Computational Sociology

Topic Modeling

Dr. Thomas Davidson

Rutgers University

March 22, 2021

Plan

1. Course updates
2. Introduction to topic modeling
3. Latent Dirichlet Allocation (LDA)
4. Structural Topic Modeling (STM)

Course updates

Feedback on proposals and homework 2

- Comments on proposals sent via Slack
- Homework 2 comments by the end of the week.

Course updates

Homework 3

- Homework 3 on NLP will be released by the end of the week.
  - Tentative due date is 4/2 at 5pm (you will have one week to complete it)

Course updates

Progress on projects

- Initial data collection due 4/9 at 5pm.
  - Preliminary analysis of some or all of the data you will use in your project; consider this as a draft of the Data section of your manuscript
    - Description of the data collection process
    - Preliminary data analysis
    - Include at least 1 summary table and 1 visualization
- Submit the following
  - A link to a Github repository (add me as a collaborator if it is private)
    - Code used to collect and analyze data (ideally in RMarkdown)
    - A rendered version of the RMarkdown document (PDF or HTML)

Introduction to topic modeling

What is topic modeling?

- A corpus of documents contains a set of latent "themes" or "topics"
- A topic model is a probabilistic algorithm is designed to inductively create a "model" of these latent topics
- This gives us an overview of the content of an entire corpus and the ability to characterize individual documents

Introduction to topic modeling

Topic modeling and sociology (Mohr and Bogdanov 2013)

- Topic modeling is a "lens" to allow reading of a corpus "in a different light and at a different scale" to traditional content analsyis.
  - Often we want to conduct an automated coding of a large corpus, but it can also be helpful for a "close reading" of a smaller corpus

Introduction to topic modeling

Topic modeling and sociology (Mohr and Bogdanov 2013)

- The approach still requires interpretation, but produces "a fundamental shift in the locus of methodological subjectivity", "from pre-counting to post-counting" (561).
  - Traditional content analysis requires us to develop our categories of analysis before coding a corpus, whereas the interpretative phase of LDA occurs after we have trained a model
  - "topic models do not remove the scholarly or the hermeneutic work from the project of analyzing a textual corpus, topic models simply move the bulk of this labor over to the other side of the data modeling procedure."

Introduction to topic modeling

Topic modeling and sociology (DiMaggio, Nag, and Blei 2013)

- Topic modeling is "an inductive relational approach to the study of culture"
  - It is *inductive* because the topics are "discovered" from the text
  - It is *relational* because meaning emerges out of relationships between words
- Topic modeling provides a high level of "substantive interpretability"
- The approach can discover different theoretical properties of language of interest to sociologists of culture
  - Framing
  - Polysemy
  - Heteroglossia

Introduction to topic modeling

Topic modeling and sociology (DiMaggio, Nag, and Blei 2013)

"Topic modeling will not be a panacea for sociologists of culture. But it is a powerful tool for helping us understand and explore large archives of texts. Used properly by subject-area experts with appropriate validation, topic models can be valuable complements to other interpretive approaches, offering new ways to operationalize key concepts and to make sense of large textual corpora."

Introduction to topic modeling

Topic modeling and sociology (Karell and Freedman 2019)

- Karell and Freedman use topic modeling to study the rhetoric of militant groups in Afghanistan
  - They characterize the process as *computational abduction* and describe "a recursive movement between the computational results, a close reading of selected corpus material, further literature on social movements, and additional theories that engendered a novel conceptualization of radical rhetorics."
  - The topic model results are then used to characterize the discourse of different groups. They then study how this discourse varies according to alliances and events.

