Latent Dirichlet Allocation

CPSC 503 - Pedagogical Project Final Presentation

Nazlı Özüm Kafaee

Outline

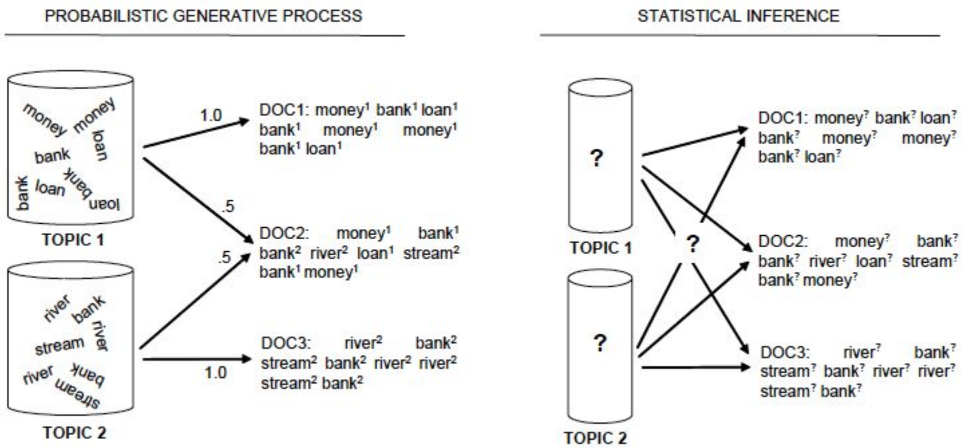
1. Motivation
2. Probabilistic Topic Modelling
3. Brief History
4. Latent Dirichlet Allocation (LDA)
    a. Intuition
    b. The Posterior Distribution
    c. Posterior Inference with Gibbs Sampling
    d. Evaluation
5. Application Examples

Motivation

- We can use topic models for
    - Data exploration
    - Information Retrieval
    - Classification (or prediction)
    - searching for relevant documents
- LDA 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

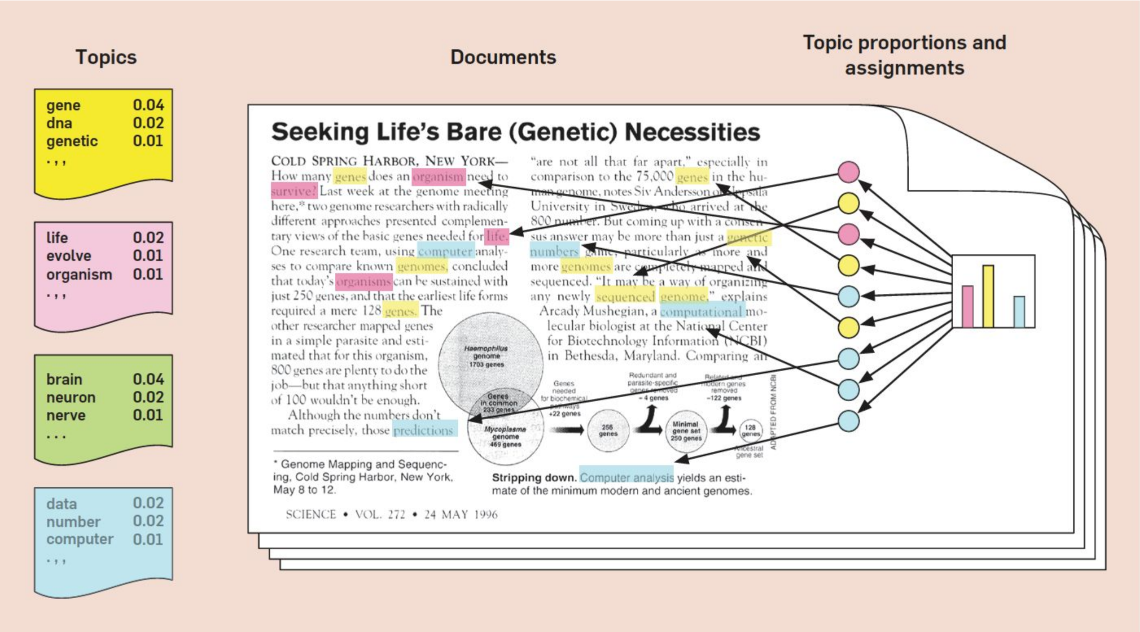
Latent Dirichlet Allocation (LDA)

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 θ → conjugate prior is the Dirichlet distribution

