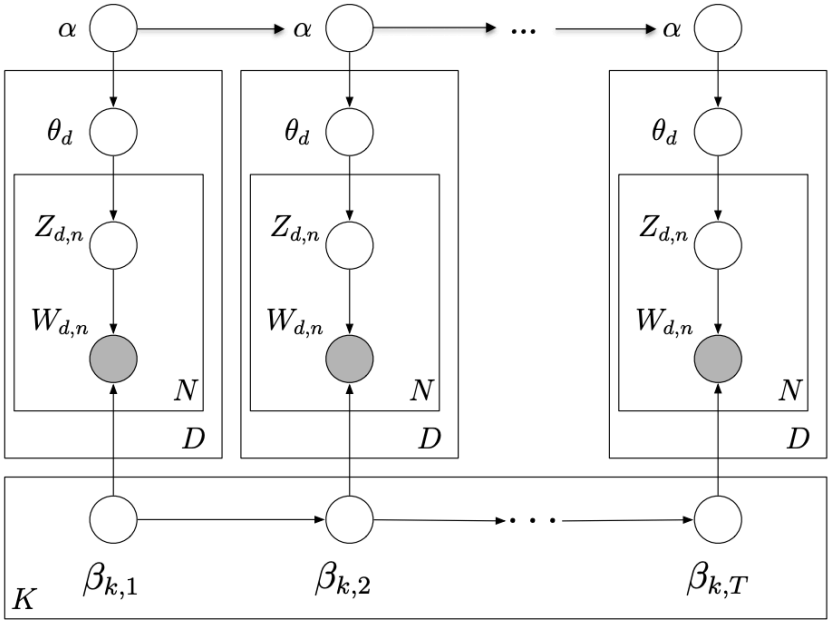
The metadata of the documents do not matter

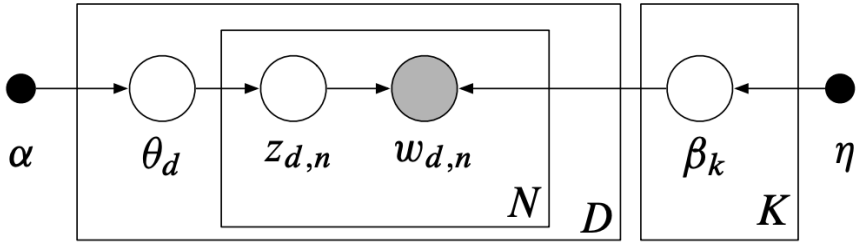
Correlated Topic Model



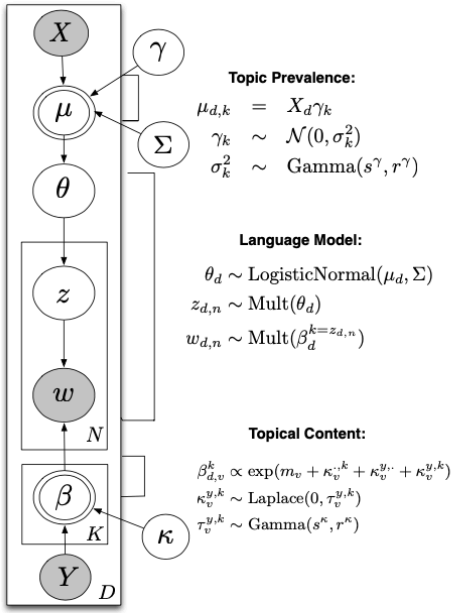
Dynamic Topic Model

The order of documents matters.





Structural Topic Model



Topic Prevalence:

$$\mu_{d,k} = X_d\gamma_k$$
$$\gamma_k \sim \mathcal{N}(0, \sigma_k^2)$$
$$\sigma_k^2 \sim \text{Gamma}(s^\gamma, r^\gamma)$$

