university of groningen

Quantitative Text Analysis – Essex Summer School

Topic models

dr. Martijn Schoonvelde

University of Groningen

Today's class

- Introduction to unsupervised and semi-supervised topic models
  - How they work, and when to use them
  - Strengths and limitations
- Lab session
  - Topic model exploration, labeling, and visualization

Why Topic Models?

- Topic modeling helps us discover latent themes in a text corpus *without pre-specifying categories*
- Useful for:
  - Summarizing large collections of text
  - Exploring new corpora
  - Informing downstream tasks like classification or measurement
- But keep in mind your research goal: discovery vs. measurement (Grimmer, Roberts, & Stewart, 2022)
  - Discovery: generating new hypotheses or understanding the corpus
  - Measurement: using topics as variables – requires careful validation and alignment with theoretical constructs

Topic models

Research goal : find a number of k topics, consisting of specific words and / or documents, that minimize the mistakes we would make if we try to reconstruct the corpus from the topics (Van Atteveldt *et al.* 2022, p. 203)

- There are many different topics models, which share the following characteristics (Grimmer & Stewart, 2013):
  - Each topic is a probability distribution over features
  - The model assumes a generative process for observed text
- Some topic models are "mixture models" (a document can consist of multiple topics), others allocate each document to one topic only (single membership model)

Why Focus on LDA?

Why focus on Latent Dirichlet Allocation (LDA; Blei, Ng & Jordan, 2003)?

- Widely used, flexible, and relatively intuitive output:
  - Topic distribution per document – how much of each topic appears in each document
  - Word distribution per topic – what words define each topic
  - E.g., "Document A is 80% topic 1, 15% topic 2; Document B is 60% topic 3..."
- Serves as a foundation for many extensions:
  - Correlated Topic Models (CTM)
  - Dynamic Topic Models (DTM)

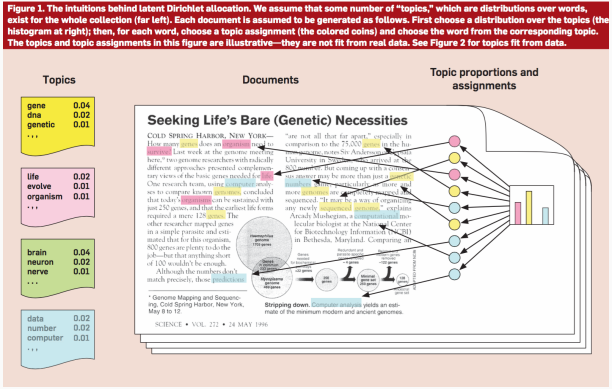
Assumptions Behind LDA

- Documents are modeled as a mixture of topics, and topics as a mixture of words
  - Each word in a document is drawn from one of the topics assigned to that document
  - This is a mixed membership model
- Number of topics k is fixed in advance (needs to be chosen by the researcher)
- LDA is a probabilistic model:
  - Running the same model multiple times may yield different results due to random initialization
  - This is known as the multimodality problem (Roberts *et al.*, 2016)
  - Solutions include multiple runs, stability checks, and model validation

Generative process of text

For each **document** in the corpus, **words** are generated in a two-stage process:

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



Inference: from words to topics

- The only thing we observe is words
  - Our document-feature matrix
- Given our generative process, we infer the topic structure that is most likely to have generated the observed words
  - Move in the opposite direction of the generative process
- To start off this inferential process we need to make some assumptions:
  - A prior distribution of topics per document θ
  - A prior distribution for words across topics β

Prior Distribution: Dirichlet

- LDA assumes that initial (i) topic distributions across documents, and (ii) word distributions across topics follow a Dirichlet distribution
- Dirichlet distribution: "distribution of multinomial distributions"
  - For D documents and K topics: D multinomial distributions of size K
  - For N words and K topics: N multinomial distributions of size K
- Hence: Latent Dirichlet Allocation
- These distributions themselves have hyperparameters α and η that govern their behavior
  - α controls topic proportions per document
  - η controls word proportions per topic

Document-topic distribution θ

	Topic 1	Topic 2	Topic 3	Topic K
Document 1	0.05	0.20	0.35	0.40
Document 2				
Document 3				
Document D				

