Computational Social Science

Topic Modeling I

Dr. Thomas Davidson

Rutgers University

October 28, 2024

Plan

1. Course updates
2. Introduction to topic modeling
3. Latent Dirichlet Allocation (LDA)

Course updates

- Homework 2 grades and proposal feedback in-progress
- Homework 3 will be released on Wednesday
  - Introduction to NLP
  - Word embeddings
  - Topic modeling

Introduction to topic modeling

What is topic modeling?

- A corpus of documents contains a set of latent "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

- Topic modeling is a "lens" to allow reading of a corpus "in a different light and at a different scale" to traditional content analysis (Mohr and Bogdanov 2013).
  - 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

Relationship to content analysis

- Content analysis is a methodology developed by social scientists to understand themes in written documents
- It involves reading a subset of documents and developing a set of themes or categories, then identify where these occur in the rest of the corpus
- In contrast to this approach, the interpretative phase of LDA occurs after we have trained a model
  - Topic modeling identifies potentially meaningful topics that we must then analyze

Introduction to topic modeling

An inductive and relational approach

- 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

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. Poetics, 41(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004

Latent Dirichlet Allocation

Terminology

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

Latent Dirichlet Allocation

Intuition

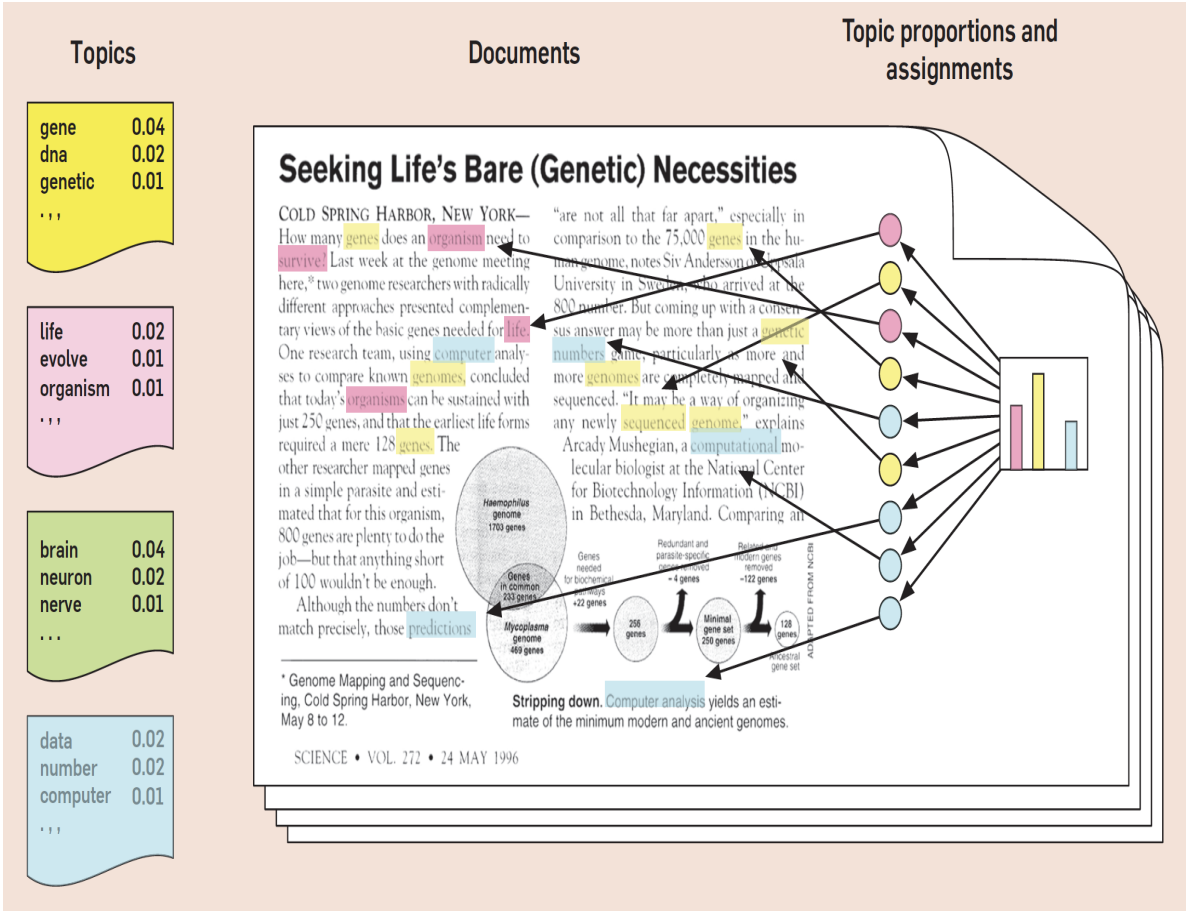
- A **topic** is a distribution over a vocabulary
  - Each word in the vocabulary has a probability of belonging to the topic
- Let's say we train an LDA on a newspaper corpus
  - We find a topic that seems to capture information about *sports*
    - The words "football" and "goal" have a high probability
    - 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*.
  - $\textit{Topic proportions}_d = [\textit{sports} : 0.63, \textit{finance} : 0.25, \textit{labor} : 0.12, ..., \textit{national security} : 0.001, \textit{arts} : 0.002, ...]$

