BA2: Digital Korea

Week 10: Topic Modeling (LDA)

Steven Denney

Korean Studies
Leiden University

April 20, 2026

Today's Agenda

1. Review & check-in: from feeling to theme
2. What is topic modeling?
3. LDA: the intuition
4. Choosing *k* (the number of topics)
5. Orange's Topic Modelling widget under the hood
6. Application: the NIKH Korean history textbook corpus
7. Interpreting topics responsibly
8. Looking ahead

Review & Check-In

From Feeling to Theme

Week 9 asked *how* a text feels; Week 10 asks *what* it is about

Last week (sentiment): each text gets *one number* — a tone.

감사 + 희망 + 위기 ⇒ a positive/negative *score*

This week (topics): each text gets *several numbers* — a mixture of themes.

One textbook ⇒ 

Sentiment: **how** does the text feel?
Topics: **what** themes run through this collection of texts?

Warm-up in English

Shout out a label for each list — no wrong answers

1. kimchi · bulgogi · ramyeon · soju · rice · banchan
2. BTS · idol · fandom · album · stage · concert
3. king · palace · scholar · dynasty · robe · sword
4. subway · cafe · apartment · delivery · convenience · night

What You Just Did

Six words in, a theme out — that is the topic-modeling move

1. kimchi · bulgogi · ramyeon · soju · rice · banchan → ***Food***
2. BTS · idol · fandom · album · stage · concert → ***K-pop***
3. king · palace · scholar · dynasty · robe · sword → ***Historical drama***
4. subway · cafe · apartment · delivery · convenience · night → ***Modern Seoul life***

You read six co-occurring words and inferred a theme. LDA does the opposite — it reads a corpus and returns lists like these for you to label.

Now in Korean

Same game — each list is the top six words of one topic

1. 고구려 백제 신라 삼국 왕 시대
2. 일제 독립 저항 운동 항일 식민
3. 경제 산업화 발전 성장 정부 수출
4. 민족 정체성 민주 시민 통일 문화

Discuss with a neighbour

What label fits each list? Are any two close to each other?

Now in Korean: Answers

Plausible labels for the Korean lists

#	Top words	A plausible label
1	고구려, 백제, 신라, 삼국, 왕, 시대	Ancient & Three Kingdoms
2	일제, 독립, 저항, 운동, 항일, 식민	Colonial era / resistance
3	경제, 산업화, 발전, 성장, 정부, 수출	Modernisation
4	민족, 정체성, 민주, 시민, 통일, 문화	Modern Korean identity

The takeaway

This *is* what LDA produces: ranked lists of words. The **model** finds co-occurring words; **you** read them and assign the label. Labels are interpretations, not facts — a list about *nation*, *democracy*, *citizenship*, *identity* could plausibly be read two ways.

Where LDA Fits Among Our Methods

Each method gives a different *kind* of answer about a document

Week	Method	What each document gets
7	Clustering (K-means, hierarchical)	*one* group (a hard label)
8	Word embeddings	a vector — a position in meaning space
9	Dictionary sentiment	*one* score (a tone)
10	**LDA (topic modeling)**	**a *mixture* over themes (many weights)**

Clustering and sentiment each give a *single* answer per document; embeddings give a fixed representation.

LDA is the first method that gives each document a *multi-valued* answer. A document isn't one thing, it is a mix of things.

What Is Topic Modeling?

From a Sea of Text to Hidden Structure

The problem: corpora too big to read, too important to ignore

- The NIKH corpus we will use today covers **67 Korean history textbooks** (1895–2016).
- Even the 11-book demo already runs past a *million characters of Korean prose*.
- You cannot close-read it all. You also cannot reduce it to a single number.

The topic modeling question

What **themes** run through this collection, **how much** does each document use each theme, and **which words** define each theme?

Topic Modeling: The Basic Idea

An unsupervised method for discovering hidden thematic structure

- A **topic** is a group of words that tend to appear together across the corpus.
- Each **document** is represented as a **mixture of topics**.
- The algorithm does not know, in advance, what the topics are. It *discovers* them.

Topic modeling answers three questions:

1. What themes appear across the documents?
2. How much of each theme does a given document contain?
3. Which words *define* each theme?

