Data Science for Digital Humanities 2

Topic Modeling

Prof. Dr. Goran Glavaš
Lennart Keller

May 22, 2023

Topic Modeling

Topic modeling is a type of **statistical modeling** that uses **unsupervised machine learning** to identify clusters or groups of similar words within a body of text.

Levity.ai

In statistics and natural language processing, a **topic model** is a type of **statistical model for discovering the abstract "topics"** that occur in a **collection of documents**. Topic modeling is a frequently used text-mining tool for **discovery of hidden semantic structures in a text body**.

Wikipedia

Topic modeling

- A (set of) explorative data analysis technique(s) for a body of text data

- In essence:
  - You have a (large) collection of (unannotated) documents
  - What is in there? What „topics“ are prominent?
  - Important: very different from document clustering!

- **Topic models**
  - Topics as probability distributions over words
  - Documents as probability distributions over topics

Latent Dirichlet Allocation

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993-1022.

- The most widely used topic model is Latent Dirichlet Allocation
- LDA is a probabilistic framework

Levity.ai

Multinomial Distribution

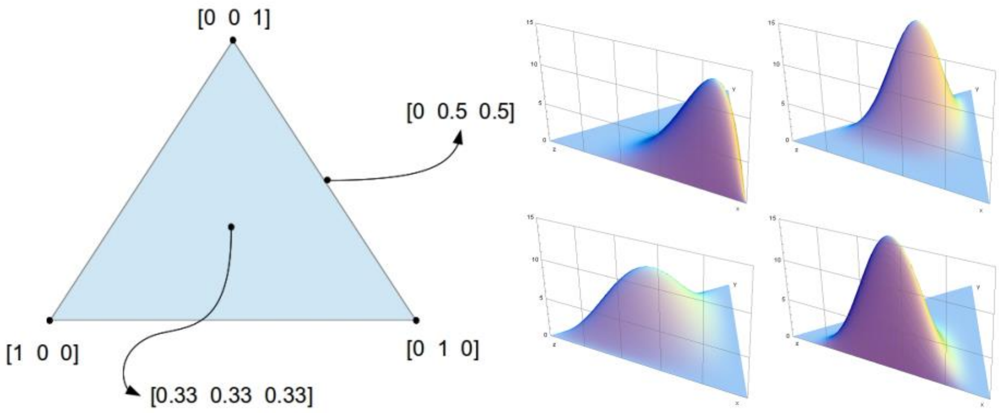
- A **multinomial (categorical) distribution** is a probability distribution over a discrete (finite) set of possible events

- The multinomial distribution over N terms, which we denote with $Mult_K(\boldsymbol{\vartheta})$ is parametrized by the vector $\boldsymbol{\vartheta}$ of N – 1 probabilities
  - Probabilities of the distribution must sum to 1, so we can compute the last probability from the given N – 1

- Each topic in LDA is going to be one multinomial distribution over all vocabulary words

Dirichlet Distribution

- **Dirichlet distribution** is a probability distribution over all vectors of length K that sum up to 1
  - A meta-distribution, a probability distribution over multinomial distributions
  - Denoted with $Dir_K(\alpha)$ Dirichlet distribution is parametrized with a parameter vector α
  - A sample $\boldsymbol{\vartheta}$ drawn from the Dirichlet distribution $Dir_K(\alpha)$ can be used to parametrize the multinomial distribution – $Mult_K(\boldsymbol{\theta})$



Latent Dirichlet Allocation

- **Latent Dirichlet Allocation (LDA)** is a latent topic model that assumes that the collection of documents was generated by a particular Dirichlet distribution
  - Collection of M documents (or „contexts”), vocabulary of N terms, K latent topics

- Each of the K latent topics is a concrete multinomial distribution over terms
- For each position in each of the M document we obtain the observed word by:
  1. Randomly selecting one of the topics (from the Dirichlet distribution)
  2. Randomly select the term from the multinomial distribution of the topic that was randomly selected in the step 1