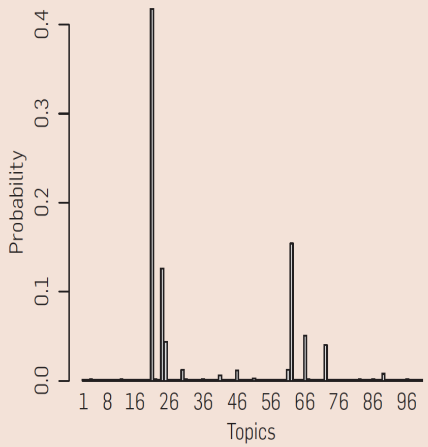
Latent Dirichlet Allocation



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 that generates 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.

* Note this is term is also used to refer to generative AI, but unlike ChatGPT and other large language models we cannot use topic models to generate plausible texts.

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

Training an LDA topic model in R

Loading the corpus

Loading a corpus of political tweets.

```
library(tidyverse)
library(tidytext)
data <- as_tibble(read_csv("../data/politics_twitter.csv"))
data$status_id <- as.character(data$status_id)
data <- data %>% distinct(text, .keep_all = TRUE) # Removing duplicates
```

Latent Dirichlet Allocation

Preprocessing

Removing hashtags, mentions and URLs using a regular expression. I initially left in hashtags and mentions but found that they tended to dominate some topics since different newspapers frequently used the same hashtags.

```
data <- data %>%
    mutate(text = gsub("#[A-Za-z0-9]+|@[A-Za-z0-9]", "", text)) %>%
    mutate(text = gsub("(http[^ ]*)|(www.[^ ]*)", "", text)) %>%
    mutate(text = gsub("’", "'", text))
```

Latent Dirichlet Allocation

Preprocessing

```
data(stop_words)

custom_stop <- tibble(word=c("amp"))

word.counts <- data %>% unnest_tokens(word, text, strip_punct = TRUE) %
  anti_join(stop_words) %>%
  anti_join(custom_stop) %>%
  count(status_id, word)

doc.freq <- word.counts %>%
  mutate(n = pmax(pmin(n, 1), 0)) %>%
  group_by(word) %>%
  summarize(df = sum(n))

words <- left_join(word.counts, doc.freq) %>%
    filter(df >= 50 & df <= 1900)
```

Latent Dirichlet Allocation

Constructing a DTM

```
DTM <- words %>%
  cast_dtm(status_id, word, n)
dim(DTM)
```

Latent Dirichlet Allocation

Training an LDA model using `topicmodels`

We can pass the DTM to the `LDA` function to train a topic model with 27 topics.

```
#install.packages("topicmodels")
library(topicmodels)
topic_model<- LDA(DTM, k=27, control = list(seed = 10980))
```

LDA analysis based on code from https://www.tidytextmining.com/topicmodeling.html and https://cbail.github.io/SICSS_Topic_Modeling.html.

Latent Dirichlet Allocation

Analyzing topics

We can use the `tidy` function to pull the beta matrix from the model. Each row corresponds to the probability of word w in topic k. In this case we can see the probabilities of the term "economy" in topics 1-10.

```
topics <- tidy(topic_model, matrix = "beta")
topics %>% filter(term == "economy") %>%
    dplyr::select(topic, beta) %>%
    head(10)
```

Latent Dirichlet Allocation

Analyzing topics

Latent Dirichlet Allocation

Analyzing topics

Latent Dirichlet Allocation

Analyzing topics

Latent Dirichlet Allocation

Analyzing documents

We can extract the document-topic proportions. Each row corresponds to the proportion of topic i in document j. Note the notation change. This matrix is called *gamma* (γ) here but was referenced as *theta* (θ) in the equation above. Unfortunately this kind of thing happens a lot as notation is used inconsistently.

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

Helper function

This function can be used to find the documents and words with highest weights in a particular topic.

Latent Dirichlet Allocation

Inspecting topics

Latent Dirichlet Allocation

Exercise

- Run code above to evaluate your assigned topic
- Add your findings to the shared Google Sheet (requires Rutgers login): https://tinyurl.com/topicmodelcoding

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

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 summarizes as corpus into k topics
  - Each document is composed of a mixture of topics
  - Each topic is a mixture of words
- These topics often capture important information about the meaning of document