Latent Dirichlet Allocation

CPSC 503 - Pedagogical Project Final Presentation

Nazlı Özüm Kafaee

Outline

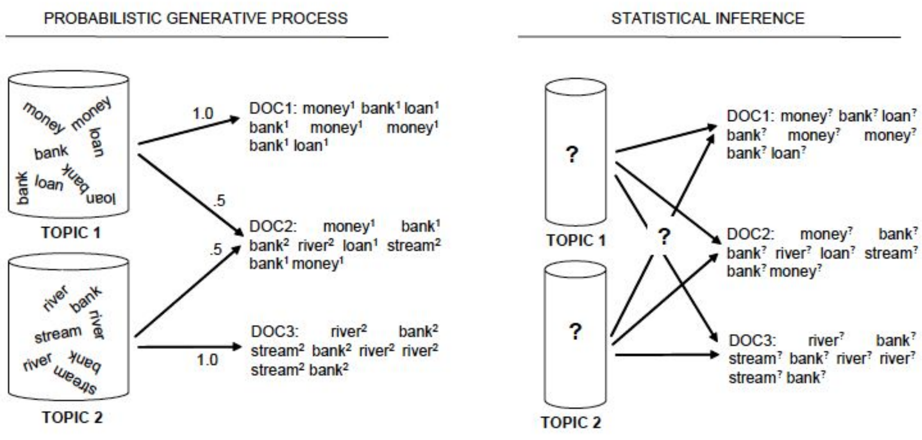
1. Motivation
2. Probabilistic Topic Modelling
3. Brief History
4. Latent Dirichlet Allocation (LDA)
    a. Intuition
    b. The Posterior Distribution
    c. Posterior Inference with Gibbs Sampling
    d. Evaluation
5. Application Examples

Motivation

- We can use topic models for
    - Data exploration
    - Information Retrieval
    - Classification / prediction
    - Searching for relevant documents
- Latent Dirichlet Allocation is the **most widely used** topic modelling method
- Latent Dirichlet Allocation by Blei et al.(2003) is one of the **most cited** machine learning papers

Probabilistic Topic Modelling

1. Treat data as observations that arise from a **generative process** that includes hidden variables
2. **Infer the hidden structure** using posterior inference
3. **Situate new data** into the estimated model



Source: Steyvers, M., & Griffiths, T. (2006). Probabilistic Topic Models. In Latent Semantic Analysis: A Road to Meaning (p. 15).

History building up to LDA

- **Latent Semantic Indexing (LSI)** → not a generative model
  - Summarize each document by its TF-IDF values
  - Run Singular Value Decomposition (SVD) on TF-IDF matrix to reduce dimension
  - Treat the principal components as the “topics”
- **Probabilistic LSI (Aspect Model)**
  - Introduced as an alternative to LSI
  - Each word w as a sample from a mixture model

$$P(w \mid d) = \sum_{z \in Z} P(w \mid z) P(z \mid d)$$

  - Mixture components are multinomial random variables z that can be viewed as "topics"
  - No probabilistic model at the level of documents

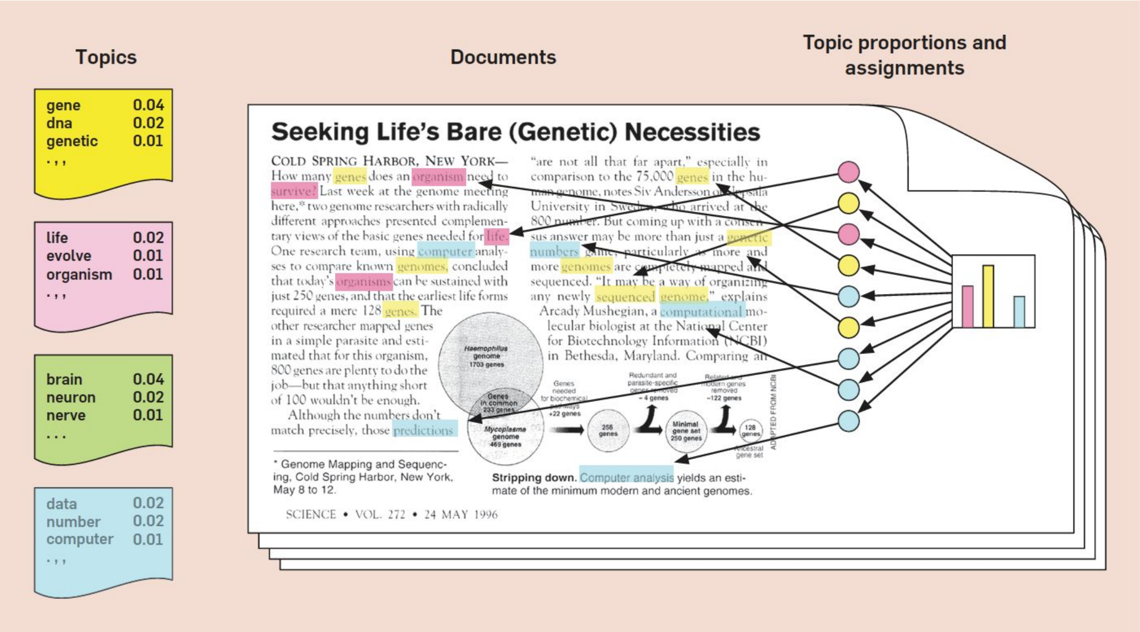
Latent Dirichlet Allocation (LDA)