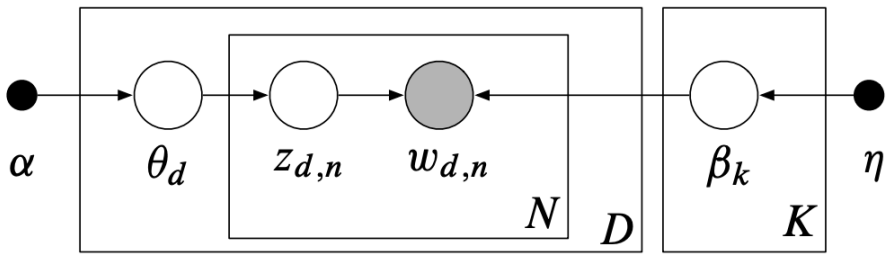
Language Model:

$$\theta_d \sim \text{LogisticNormal}(\mu_d, \Sigma)$$
$$z_{d,n} \sim \text{Mult}(\theta_d)$$
$$w_{d,n} \sim \text{Mult}(\beta_d^{k=z_{d,n}})$$

Topical Content:

$$\beta_{d,v}^k \propto \exp(m_v + \kappa_v^{\cdot,k} + \kappa_v^{y_\cdot,k} + \kappa_v^{y,k})$$
$$\kappa_v^{y,k} \sim \text{Laplace}(0, \tau_v^{y,k})$$
$$\tau_v^{y,k} \sim \text{Gamma}(s^\kappa, r^\kappa)$$

Figure 1: Plate Diagram for the Structural Topic Model



Structural Topic Model (STM)

The goal of the Structural Topic Model is to allow researchers to discover topics and estimate their relationship to document metadata.

Topic prevalence: the proportion of a document devoted to a topic; *topical content:* the word rates used in discussing a topic.

Three critical differences between STM and LDA:

1. topics can be correlated;
2. each document has its own prior distribution over topics, defined by covariate *X* rather than sharing a global mean;
3. word use within a topic can vary by covariate *U* .

Roberts et al. 2014 2016

The generative process for each document (indexed by d) with vocabulary of size V for a STM model with K topics can be summarized as:

1. Draw the document-level attention to each topic from a logistic-normal generalized linear model based on a vector of document covariates X_d.

$$\vec{\theta}_d | X_d\gamma, \Sigma \sim \text{LogisticNormal}(\mu = X_d\gamma, \Sigma) \tag{1}$$

   where X_d is a 1-by-p vector, γ is a p-by-$K-1$ matrix of coefficients and Σ is $K-1$-by-$K-1$ covariance matrix.

2. Given a document-level content covariate y_d, form the document-specific distribution over words representing each topic (k) using the baseline word distribution (m), the topic specific deviation $\kappa_k^{(t)}$, the covariate group deviation $\kappa_{y_d}^{(c)}$ and the interaction between the two $\kappa_{y_d,k}^{(i)}$).

$$\beta_{d,k} \propto \exp(m + \kappa_k^{(t)} + \kappa_{y_d}^{(c)} + \kappa_{y_d,k}^{(i)}) \tag{2}$$

   m, and each $\kappa_k^{(t)}$, $\kappa_{y_d}^{(c)}$ and $\kappa_{y_d,k}^{(i)}$ are V-length vectors containing one entry per word in the vocabulary. When no convent covariate is present β can be formed as $\beta_{d,k} \propto \exp(m + \kappa_k^{(t)})$ or simply point estimated (this latter behavior is the default).

3. For each word in the document, ($n \in 1, \ldots, N_d$):

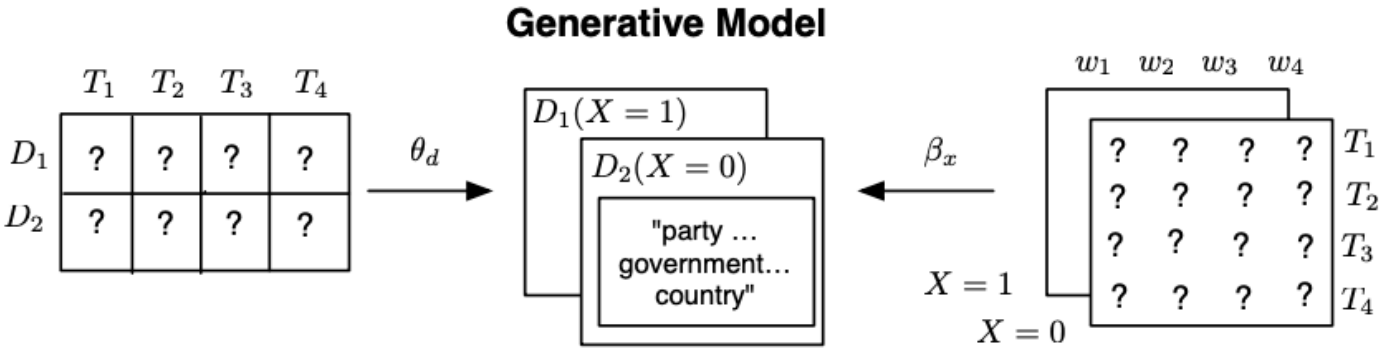
   - Draw word's topic assignment based on the document-specific distribution over topics.