Latent Dirichlet Allocation

Naming

- We observe the documents but the topics are **latent**
- The model is based on a probability distribution called the **Dirichlet** distribution
- Using this model we **allocate** words to topics

Latent Dirichlet Allocation

Intuition

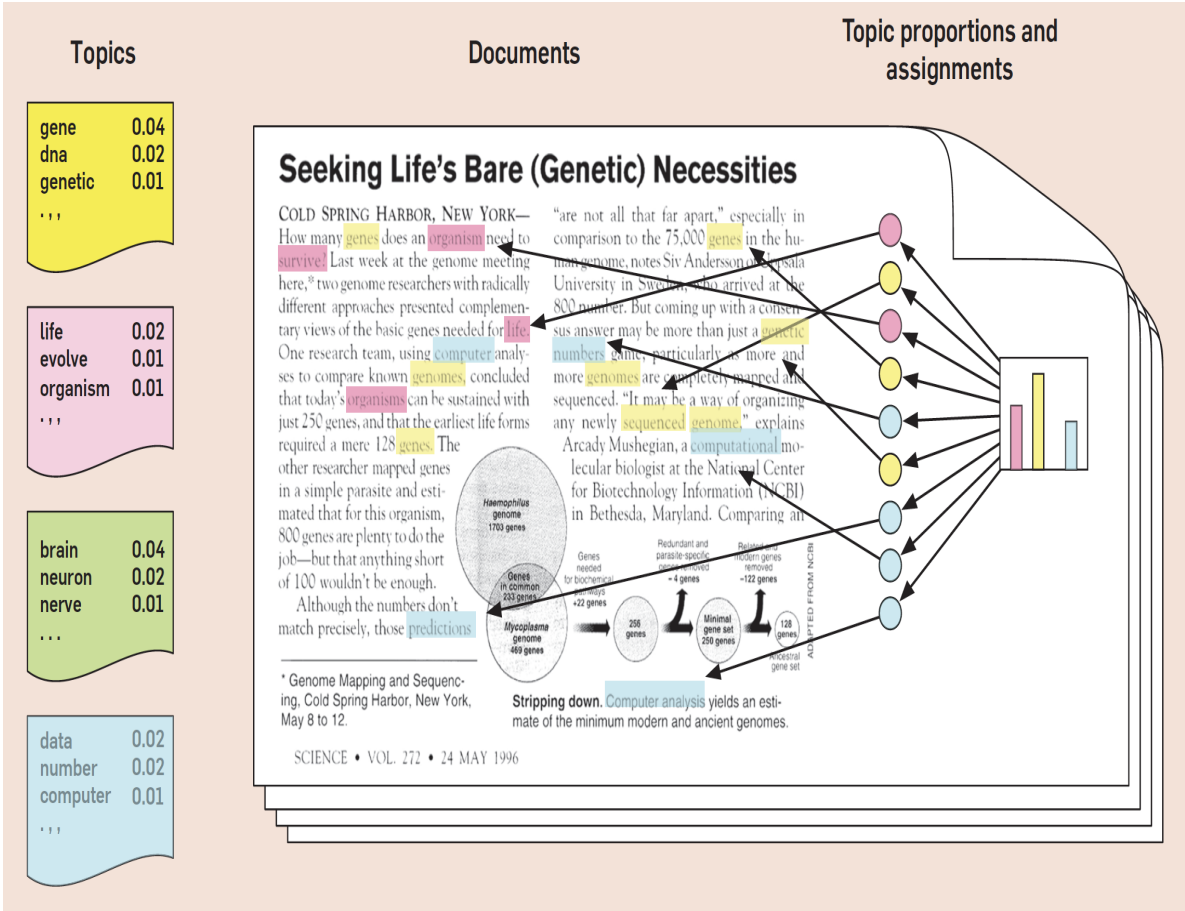
- A **topic** is a distribution over a vocabulary
  - Each word in the vocabulary has a probability of belonging to the topic
- Let's say we train an LDA on a newspaper corpus
  - We find a topic that seems to capture information about *sports*
    - The words "football" and "goal" have a high probability
    - The words "literary" and "helicopter" have a low probability
  - This topic can be represented as a distribution over all words in the vocabulary
    - $\textit{Topic distribution}_k = [\textit{football} : 0.34, \textit{goal} : 0.23, ..., \textit{literary} : 0.0001, \textit{helicopter} : 0.0002, ...]$