An **extension of pLSI** bringing a solution to the computation of per-document topic distributions (θ)

A **hierarchical Bayesian model** of **each word** in a document

- Puts a prior on $\theta \rightarrow$ conjugate prior is the Dirichlet distribution

The intuition behind LDA

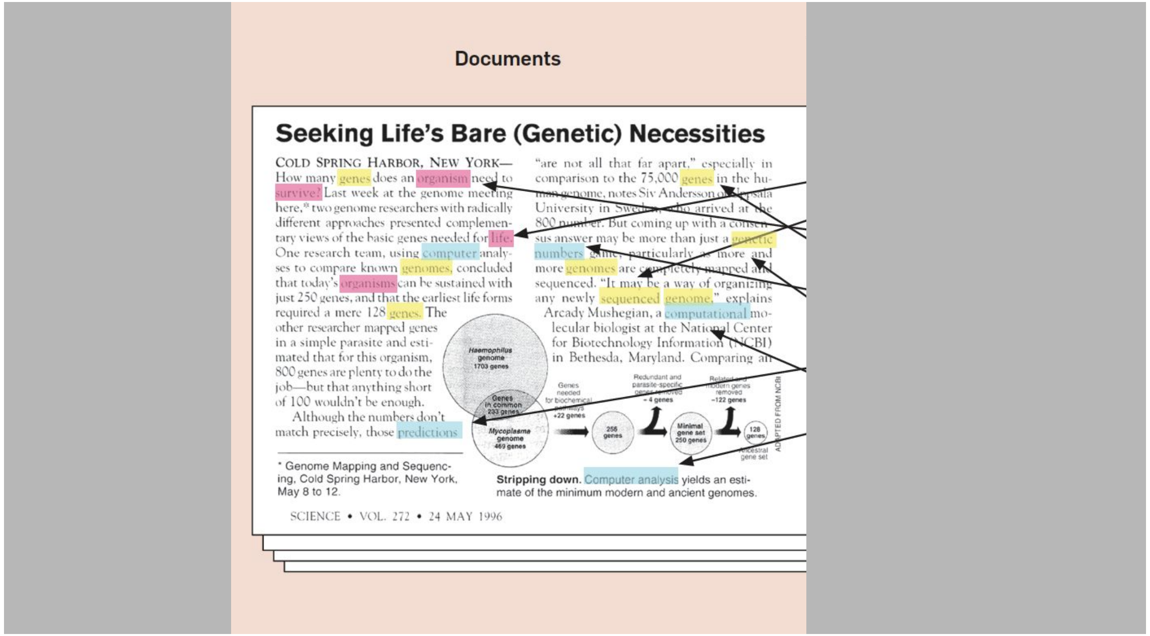


Each document is a random mixture of corpus-wide topics.

Each word is drawn from one of those topics.

Source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77. https://doi.org/10.1145/2133806.2133826