- Vocabulary of N terms
  - Each topic is a concrete multinomial distribution with N – 1 parameters

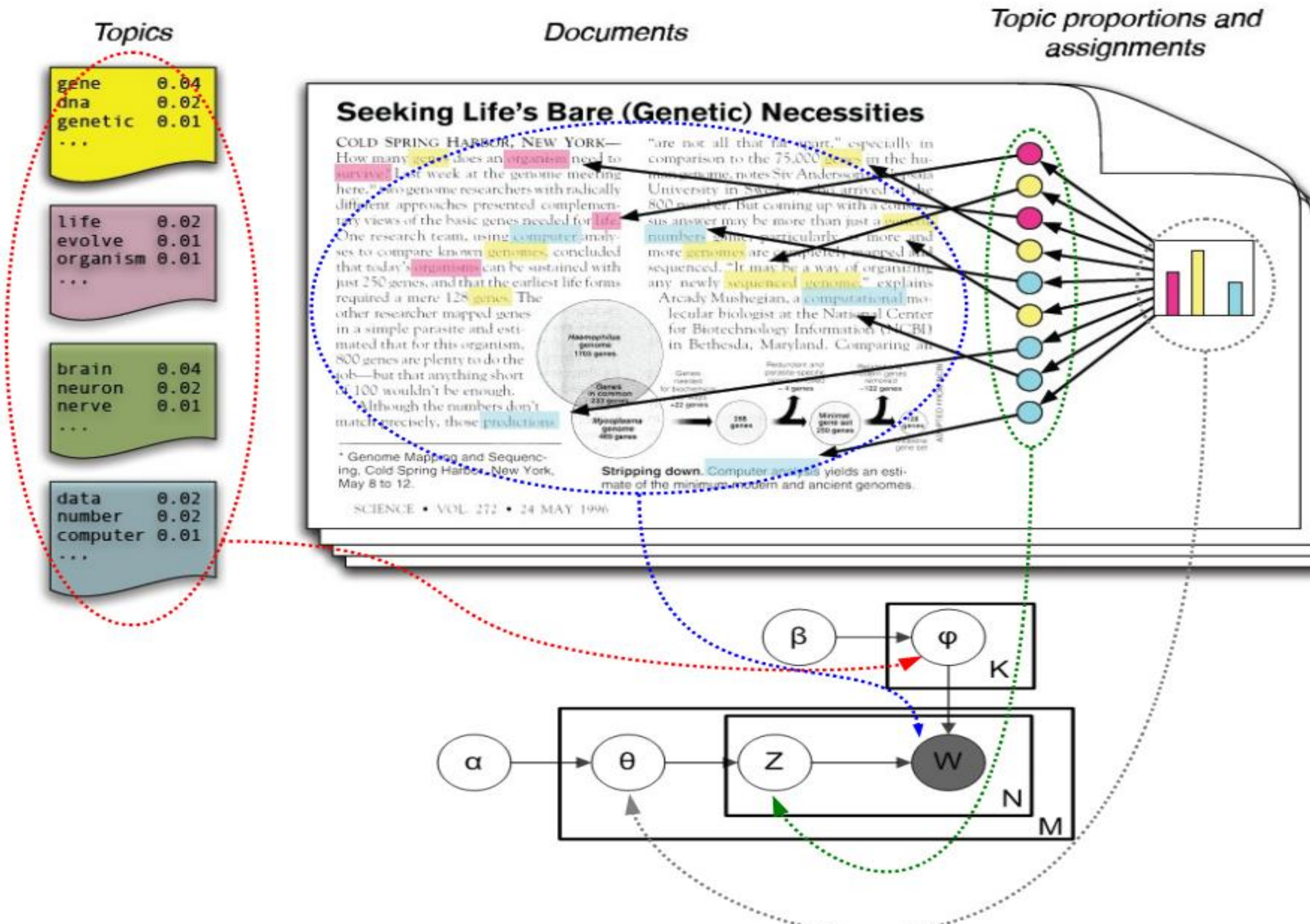
LDA – Generative View

1. For each topic k (k = 1, ..., K):
   - Draw parameters of a multinomial distribution $\boldsymbol{\varphi}_k$ (over terms) for topic k from a Dirichlet distribution $Dir_N(\beta)$