How LDA Works

Latent Dirichlet Allocation, in Plain Words

The most widely used topic model; Orange runs it

LDA (Blei, Ng & Jordan, 2003) makes two simple assumptions:

1. Each **document** is made up of several **topics**, in some proportion.
2. Each **topic** is made up of **words that tend to occur together**.

Input: a term–document matrix of word counts — a bag of words, not TF–IDF weights.

Unpacking the name:

- **Latent** — topics are hidden; we only observe words.
- **Dirichlet** — the math for proportions (e.g., 70% + 20% + 10%). You don't work with it directly.
- **Allocation** — each word in each document gets allocated to one of the topics.

You, the researcher, specify one number: *k*, the **number of topics**.

The Generative Story

A thought experiment about how LDA imagines a document is written

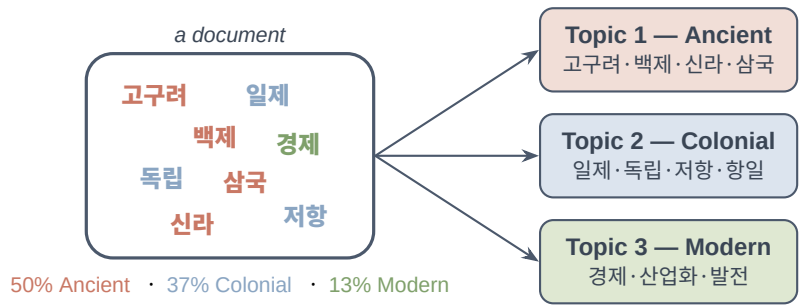
Pretend you are about to write a document. LDA imagines you do this:

1. Pick a **mixture** of topics (“this book will be 70% ancient, 20% colonial, 10% modern”).
2. For each word in the document:
   - pick one *topic* from the mixture,
   - pick one *word* from that topic’s vocabulary,
   - write it down.

LDA runs this story **backwards**: given the words we actually see, it infers the topics and the per-document mixtures.

Drawing Words from a Bag into Topic Bins

Running the story backwards: each word in a document is assigned to a topic



LDA reads each word, assigns it to a topic (the colour), and reports the proportions — that is the document's topic *mixture*.

Two Key Probability Distributions

What LDA estimates — every Orange output is a view of one of these

$$\underbrace{\phi_k(w)}_{\text{weight of word } w \text{ in topic } k} = \underbrace{p(w \mid k)}_{\text{probability of } w \text{ given topic } k}$$

Defines the topics.

	T1	T2	T3
고구려	0.08	0.00	0.00
일제	0.00	0.09	0.01
산업화	0.00	0.00	0.07
시대	0.04	0.02	0.02

$$\underbrace{\theta_d(k)}_{\text{weight of topic } k \text{ in document } d} = \underbrace{p(k \mid d)}_{\text{probability of topic } k \text{ given document } d}$$

Describes the documents.

Book	T1	T2	T3
1940 elementary	0.05	0.85	0.10
1981 middle school	0.30	0.15	0.55
2002 high school	0.35	0.10	0.55

Choosing the Number of Topics

How Many Topics (k)?

You pick k; no universal rule — research question and corpus size guide you

- k is a research choice. A coarse $k = 2$ can be plenty if you only want to distinguish two broad things; a fine-grained $k = 20$ makes sense on a big, heterogeneous corpus.
- Practical move: start small, *read the topics*, and raise k only if they feel too broad.
- Larger corpora can support more topics; smaller corpora usually shouldn't be asked for many.

For today's demo

11-book NIKH sample: we will use $k = 3$ (preselected).
Full 67-book corpus: $k = 5$ or 6 would be a reasonable starting point.

The Orange Pipeline

The Topic Modelling Widget in Orange

Input a Corpus, set k, read topics out the other side

What Orange does

- Fits an LDA model to your preprocessed corpus
- Shows top words per topic in a side panel
- Exports the two tables (ϕ and θ) as data you can pipe into other widgets

What you set

- `Number of topics` *k*: start with 4 or 5 for our demo
- Click *Commit*, or enable *Commit Automatically*, to refit
- Everything else (alpha, iterations, seed) stays at defaults

A note on methods

LDA is not the only way to do topic modeling — there are a couple of other methods available in the same widget. We use LDA throughout this course because it produces readable topic *mixtures* and lets you choose *k* directly. The others are described in an appendix slide if you are curious.