$$z_{d,n} | \vec{\theta}_d \sim \text{Multinomial}(\vec{\theta}_d) \tag{3}$$

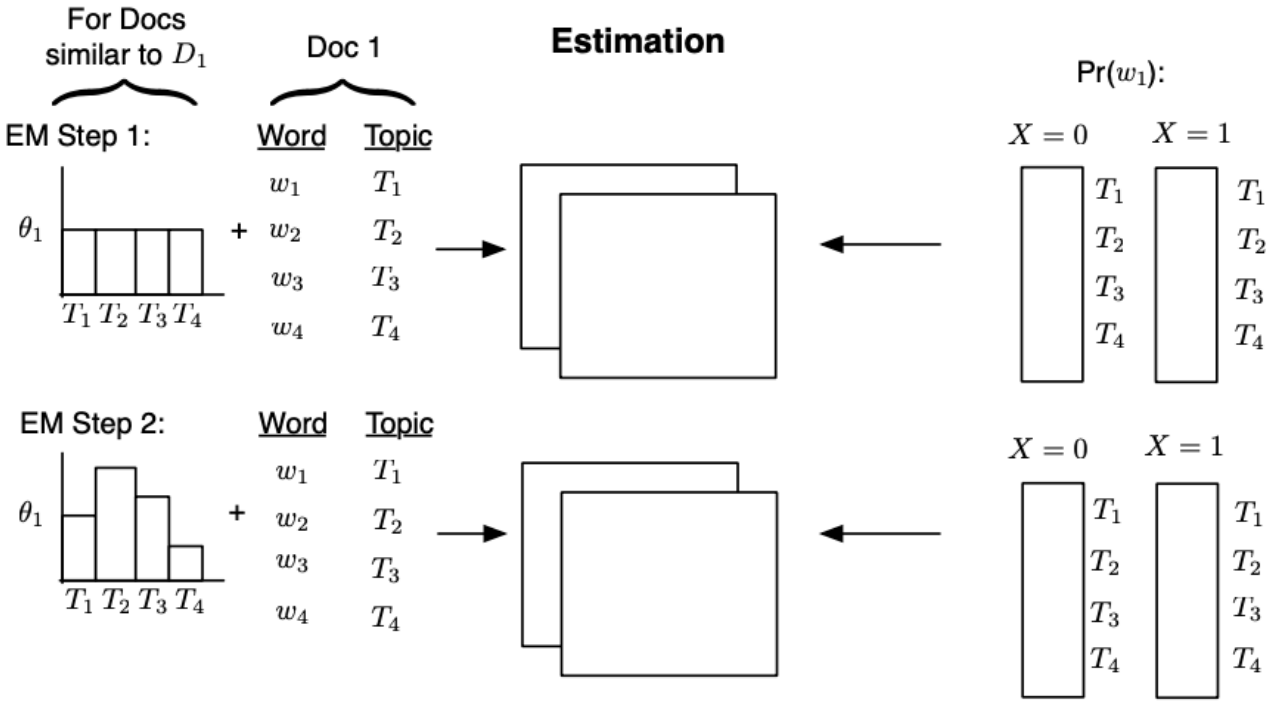
   - Conditional on the topic chosen, draw an observed word from that topic.

$$w_{d,n} | z_{d,n}, \beta_{d,k=z_{d,n}} \sim \text{Multinomial}(\beta_{d,k=z_{d,n}}) \tag{4}$$