$D \times K$ document-topic distribution

Word-topic distribution β

	Topic 1	Topic 2	Topic 3	Topic K
Word 1	0.25	0.20	0.15	0.40
Word 2				
Word 3				
Word N				

$N \times K$ word-topic distribution

From Prior to Posterior

- We begin with a prior distribution:
  - Randomly drawn Dirichlet distributions
  - These are assumptions about topic/word distributions *before* seeing the data
- After observing the corpus, we want a posterior distribution:
  - The updated topic and word distributions that are *most likely to have generated the observed text*
- This posterior cannot be computed exactly – we use approximate inference:
  - Gibbs sampling (sampling-based approach)
  - Variational inference (optimization-based approach)

Gibbs Sampling Algorithm

Gibbs sampling is a way to iteratively approximate the posterior distribution by resampling topic assignments.

Step 1: **Initialize**

- Assign each word in each document to a random topic

Step 2: **Iterate**

For each document:

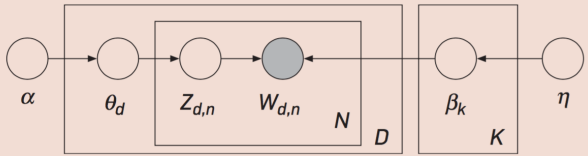
- For each word:
  - Reassign the word to a topic based on two things:
    1. How often this topic appears in the document (document-topic count)
    2. How often this word appears in each topic (topic-word count)

Step 3: **Repeat**

- Run for a number of iterations (chosen by the researcher)
- After a "burn-in" period, topic distributions begin to stabilize

Graphical model for LDA (Blei, 2012)

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.



Determining k

- Number of topics k is determined by the researcher
- One approach to the right number of topics: perplexity criterion
- Perplexity is a measure of the likelihood of a hold-out test set given the model
- Procedure:
  - Estimate topic models with various values of k
  - Calculate perplexity score.
  - Choose topic model with lower perplexity

Validating topic models in context of measurement (Grimmer & Stewart 2013)

- Semantic validity: extent to which topics are coherent
  - Absence of random, unrelated words
  - Topics that are specific enough and not overly general
  - Can be evaluated using coders – check out the **oolong** library in R (Chan, 2021)
- Predictive validity: how well does variation in topic usage correspond with predicted events
  - E.g, a terrorism topic in media reports should peak after a terrorist incident
- Convergent validity: extent to which model output can be validated with other approaches

Structural topic models

We often have metadata in our corpus

- For a newspaper corpus: *year*, *source*, *section*, *etc.*
- For a speech corpus: *speaker*, *party*, *etc.*
- For a social media corpus: *platform*, *account*, *etc.*

Structural topic models (Roberts *et al.*, 2014) allow a topic model to use that data to infer topical content and topical prevalence

- Accompanying website `structuraltopicmodel.com` provides extensive resources and vignettes

Original application was on open-ended survey data in the US political context. Do demographics and partisan preferences of respondents affect their responses?

Generative process

- Topics are initialized deterministically (if you run the same stm on the same data twice you get the same outcome)
- Topic proportions θ drawn from multinomial logistic normal distribution with covariates
  - Topical prevalence per document can correlate with covariates
- Topic words β also drawn from multinomial logistic normal distribution with covariates
  - Topical content can correlate with covariates

Semi-supervised topic models

Some types of topic models are semi-supervised in that they rely on an *ex ante* mapping of words to topics. As such they combine both inductive and deductive aspects

- **keyATM** (Eshima *et al.* 2023) and **seeded LDA** (Watanabe & Baturo, 2023)

Workflow consists of determining a set of keywords for a set of theoretically grounded topics and use these keywords to inform (some or all) of the topics

- Quality of the topics depends of in large part on the quality of the seedwords

Other types of topic models

- Many other types of topic models – with variations on model assumptions. But their accessibility in R varies:
  - **Correlated topic models** (Blei & Lafferty, 2005) – allow for the possibility that certain topics are correlated with each other
  - **Dynamic topic models** (Blei & Lafferty, 2006) – models over time variation in topical content (high loading words on a certain topic may vary) and topical prevalence
  - **Hierarchical topic models** (e.g., Grimmer, 2010) – does not require the researcher to set the number of topics in advance