Latent Dirichlet Allocation

Intuition

- A **document** is a distribution over topics
  - *All* documents contain *all* topics, but in different proportions
- Let's say we take an article about a new player a football team hired and look at the topics based on the newspaper model
  - The highest probability topic might be *sports*, but the article also disusses contract and the position of the player in the labor market
    - Thus the document may also contain the topics *finance* and *labor*.
  - The article is irrelevant to other issues in discussed in newspapers, so has a low probability of containing the topic *national security* or *arts*.
  - $\textit{Topic proportions}_d = [\textit{sports} : 0.63, \textit{finance} : 0.25, \textit{labor} : 0.12, ..., \textit{national security} : 0.001, \textit{arts} : 0.002, ...]$

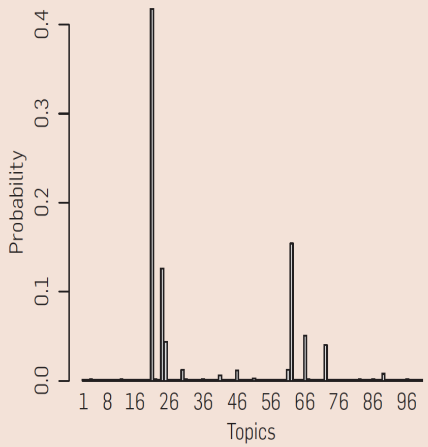
Latent Dirichlet Allocation



Blei, David M. 2012. "Probabilistic Topic Models." Communications of the ACM 55 (4): 77. https://doi.org/10.1145/2133806.2133826.

Latent Dirichlet Allocation

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Blei, David M. 2012. "Probabilistic Topic Models." Communications of the ACM 55 (4): 77. https://doi.org/10.1145/2133806.2133826.

Latent Dirichlet Allocation

Intuition

- Topic modeling is a *generative* process
  - The goal is to create a plausible model that can mimic the hidden structure and *generate* the observed documents
  - "“The utility of topic models stems from the property that the inferred hidden structure resembles the thematic structure of the collection.” Blei, 2012.

Latent Dirichlet Allocation

Mathematical formulation

- There are four components that we need to compute the LDA over a corpus of documents
  - Topics $\beta_{1:K}$, where β_k is a distribution over the vocabulary
  - Topic proportions $\theta_{1:D}$, where $\theta_{d,k}$ is proportion for topic k in document d
  - Topic assignments $z_{1:D}$, where $z_{d,n}$ is topic assignment of n^{th} word in document d.
  - Observed words $w_{1:D}$, where $w_{d,n}$ is the n^{th} word in document d.

Latent Dirichlet Allocation

Mathematical formulation

- The relationship between these variables is expressed as a joint-distribution:

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

Latent Dirichlet Allocation

Algorithm

- We use this formula to compute the *posterior distribution* of the variables
- This is a computationally-intensive process and requires some probabilistic short-cuts
  - *Sampling-based* methods approximate the posterior distribution by random sampling
  - *Variational* methods find an approximation of the posterior distribution that fits well

Latent Dirichlet Allocation

Exercise

- Each group will analyze 8 topics.
  - Modify the code above to look at top words and documents in each topic.
  - For each topic:
    - Can you come up with a short name and description of the topic?
    - Does the topic capture something meaningful about the documents?
    - Is there anything else you notice about the topic?
- Add your findings to the shared Google Sheet

Latent Dirichlet Allocation

Interpreting the results

"Producing an interpretable solution is the beginning, not the end, of an analysis. The solution constructs meaningful categories and generates corpus-level measures (e.g., the percentage of documents in which a given topic is highly represented) and document-level measures (e.g., the percentage of words in each document assigned to each topic) based on these categories. It remains for the analyst to use this information to address the analytic questions that motivated the research. The analyst must also validate the solution by demonstrating that the model is sound and that his or her interpretation is plausible" (DiMaggio et al. 2013: 586).

