Text as Data: Topic Models

Guest Course – January 2026

Germain Gauthier, Philine Widmer[1]

[1]Bocconi Unversity, Paris School of Economics

USI Lugano

Today: Unsupervised learning with topic models

- **Supervised learning** (yesterday): predict labels from text
  - Great for prediction once target is defined
  - Limited for discovery: "What are the themes in this corpus?"
- **Unsupervised learning** (today): discover latent structure
  - No labels required
  - Goal: find interpretable patterns/topics in text
- Focus: **Latent Dirichlet Allocation (LDA)**[1]
  - Matrix decomposition perspective
  - Generative model interpretation
  - Estimation and hyperparameters
  - Applications in economics and social sciences

Recall: Supervised text classification

Yesterday we had:

- Documents $i = 1, \ldots, n$
- Labels y_i (e.g., sentiment, political party)
- Features x_i (bag-of-words, tf-idf)
- Goal: learn $f(x_i) \to y_i$

Today: **No labels!**

- Same documents, same features
- Goal: discover latent themes/topics that explain word patterns
- Output: interpretable groupings of words (topics) and documents

What is a "topic"?

Intuitively, a **topic** is a recurring pattern of co-occurring words.

Examples:

- **Topic 1 (Economics)**: *growth, inflation, GDP, unemployment, economy*
- **Topic 2 (Politics)**: *election, vote, party, government, president*
- **Topic 3 (Health)**: *patient, hospital, treatment, medical, doctor*

Formally, a topic is a **distribution over words**.

- Each topic assigns probability to every word in vocabulary
- High-probability words characterize the topic

Why topic models?

Applications:

- **Exploratory data analysis**: What are documents about?
- **Dimensionality reduction**: Represent documents by topic mixtures instead of high-dimensional word counts
- **Feature extraction**: Use topic proportions as features for downstream tasks (e.g., regression, classification)

Economics/social science examples:

- Policy documents: identify issue dimensions
- Congressional speeches: track political agendas
- Central bank communications: detect shifts in policy focus

The document-term matrix

Recall the bag-of-words representation:

- n documents, V vocabulary size
- $X \in \mathbb{N}^{n \times V}$: each entry X_{ij} = count of word j in document i

	word 1	word 2	$\cdots$	word V
doc 1	5	0	$\cdots$	2
doc 2	1	8	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
doc n	0	3	$\cdots$	1

Problem: X is high-dimensional ($V \sim 10^4$) and sparse.

Matrix decomposition perspective

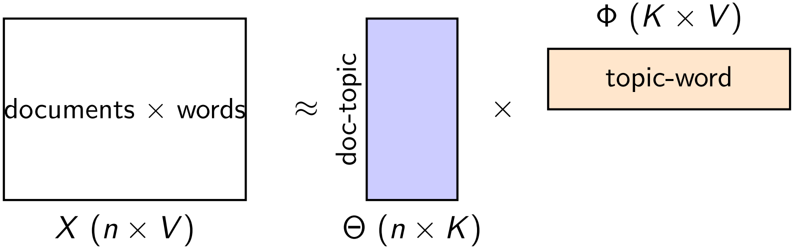
Key idea: Approximate X as a product of two lower-dimensional matrices.

$$\underbrace{X}_{n \times V} \approx \underbrace{\Theta}_{n \times K} \times \underbrace{\Phi}_{K \times V}$$

where $K \ll V$ (e.g., $K = 10$–100 topics).

- Θ: **document-topic matrix**
  - $\theta_i = (\theta_{i1}, \ldots, \theta_{iK})$: topic proportions in document i
  - $\sum_{k=1}^{K} \theta_{ik} = 1$, $\theta_{ik} \geq 0$
- Φ: **topic-word matrix**
  - $\phi_k = (\phi_{k1}, \ldots, \phi_{kV})$: word distribution for topic k
  - $\sum_{v=1}^{V} \phi_{kv} = 1$, $\phi_{kv} \geq 0$

Visualizing the decomposition



Each document is a **mixture of topics**, each topic is a **distribution over words**.

LDA: Generative model

LDA posits the following generative process for each document i:

1. Draw topic proportions: $\theta_i \sim \text{Dirichlet}(\alpha)$
2. For each word position $j = 1, \ldots, N_i$ in document i:
   2.1 Draw a topic: $z_{ij} \sim \text{Categorical}(\theta_i)$
   2.2 Draw a word: $w_{ij} \sim \text{Categorical}(\phi_{z_{ij}})$

Parameters:

- $\alpha \in \mathbb{R}_+^K$: Dirichlet prior for document-topic distributions
- $\beta \in \mathbb{R}_+^V$ (or η): Dirichlet prior for topic-word distributions
- $\phi_k \sim \text{Dirichlet}(\beta)$ for each topic $k = 1, \ldots, K$

What does LDA learn?

Given a corpus (observed word counts), LDA inference produces:

1. **Topic-word distributions** ϕ_k for $k = 1, \ldots, K$
   - Each topic's vocabulary signature
   - Typically display top 10–20 words per topic
2. **Document-topic distributions** θ_i for $i = 1, \ldots, n$
   - What topics are present in each document?
   - Can be used as features for downstream tasks

Inference problem

Goal: Given observed words $\mathbf{w}$, infer latent variables $\theta, \phi, \mathbf{z}$.