Inputs, Outputs, and Where Each Goes

One input port, three output ports, four downstream widgets

Port	What it carries	Plug into …
In: *Corpus*	Preprocessed documents (tokens)	—
Out: *Corpus*	Corpus with **topic weight** columns appended	Data Table, Box Plot by era
Out: *Topics*	Top words of the *selected* topic	Word Cloud
Out: *All Topics*	Full topic–word matrix (ϕ)	Heatmap, LDAvis

Clicking a topic row in the widget updates the *Topics* output; the other two ports always carry the full fit. Results change from run to run because LDA uses random initialisation.

A Minimum Orange Workflow

Load, tokenise, preprocess, model, visualise

Nouns Only for Topic Modeling

Why our preprocessing differs from Week 9

Week 9 sentiment kept **nouns, verbs, and adjectives**, because feeling lives in verbs and adjectives too. For topic modeling we keep **nouns only**.

POS tag	Meaning	Example
NNG	Common noun (일반명사)	독립 *(independence)*
NNP	Proper noun (고유명사)	고구려 *(Goguryeo)*

Why not verbs?

Topics are about *what a text is about*, and in Korean, *what* is carried by nouns. Verbs like 하다 (*do*) and 되다 (*become*) appear everywhere; they blur topics rather than separate them. Drop them, and the themes sharpen.

LDAvis: The Interactive View

A sibling widget in orange3-text, not an add-on you install separately

Pipe the *All Topics* output into the **LDAvis** widget and you get an interactive view of the model.

- **Map** on the left: each circle is a topic, circle size is prevalence, distance is dissimilarity.
- **Bar chart** on the right: red = how often the word appears in the selected topic, grey = how often across the whole corpus.
- λ **slider** blends raw frequency ($\lambda = 1$) with distinctiveness ($\lambda = 0$).

Practical tip

At $\lambda = 1$ the top words are whatever is frequent in the topic, which is often the same across topics. At $\lambda \approx 0.2$–0.35 you see the words that make each topic *distinctive*. Start at 0.3.

Application: The NIKH Corpus

The NIKH Corpus

Korean history textbooks from three political eras

The National Institute of Korean History corpus (국사편찬위원회) collects Korean history textbooks spanning 1895–2016. For today's demo we use the **11-book clustering sample** already on the course website.

Era	Books	Examples
Colonial (1940)	3	심상소학국사보충아동용 1–2; 교수참고서
Authoritarian (1973–1987)	4	중학교국사 3 차 (1973); 고등학교국사 4 차 (1981); …
Democratic (1995–2002)	4	고등학교국사 6 차 (1995); 초등학교사회 6-1 7 차 (2002)

The corpus is *small in document count* (11) but *large in prose* (tens to hundreds of thousands of characters per book). That is exactly the shape of corpus LDA was designed for.

Research Questions for the Demo

What we will ask of the NIKH corpus in class

1. What themes run through Korean history textbooks as a group?
2. Do Colonial-era textbooks emphasise different topics than post-1987 Democratic-era ones?
3. Is there a topic that is more prominent in elementary versus high-school textbooks?
4. Are some topics near-universal (every textbook uses them) and others era-specific?

A research choice, not a given

The `era` labels (Colonial / Authoritarian / Democratic) are an *editorial* split from the corpus metadata. Different historians might draw the lines differently. The periodisation is part of your analysis, not a fact handed to you by the algorithm.

Interpreting Topics Responsibly

Strengths of LDA

What the method is good for

- **Exploratory.** Reveals structure in corpora too large to close-read.
- **Mixture-aware.** A textbook can be 60% colonial-era and 40% modernisation; clustering could not say that.
- **Comparative.** Topic weights are numbers: you can group by era, decade, author, level.
- **Portable.** The same widget works on tweets, speeches, newspaper articles, or textbook chapters.