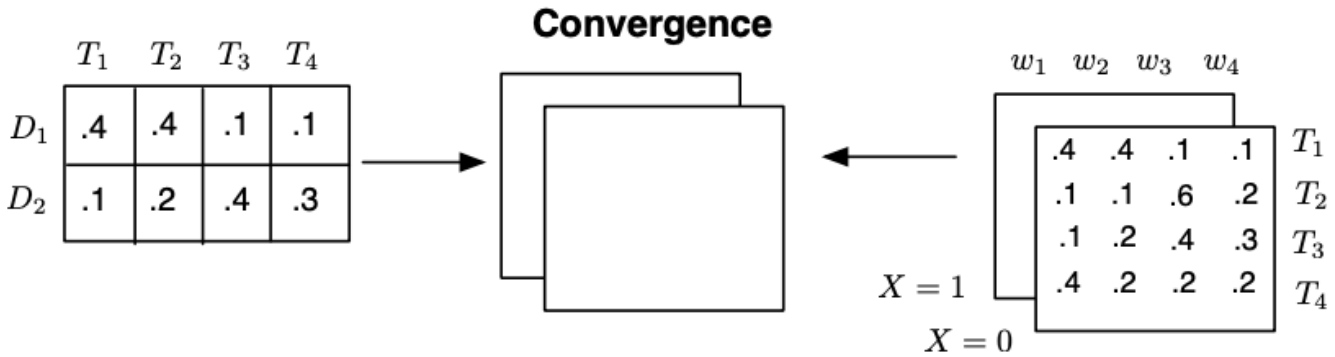
Graphical Representation of STM



Graphical Representation of STM



Graphical Representation of STM



Examples of Using Topic Model

Survey Responses with Experimental Design: You are assuming that open-ended response is a mixture of topics.

Structural Topic Models for Open-Ended Survey Responses

Margaret E. Roberts University of California, San Diego
Brandon M. Stewart Harvard University
Dustin Tingley Harvard University
Christopher Lucas Harvard University
Jetson Leder-Luis California Institute of Technology
Shana Kushner Gadarian Syracuse University
Bethany Albertson University of Texas at Austin
David G. Rand Yale University

Collection and especially analysis of open-ended survey responses are relatively rare in the discipline and when conducted are almost exclusively done through human coding. We present an alternative, semiautomated approach, the structural topic model (STM) (Roberts, Stewart, and Airoldi 2013; Roberts et al. 2013), that draws on recent developments in machine learning based analysis of textual data. A crucial contribution of the method is that it incorporates information about the document, such as the author's gender, political affiliation, and treatment assignment (if an experimental study). This article focuses on how the STM is helpful for survey researchers and experimentalists. The STM makes analyzing open-ended responses easier, more revealing, and capable of being used to estimate treatment effects. We illustrate these innovations with analysis of text from surveys and experiments.

Advance Access publication February 4, 2015

Political Analysis (2015) 23:254–277
doi:10.1093/pan/mpu019

Computer-Assisted Text Analysis for Comparative Politics

Christopher Lucas
Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge St., Cambridge MA 02138, USA
e-mail: clucas@fas.harvard.edu