2. For each document d in the collection:
   - Draw parameters of a multinomial distribution of topics for the document d, $\boldsymbol{\theta}_d$, from a Dirichlet distribution $Dir_K(\alpha)$

   - For each term position w_{dn} in the document d:
     a) Draw a topic assignment (i.e., a concrete multinomial distribution over terms) z_{dn} from $Mult_K(\boldsymbol{\theta}_d)$
     b) Draw a concrete term w_{dn} from the multinomial distribution over terms of the topic z_{dn} (drawn in a)), $Mult_N(\boldsymbol{\varphi} z_{dn})$

LDA – Generative View



LDA – Parameters and estimation

- Parameters of the LDA are variables/probabilities that we cannot directly observe
- Probabilities of all multinomial distributions that are sampled in the generative algorithm
    1. Term probabilities (vector of N probabilities) for each of the K latent topics
    φ_k for k = 1, ..., K (so, total of K * N(-1) parameters)
    2. Topic probabilities (vectors of K probabilities) for each of the M documents
    θ_d for d = 1, ..., M (so, total of M * K(-1) parameters)

- **Optimization** (learning model's parameters):
    1. Start from random multinomial distributions
    2. Update parameters to maximize probability of observed terms in documents
    - Direct maximization is intractable
    - Approximate inference (maximization): (1) variational or (2) sampling methods

Latent Dirichlet Allocation

- Once the model is trained (parameters optimized based on observed text), we represent documents/contexts and terms as follows:

1. Document d – simply the multinomial distribution vector over topics for that documents, $\boldsymbol{\theta}_d$
2. Topic k – simply the learned multinomial distribution over terms
3. Term t_i (i = 1, ..., N) – for each of the K topics we take the probability of t_i from the multinomial distribution (over terms) of that topic $[\boldsymbol{\varphi}_k]^i$, that is term's probability from multinomial distributions of all topics

Latent Dirichlet Allocation

- The topics are generally interpretable – the terms with largest probabilities within the multinomial distribution of the topic tend to be semantically related
- Example – topics obtained on 1.8M New York times articles:

music	book	art	game	show
band	life	museum	knicks	film
songs	novel	show	nets	television
rock	story	exhibition	points	movie
album	books	artist	team	series
jazz	man	artists	season	says
pop	stories	paintings	play	life
song	love	painting	games	man
singer	children	century	night	character
night	family	works	coach	know

theater	clinton	stock	restaurant	budget
play	bush	market	sauce	tax
production	campaign	percent	menu	governor
show	gore	fund	food	county
stage	political	investors	dishes	mayor
street	republican	funds	street	billion
broadway	dole	companies	dining	taxes
director	presidential	stocks	dinner	plan
musical	senator	investment	chicken	legislature
directed	house	trading	served	fiscal

Questions?


Awọn ibeere?
ਸਵਾਲ?
Küsimusi?
Turite klausimų?
Pitanja?
Sorusu olan?
Dúvidas?
質問は？
有问题吗？
Fragen?
¿Preguntas?
Questions?
Frågor?
Pytania?
Domande?
Питання?
Vragen?
Ερωτήσεις;
Porandukuéra?
คำถาม?
أسئلة؟

Homework #1

1. Find an interesting corpus of documents
    - E.g., collection of poems of some poet
    - Set of chapters or sections from a book you like
    - Collection of movie reviews
    - …
2. Preprocess the corpus (lemmatize, remove stopwords, etc.)
3. Run LDA (with gensim) on this document collection
    - Play with different number of topics
4. Make a thorough analysis of induced topics
5. Present your results in a couple of weeks