The intuition behind LDA



In **reality**,

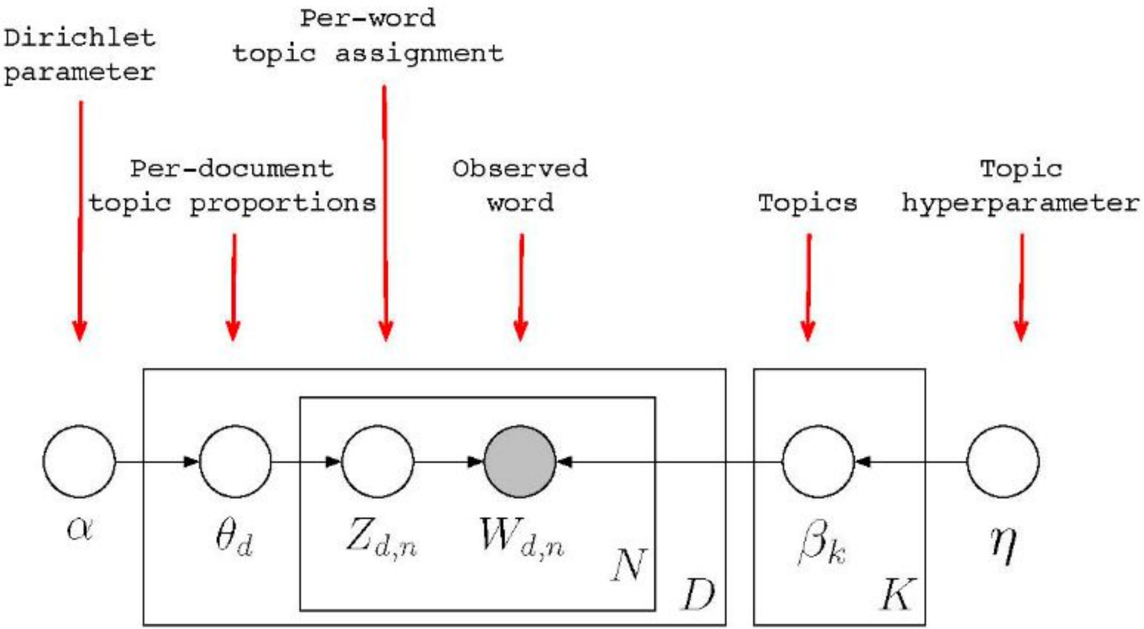
we only **observe the documents**

and

aim to **infer the topic structure**.

Source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77. https://doi.org/10.1145/2133806.2133826

Latent Dirichlet Allocation (LDA)



Source: Topic Model lecture by David Blei on http://videolectures.net/mlss09uk_blei_tm/

Approximate Posterior Inference

Given observations (words in documents), we want to **infer the hidden structure (topics)** from the posterior distribution.

$$p(\underbrace{Z}_{\text{topics}}, \underbrace{\pi}_{\text{word prob.}}, \overbrace{\theta}^{\text{topic prop.}} \mid \underbrace{X}_{\text{data (words)}}, \underbrace{\alpha}_{\text{prior on topic proportions}}, \underbrace{\beta}_{\text{prior on word probabilities}}) = \left[\prod_{i=1}^{n} \underbrace{p(\theta^i \mid \alpha)}_{\text{topic proportion probability (document 'i')}} \prod_{j=1}^{d} \underbrace{p(z_j^i \mid \theta^i)}_{\text{topic probability (topic at position 'j' in document 'i')}} \underbrace{p(x_j^i \mid z_j^i, \pi_j)}_{\text{word probability (word at position 'j' in document 'i')}}\right]\left[\prod_{c=1}^{k} \underbrace{p(\pi_c \mid \beta)}_{\text{word probability parameters (topic 'c')}}\right]$$

Source: CPSC 540 slides by Mark Schmidt at https://www.cs.ubc.ca/~schmidtm/Courses/540-W18/L33.pdf

Posterior probability is **computationally intractable** → approximate inference

- Sampling-based algorithms → MCMC (Markov Chain Monte Carlo)
  - Gibbs sampling
- Variational algorithms → optimization
  - Variational EM algorithm

Approximate inference: Gibbs Sampling

- Define a **Markov chain** → stationary distribution is the posterior
- Collect **independent samples** from the stationary dist.
- The space of the MC is the space of possible configurations of the hidden variables
  - The chain is run by **iteratively sampling** from the conditional dist. of each hidden variable **given observations and the current state of the other hidden variables**
- Once a chain is burned in, collect samples at a lag to approximate the posterior

Evaluation: Perplexity

The **most typical evaluation** of topic models

The predicted number of equally likely words for a word position on average

- A monotonically decreasing function of the log-likelihood → **lower perplexity** over a held-out document → higher log-likelihood → **better predictive performance**

$$perplexity(D_{\text{test}}) = \exp\left(- \sum_{d=1}^{M} \log p(\mathbf{w}_d) \Big/ \sum_{d=1}^{M} N_d \right)$$

Applications in Informations Systems (IS)

- Developing an automated system to raise **red flags for financial fraud based on social media posts of companies.** → Dong, Wei, Shaoyi Liao, and Zhongju Zhang. 2018. "Leveraging Financial Social Media Data for Corporate Fraud Detection." Journal of Management Information Systems 35 (2): 461–87. doi:10.1080/07421222.2018.1451954.
- Analyzing how the **topic discussions in Denial of Service Attack (DDoS) forums** can predict actual DDoS attacks. → Yue, Wei T., Qiu-Hong Wang, and Kai-Lung Hui. 2019. "See No Evil, Hear No Evil? Dissecting the Impact of Online Hacker Forums." MIS Quarterly 43 (1): 73–95. doi:10.25300/MISQ/2019/13042.
- Examining how personality traits of social media users affect the effectiveness of word-of-mouth (WOM), using **LDA to measure for the similarity of interests and topics discussed in social media posts** by the recipient and the sender. → Adamopoulos, Panagiotis, Anindya Ghose, and Vilma Todri. 2018. "The Impact of User Personality Traits on Word of Mouth: Text-Mining Social Media Platforms." Information Systems Research 29 (3): 612–40. doi:10.1287/isre.2017.0768.

References

- Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," Journal of Machine Learning Research (3), pp. 993-1022.
- CPSC 540: Machine Learning slides by Mark Schmidt on Topic Models
- University of Waterloo lecture slides