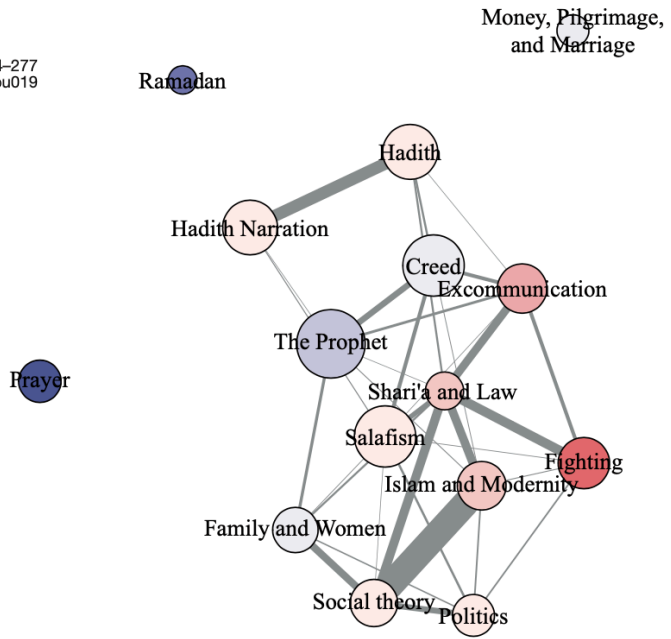


Fig. 2 The network of correlated topics for a 15-topic Structural Topic Model with Jihadi/not-Jihadi as the predictor of topics in Arab Muslim cleric writings. Node size is proportional to the number of words in the corpus devoted to each topic. Node color indicates the magnitude of the coefficient, with redder nodes having more positive coefficients for the Jihadi indicator and blue nodes having more negative coefficients. Edge width is proportional to the strength of the correlation between topics.

Visualizing Topical Content

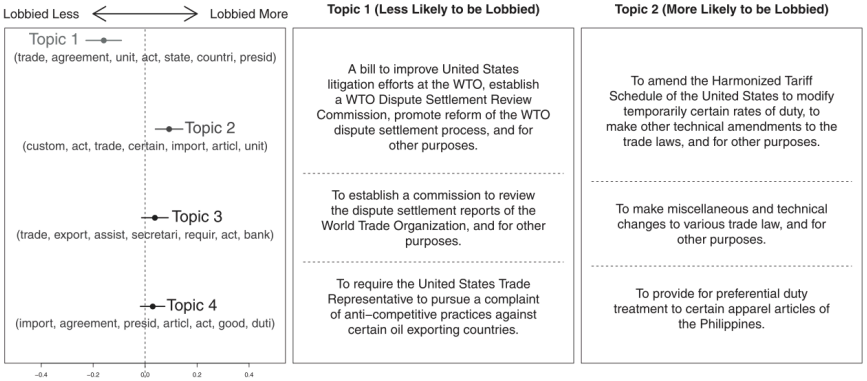
American Political Science Review (2017) 111, 1, 1–20
doi:10.1017/S0003055416000654

Political Cleavages within Industry: Firm-level Lobbying for Trade Liberalization

IN SONG KIM *Massachusetts Institute of Technology*

Existing political economy models explain the politics of trade policy using inter-industry differences. However, this article finds that much of the variation in U.S. applied tariff rates in fact arises within industry. I offer a theory of trade liberalization that explains how product differentiation in economic markets leads to firm-level lobbying in political markets. High levels of product differentiation eliminates the collective action problem faced by exporting firms while import-competing firms need not fear product substitution. To test this argument, I construct a new dataset on lobbying by all publicly traded manufacturing firms from reports filed under the Lobbying Disclosure Act of 1995. I find that productive exporting firms are more likely to lobby to reduce tariffs, especially when their products are sufficiently differentiated. I also find that highly differentiated products have lower tariff rates. The results challenge the common focus on industry-level lobbying for protection.

FIGURE 8. Topic of Trade Bills Likely to be Lobbied



Notes: The first panel compares the estimated topic prevalence for trade bills that are lobbied vs. not lobbied. It shows that Topic 2 is estimated to be associated with lobbied bills, whereas bills with Topic 1 are less likely to be lobbied. The words inside parenthesis represents top seven words associated with each topic. The next two panels display the titles of three example bills for Topic 1 and Topic 2 respectively. It reveals that Topic 2 bills tend to deal with modifying duties and other technical aspects of product-specific trade policies. The analysis is conducted based on a structural topic model (Roberts, Stewart, and Tingley 2016) by treating the presence of lobbying as an observed document-level covariate.

