Quantitative Text Analysis – Essex Summer School

Scaling methods

dr. Martijn Schoonvelde

University of Groningen

Today's class

- Topic models (LDA and STM)
- Lab session
- Flash talks: Nicholas, Paula, Julie

Topic models

- Why topic modeling?
  - Discover themes in a corpus without specifying them in advance
  - Summarize regularities in a corpus
- But keep in mind the research goal: discovery versus measurement (Grimmer, Roberts, & Stewart, 2022)
  - When using topic models in a measurement framework we need to make certain to validate these models extensively in light of our specific research goals

Topic models

Research goal : find a number of topics, consisting of specific words and / or document, 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):
  - Topics are represented as a probability distribution over words
  - Assume 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

LDA topic model

Why focus on the LDA topic model (Blei, Ng & Jordan 2003)?

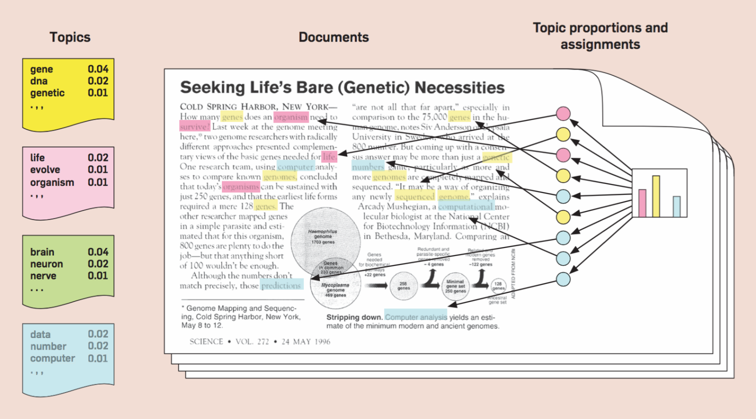
- Output of a LDA topic model is intuitive: topic probabilities for each document, and word probabilities for each topic
  - Allows for statements like: 80% of words in document A is assigned to topic 1, and 10% of words in document B is assigned to topic 1
- Forms basis for extensions such as correlated topic models or dynamic topics models (Blei & Lafferty, 2009)

Assumptions behind the LDA topic model

- Documents exhibit multiple topics
    - "mixed membership model" (as opposed to single membership topics model)
- Number of topics is fixed in advance
- LDA is a probabilistic model:
    - When you run a LDA topic model on the same data with the same parameter settings you may get different results
    - Problem of multimodality (Roberts et al. 2016)

Generative process of text

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.



First there is a distribution of topics, and then the rest follows

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 ϕ

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 K topics and N words: K multinomial distributions of size N
- Hence: Latent Dirichlet Allocation
- These distributions themselves have hyperparameters α and η that govern their behavior
  - For example, α smooths the observed distribution of topics per document

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

- Now we need to go from prior distributions to posterior distributions
  - Prior distribution: randomly selected Dirichlet distributions for topics per document and words per topic, based on hyperparameters α and η
  - Posterior distribution: estimate Dirichlet distribution that is most likely to have generated the observed words
- Inference can be can be done in different ways: Gibbs sampling or variational inference

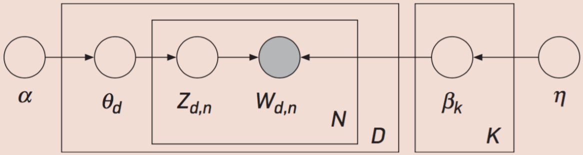
Gibbs sampling algorithm

Initialize topic assignment randomly according. For each iteration:

- For each document:
  - For each word:
    - Resample topic for word, given all other words and their topic assignments
    - Depends on (i) how many words in that document are assigned to a topic, and (ii) how often the word is assigned to each topic
- Number of iterations is determined by the researcher

Graphical model for LDA (Blei, 2012)

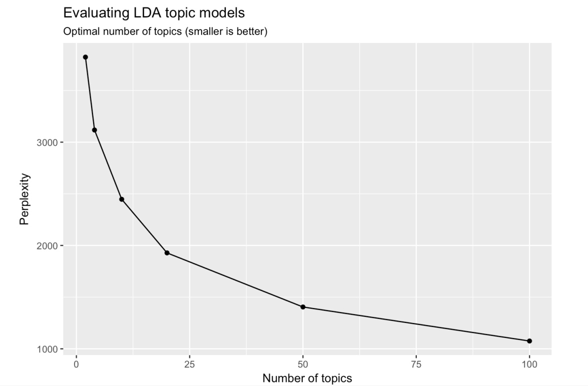
Figure 4. The graphical model for latent Dirichlet allocation. Each node is a random variable and is labeled according to its role in the generative process (see Figure 1). The hidden nodes—the topic proportions, assignments, and topics—are unshaded. The observed nodes—the words of the documents—are 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

Determining k



Perplexity of topic model on Associated Press dataset. From: http://cfss.uchicago.edu/fall2016/text02.html

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 (and with each other)
- Topic words β also drawn from multinomial logistic normal distribution with covariates
    - Topical content can correlate with covariates

Extensions to LDA

- Many other extensions to LDA – with variations on model assumptions.
  - 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