Latent Dirichlet Allocation

Validation

- "there is no statistical test for the optimal number of topics or for the quality of a solution" (DiMaggio, Nag, and Blei 2013: 582)
- They suggest three forms of validation
  - *Statistical*: Calculating various measures of fit
  - *Semantic*: Hand-coding / close reading of documents
  - *Predictive*: Do events change the prevalence of topics?
- Not all topics will be meaningful, some capture residual "junk" and can be ignored (Karell and Freedman 2019)

Latent Dirichlet Allocation

How does it differ from NLP approaches covered so far?

- Document representation
  - A document is represented as a probability distribution over topics, not just a bag-of-words or an embedding
  - Closest approach is *latent semantic analysis*, where a document is a set of weights over latent dimensions. Indeed, LDA was developed as an extension of LSA.
- Document retrieval
  - We can select documents by topic content rather than keywords
  - Words can be shared by multiple topics, unlike conventional keyword detection
- Document comparisons
  - We can compare documents based on topic content rather than text similarity

Latent Dirichlet Allocation

Extensions of LDA

- LDA is considered the "vanilla" topic model. Subsequent approaches have relaxed some of the assumptions of LDA (Blei, 2012):
  - *Assumption 1*: Documents are treated as bags-of-words
    - Language models can be incorporated to better account for linguistic structure
  - *Assumption 2*: Document order does not matter
    - Dynamic topic models account for how topics can change over time
  - *Assumption 3*: The number of topics, K, is known
    - Bayesian non-parametric topic models discover K during the inference procedure
- *Structural topic modeling* (STM) has recently become a popular extension of LDA