The intuition behind LDA

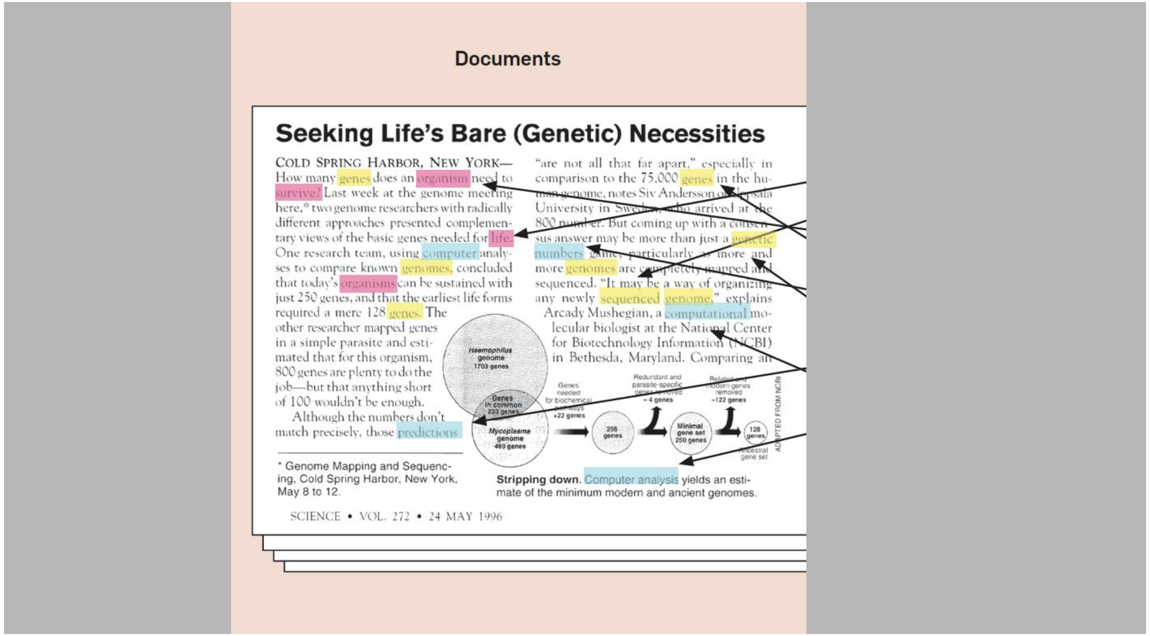


Each document is a random mixture of corpus-wide topics.

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



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

In **reality**,

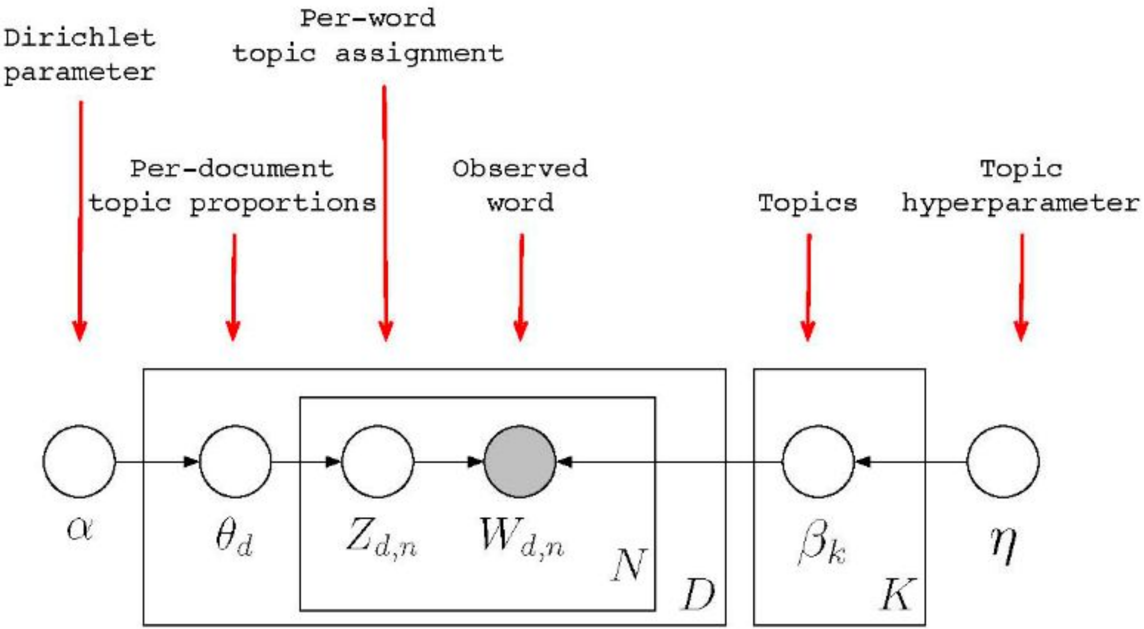
we only <u>observe the documents</u>

and

aim to <u>infer the topic structure</u>.

<u>Source:</u> 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.

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

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 with MCMC

With Gibbs sampling we alternate between:

- Sampling topics given word probabilities and topic proportions.
- Sampling topic proportions given topics and prior parameters α.
- Sampling word probabilities given topics, words, and prior parameters β.

Have a burn-in period, use thinning, try to monitor convergence, etc.

Finally we use posterior samples to do inference:

- Distribution of topic proportions for sample ‘i’ is frequency in samples.
- To see if words come from same topic, check frequency in samples.

Evaluation: Perplexity

The most typical evaluation of topic models

The predicted # 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

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

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 whether personality traits of social media users attenuate or accentuate the effectiveness of word-of-mouth (WOM) → use LDA model 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