Posterior distribution:

$$p(\theta, \phi, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\mathbf{w}, \theta, \phi, \mathbf{z} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}$$

Problem: The denominator (marginal likelihood) is intractable.

$$p(\mathbf{w} \mid \alpha, \beta) = \int_{\theta,\phi} \sum_{\mathbf{z}} p(\mathbf{w}, \theta, \phi, \mathbf{z} \mid \alpha, \beta)\, d\theta\, d\phi$$

Summing over all possible topic assignments $\mathbf{z}$ is exponential in document length.

Two main inference methods

1. **Variational Inference** (Blei, Ng, Jordan 2003)
   - Approximate posterior with simpler distribution $q(\theta, \phi, \mathbf{z})$
   - Minimize KL divergence: $\mathrm{KL}(q \| p)$
   - Fast, deterministic
   - Used in: `gensim`, `sklearn`
2. **Gibbs Sampling** (Griffiths & Steyvers 2004)
   - MCMC method: iteratively sample topic assignments z_{ij}
   - Integrate out θ, ϕ (collapsed Gibbs sampling)
   - Slower, but often more accurate
   - Used in: `MALLET`, `tomotopy`

Hyperparameters: α (document-topic)

α controls how many topics each document uses.

Small α (e.g., 0.1): Sparse topic mixtures

- Each document uses few topics
- More interpretable (documents are "about" one or two things)
- Default in many implementations: $\alpha = 50/K$

Large α (e.g., 10): Dense topic mixtures

- Documents use many topics
- Less interpretable
- May be appropriate for very short documents

Hyperparameters: β (topic-word)

β (sometimes η) controls how many words each topic uses.

Small β (e.g., 0.01): Sparse word distributions

- Each topic concentrated on few words
- More interpretable topics
- Default in many implementations: $\beta = 0.01$ or $\beta = 1/V$

Large β (e.g., 1.0): Dense word distributions

- Topics spread over many words
- Less distinct topics
- Rarely used

Choosing the number of topics K (1/2)

No single correct answer! Trade-offs:

Small K **(e.g.,** 5–10**)**:

- Broad, general topics
- Easier to interpret
- May miss fine-grained distinctions

Large K **(e.g.,** 50–100**)**:

- More specific topics
- Captures more detail
- Harder to interpret, potential redundancy

Choosing the number of topics K (2/2)

Approaches:

- **Perplexity**: held-out log-likelihood (often keeps increasing with K)
- **Coherence**: do top words co-occur in documents? (better metric)
- **Human evaluation**: read topics, pick K that makes sense
- **Sensitivity analysis**: try multiple K, compare results

Interpreting topics

1. **Top words**: Look at highest-probability words in ϕ_k
   - Typically display top 10–20 words
   - Do they cohere? Can you give the topic a label?
2. **Representative documents**: Which documents have high θ_{ik}?
   - Read documents where topic k is dominant
   - Validates topic interpretation

Using topics for downstream tasks

Topics as features for prediction:

Example: Predict stock returns from earnings call transcripts

1. Run LDA on all transcripts $\rightarrow$ get θ_i for each document
2. Use θ_i as features in regression: $\text{return}_i = \beta^\top \theta_i + \varepsilon_i$
3. Interpret: which topics predict positive/negative returns?

Advantages:

- Lower-dimensional representation ($K \ll V$)
- Interpretable features (topic $=$ theme)
- Can capture semantic similarity (documents with similar topics)

LDA variants and extensions

Supervised LDA (sLDA)[2]:

- Include document-level response variable in the model
- Topics optimized for prediction, not just description

Structural Topic Model (STM)[3]:

- Include document-level covariates (e.g., author, year)
- Topic prevalence and content can vary with covariates

Summary: Topic models

Key concepts:

- Topic models discover latent themes in text collections
- Documents = mixtures of topics, topics = distributions over words
- Inference via variational methods or Gibbs sampling

Hyperparameters:

- K: number of topics (most important choice!)
- α: controls sparsity of document-topic distributions (default: $50/K$)
- β: controls sparsity of topic-word distributions (default: 0.01)

Practical advice:

- Always inspect topics qualitatively
- Try multiple values of K

Next: Word embeddings

Today: Topics = discrete mixtures

Next session: Word embeddings = continuous representations

- Represent words as vectors in $\mathbb{R}^d$ (e.g., $d = 100$–300)
- Semantic relationships: $\vec{king} - \vec{man} + \vec{woman} \approx \vec{queen}$
- Learn from word co-occurrence (Word2Vec, GloVe)
- Foundation for modern NLP (precursor to transformers)

Topics and embeddings are complementary:

- Topics: interpretable themes, document-level
- Embeddings: semantic similarity, word-level

References I

- Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.
- Mcauliffe, Jon and David Blei (2007). "Supervised topic models". In: *Advances in neural information processing systems* 20.
- Roberts, Margaret E, Brandon M Stewart, and Edoardo M Airoldi (2016). "A model of text for experimentation in the social sciences". In: *Journal of the American Statistical Association* 111.515, pp. 988–1003.