Structural Topic Modeling

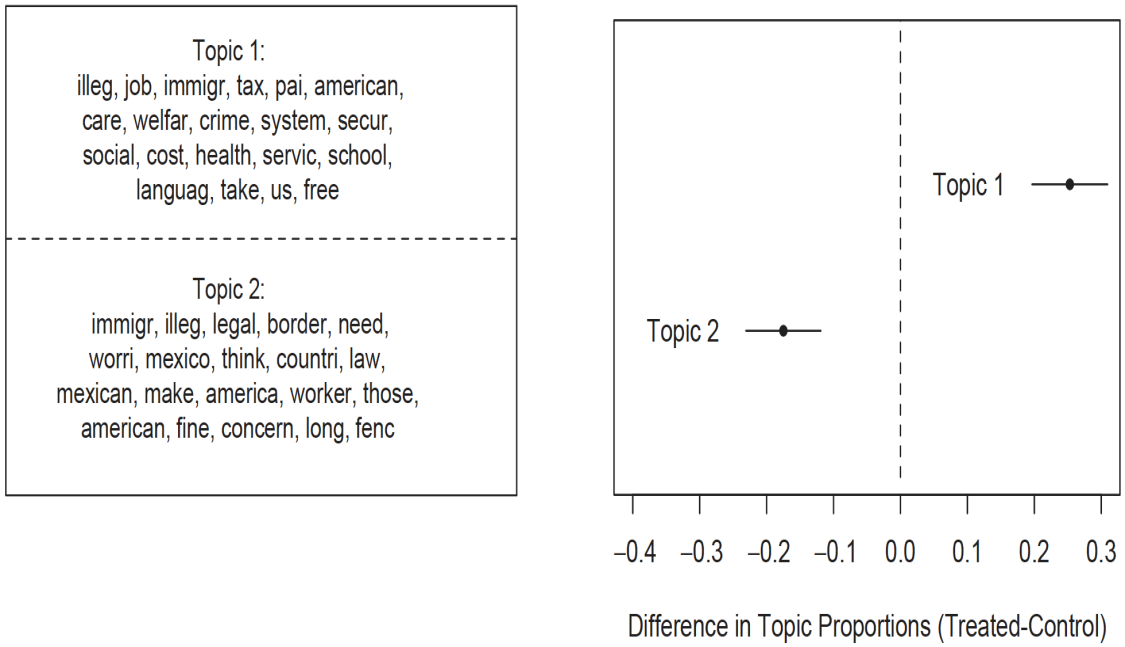
Background

- LDA assumes *topic prevalence* (frequency topic is mentioned) and *topic content* (the words used to discuss a topic) are constant across documents
  - e.g. In the previous example, we assume that *NYT* and *WSJ* devote equal coverage to topics and discuss topics in the same way.
- STM extends LDA by "allowing for the inclusion of covariates of interest into the prior distributions for document-topic proportions and topic-word distributions" (Roberts et al. 2014).
  - This allows analysis of how topics vary according to other factors, for example the treatment in a survey experiment may alter open responses.

Structural Topic Modeling

Analyzing open-ended survey responses using an STM

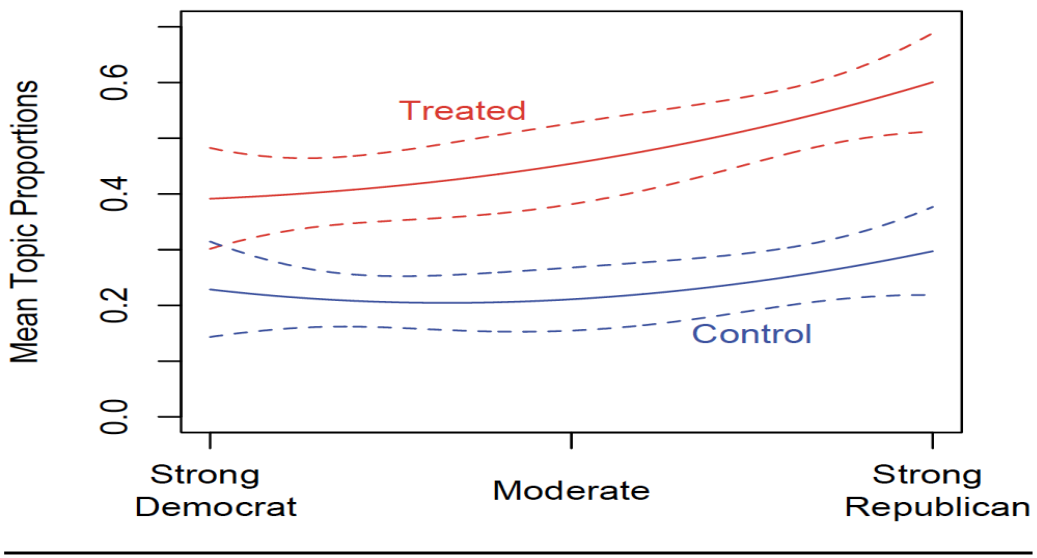
FIGURE 7 Words and Treatment Effect Associated with Topic 1



Structural Topic Modeling

Analyzing open-ended survey responses using an STM

FIGURE 8 Party Identification, Treatment, and the Predicted Proportion in Topic 1



Structural Topic Modeling

- Resources
  - The STM website contains information on various tools and research papers that use the approach
    - There are several packages including `stmBrowser`, `stmCorrViz` and `stminsights` that enable more interactive visualization.
  - The vignette provides a closer description of the methodology and a hands-on guide to using the `stm` package.

Summary

- Topic modeling is an inductive approach for the summary of large text corpora
  - Analysis of topic models involves the interpretation of topics
  - A key challenge is selecting an appropriate number of topics
- LDA algorithm summarize as corpus into K topics
  - Each document is composed of a mixture of topics
  - Each topic is a mixture of words
- STM improves on LDA by allowing topic prevalence and content to vary by covariates
  - This is particularly useful for social scientific applications