Limitations of LDA

What the method will *not* do for you

- **Topics are statistical, not semantic.** They are co-occurrence patterns; *you* give them names.
- **Sensitive to preprocessing.** Change stopwords, POS filter, or document length: topics change.
- **Sensitive to *k*.** Re-running with $k = 3$ vs $k = 7$ gives different themes.
- **Non-deterministic.** Two runs with the same *k* can produce different topic orderings (and sometimes different topics).
- **Junk topics happen.** Expect one or two topics to be incoherent at any given *k*.

A useful reminder

Grimmer, Roberts & Stewart: *"All quantitative models of language are wrong, but some are useful."* LDA is a **reading aid**, not an oracle.

Looking Ahead

Final Assignment

Reproduce today's pipeline and interpret the topics — push to GitHub

This week's submission is your **final assignment** for the course.

1. **Use** one of the NIKH corpora from the *Data & Scripts* page — the 9-book or the 11-book sample.
2. **Reproduce** today's Orange pipeline: File → Corpus → Python Script (Kiwi) → Corpus → Preprocess Text → Topic Modelling → LDAvis.
3. **Interpret.** Pick one (or both):
   - A modification of the **LDAvis** view — adjust λ, select a topic, and describe what you see.
   - A reflection on the **topic output** — the labels you gave each topic, which eras use which, what surprised you.
4. **Submit** by pushing your work to GitHub this week.

Where This Leaves Us

A map of the semester so far

Week	Method	What it gives you
4–5	BoW / TF–IDF	descriptive word frequencies
7	Clustering (K-means, hierarchical)	one group per document
8	Word embeddings	meaning as vectors
9	Dictionary sentiment	a tone score per document
10	**LDA topic modeling**	**a theme mixture per document**

Recommended reading. Grimmer, Roberts & Stewart, Chapter 13: *Topic Models*.

Week 11: Final Review & Assessment

May 11 — covering Weeks 7–10

On the day:

- **10-question assessment** via Qualtrics (same platform as the midterm).
- **Application exercise.** Apply *at least two* of the methods you have learned in the second half of the course — clustering, word embeddings, sentiment, or LDA — to a question I will pose, using a dataset I will provide on the day.
- Come prepared to use *any* of them. You will choose which to combine once you see the task.
- Followed by a review of the term and a look ahead to the Research Methods Project.

Between today and May 11

We do not meet. I will email you more information about the Research Methods Project in the meantime.

Week 12: Research Methods Project Workshop

May 18 — our last meeting

A working session for your final paper.

- In-class assistance from me and peer support from your classmates.
- Bring whatever you are stuck on: a research question, a dataset, a draft, a half-finished pipeline.
- Troubleshoot together, plan the write-up, ask questions.
- This is our last chance to come together as a group.

What to expect

Attendance is expected. The environment is supportive — not an assessment. Come prepared to work on your project and to help a classmate with theirs.

Appendix: Other Topic Modeling Methods

Appendix: Other Topic Modeling Methods

LSI, HDP, NMF — three alternatives you may see in the literature

Method	How it works	When useful
LSI	A linear-algebra trick (SVD) on the term–document matrix. Produces “latent dimensions,” not probability distributions.	Search and information retrieval; fast, but topics are hard to read.
NMF	Factorises the (non-negative) term–document matrix into a word part and a document part. Not probabilistic, but topics often look clean.	Often produces sharper, more coherent topics than LDA; common in recent NLP work.
HDP	A non-parametric cousin of LDA: you don’t pick *k* — the model infers the number of topics from the data.	When you genuinely don’t know how many topics to ask for.

All three are available in Orange’s Topic Modelling widget. We stay with LDA in this course because it produces probability mixtures that are easy to teach and because you pick *k* deliberately.

Appendix: Quick Comparison

How each method answers the same two questions LDA answers

	LDA	LSI	NMF
Probabilistic?	yes	no	no
Output: topic–word	$\phi_k(w)$, probabilities	loadings	non-negative weights
Output: doc–topic	$\theta_d(k)$, mixture	coordinates	non-negative weights
You choose *k*?	yes	yes	yes

HDP sits outside this table because *k* is inferred rather than chosen. Its output is still a topic mixture per document, like LDA.

Practical advice. If LDA gives you incoherent topics no matter what *k* you choose, NMF is the obvious next thing to try. It uses the same preprocessing pipeline.