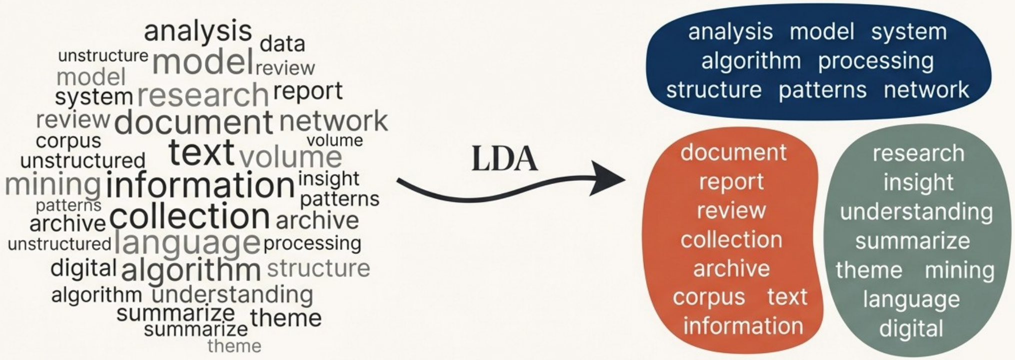
Digital Humanities: Text-as-Data

Week 6 – Topic Modeling & Latent Dirichlet Allocation (LDA)

Steven Denney
Leiden University

BA3 Korean Studies
November 21, 2025

WHAT IS TOPIC MODELING?



From a Sea of Text to Hidden Structures

- Massive collections of documents—articles, reviews, reports—are a goldmine of information, but their sheer volume makes them impenetrable.
- How can we automatically organize, understand, and summarize these large archives without reading every single word?
- The goal is to discover the hidden thematic structure—the "topics"—that run through the collection.

Topic Modeling: The Basic Idea

- Topic modeling is an **unsupervised** method for discovering **hidden thematic structure** in a corpus.
- It identifies **groups of words** that tend to appear together.
- Each document is represented as a **mixture of topics.**
- Topic modeling answers:
  - *What themes appear across these documents*?
  - *How much of each theme does a document contain*?
  - *Which words define each theme*?

Why We Use Topic Modeling

- Summarize long or numerous documents.
- Identify thematic patterns in large corpora.
- Compare themes across time, authors, genres, and periods.
- Reduce complex text into interpretable components.
- Particularly useful for:
  - textbook analysis,
  - historical periodization,
  - political discourse.

What a “Topic” Looks Like

- A **topic** is a set of words that tend to appear together.
- Examples (fictional but realistic):
  - **T1 (Ancient Korea)**: 고구려, 삼국시대, 백제, 신라
  - **T2 (Colonial Era)**: 일제, 독립, 저항, 항일
  - **T3 (Modernization)**: 산업화, 발전, 민주화, 경제성장
- Documents are mixtures of these topics in different proportions.

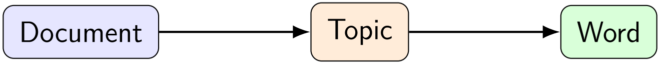
LATENT DIRICHLET ALLOCATION (LDA)

LDA: The Standard Model

- LDA is a model that uncovers the **hidden themes** that run through a collection of documents.
- It makes two simple assumptions:
  1. Each **document** is made up of several **topics**.
  2. Each **topic** is made up of several **words** that tend to occur together.
- To "generate" a document, LDA imagines that each word:
  - first chooses a topic,
  - then chooses a word from that topic's vocabulary.
- LDA works **backwards** from the real text to infer:
  - what the topics are, and
  - how each document mixes these topics.
- You specify the **number of topics** k.

The LDA Generative Story (Intuition Only)

- LDA imagines:
  - Documents are made of topics.
  - Topics are made of words.
- To create a document:
  - pick a topic,
  - pick a word associated with that topic,
  - repeat many times.



LDA then “reverses” this process to uncover the hidden topics in your data.

Two Key Probability Distributions

$$\phi_k(w) = p(w \mid k)$$

How strongly topic k uses word w.

$$\theta_d(k) = p(k \mid d)$$

How strongly document d uses topic k.

ϕ_k defines topics; θ_d describes documents.

Example Output

Topic Word Lists

T1 고구려, 백제, 신라, 삼국시대
T2 일제, 독립, 투쟁, 저항
T3 산업화, 민주화, 발전, 성장

Document–Topic Proportions

Doc	$T1$	$T2$	$T3$
1	0.90	0.05	0.05
2	0.10	0.85	0.05
3	0.05	0.10	0.85

Documents mix topics — they do not belong to just one.

Choosing k (Number of Topics)

- **Too few** topics $\rightarrow$ overly broad themes.
- **Too many** topics $\rightarrow$ noisy, incoherent themes.
- Rules of thumb:

Corpus Size	**Suggested k**
10–50 documents	3–6 topics
50–200 documents	5–12 topics
200+ documents	10–20 topics

Try several values of k and compare in LDAvis.

HOW TO DO LDA IN ORANGE

The Orange LDA Pipeline

1. **Import Text** (or use Corpus widget)
2. **Preprocess Text**
   - Tokenize
   - Remove stopwords (KR/EN)
   - Keep POS (nouns, verbs, adjectives)
3. **Topic Modeling (LDA)**
   - Choose number of topics k
   - Outputs: topic words, document–topic proportions
4. **LDAvis**
   - Explore topics interactively
   - Inspect relevance and distinctiveness
5. **Data Table**
   - Inspect topic weights per document

Topic Modeling Widget Output

- List of topics with top words + weights
- Topic–term matrix
- Document–topic proportions
- Send to:
  - Data Table
  - LDAvis

What LDAvis Shows You

- **Left panel:**
  - Topics list
  - **Relevance slider** (λ)
- **Right panel (bar chart):**
  - **Red bars** = how often a word appears within the topic
  - **Grey bars** = how often the word appears in the whole corpus
  - Difference shows **distinctiveness.**
- **Map view** (not shown here):
  - Each circle = a topic
  - Distance = similarity/difference
  - Circle size = prevalence

LDAvis links topic distinctiveness, frequency, and structure in one place.

The λ Slider in LDAvis

- $\lambda = 1.0$: ranks by **raw frequency** within the topic.
- $\lambda = 0.0$: ranks by **distinctiveness** (red » grey).
- Best practice: $\lambda \approx 0.2$–0.35.

Low λ reveals the words that make a topic unique.

INTERPRETING TOPICS RESPONSIBLY

Strengths of LDA

- Reveals hidden structure in large text corpora.
- Summarizes documents via a small set of themes.
- Allows documents to express multiple topics.
- Works well with interactive tools like LDAvis.

Limitations of LDA

- Topics reflect **statistical patterns**, not "true" meanings.
- Highly sensitive to preprocessing (stopwords, POS, etc.).
- Highly sensitive to number of topics k.
- May produce "junk topics" or mixed themes.
- Running LDA on subsets produces different topics.

How to Interpret Topics (Best Practices)

- Examine topics in LDAvis at $\lambda \approx 0.2$.
- Inspect documents with high weight for each topic.
- Use domain knowledge to label topics.
- Look for patterns across:
  - time,
  - authorship,
  - genre,
  - political era.
- Treat topic modeling as **a starting point**, not a final analysis.