Stony Brook University

POLITICAL RESEARCH EXCHANGE
2019, VOL. 1, 1646102
https://doi.org/10.1080/2474736X.2019.1646102

ecpr Routledge Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Whose ideas are worth spreading? The representation of women and ethnic groups in TED talks

Carsten Schwemmer [a,b] and Sebastian Jungkunz [a,c]

[a]Chair of Political Sociology, University of Bamberg, Germany; [b]Weizenbaum Institute for the Networked Society, Germany; [c]Zeppelin University, Germany

ABSTRACT

We investigate the representation of women and ethnic groups in TED talks, which reach a large online audience on YouTube with science-related content and topics on societal change. We argue that gaps in representation can create a misleading perception of science and the respective topics discussed in these talks. We validate annotations from an image recognition algorithm for identifying speaker ethnicity and gender to compile a data set of 2333 TED talks and 1.2 million YouTube comments. Findings show that more than half of all talks were given by white male speakers. While the share of women increased over time, it is constantly low for non-white speakers. Topic modelling further shows that the share of talks addressing inequalities which affect both groups is low, but increasing over time. However, talks about inequalities and those given by female speakers receive substantially more negative sentiment on YouTube than others. Our findings highlight the importance of speaker and topic diversity on digital platforms to reduce stereotypes about scientists and science-related content.

ARTICLE HISTORY
Received 22 March 2019
Accepted 12 July 2019

KEYWORDS
Representation; women; ethnic groups; computational social science; YouTube

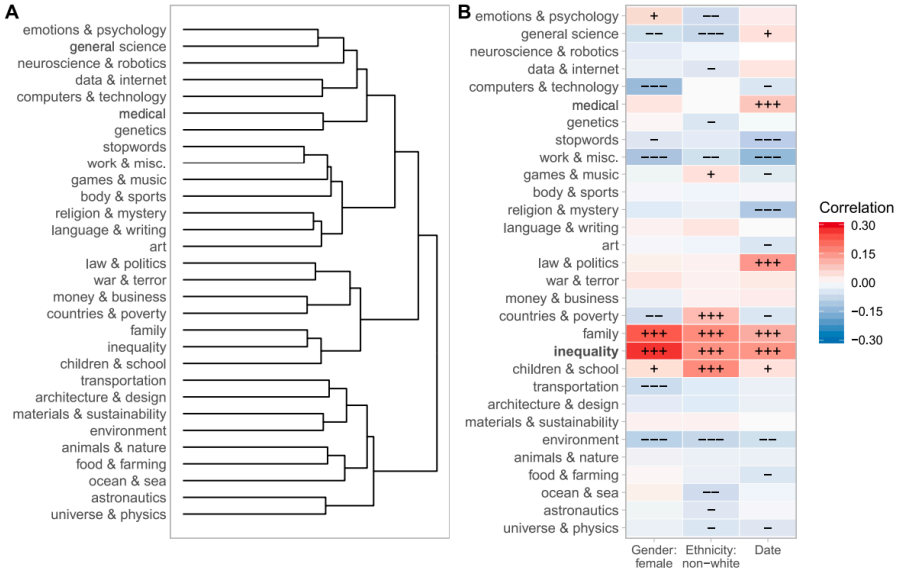


Figure 3. Topic clusters and correlations. (a) Hierarchical ward clustering of topics. (b) Pearson correlations between prevalence variables of the structural topic model and proportions of all topics. Characters denote *p* values for positive (+) and negative (–) correlations, adjusted for multiple comparisons (3=$p<0.001$, 2=$p<0.01$, 1=$p<0.05$).

Some Caveats of Using LDA Topic Models

We should write an article, titled "Limitations of Using Topic Models in Social Science."

1. The number of documents plays perhaps the most important role; it is theoretically *impossible* to guarantee identification of topics from a small number of documents, no matter how long.
2. The length of documents also plays a crucial role: poor performance of the LDA is expected when documents are too short, even if there is a very large number of them.
3. When a very large number of topics than needed are used to fit the LDA, the statistical inference may become inescapably inefficient.
4. The LDA performs well when the underlying topics are well-separated in the sense of Euclidean metric.
5. If it is believed that each document is associated with few topics, the Dirichlet parameter of the document-topic distributions should be set small (e.g. $\alpha \approx 0.1$). If the topics are known to be word-sparse, the Dirichlet parameter of the word distributions β is set small (e.g., 0.01), in which case learning is efficient.

Tang et al. 2014

Thank you!

Yongjun Zhang, Ph.D
Assistant Professor
Dept of Sociology and
Institute for Advanced Computational Science
Stony Brook University
Yongjun.Zhang@stonybrook.edu
https:yongjunzhang.com