Distributional Semantics

Computational Methods for Text Analysis

Alena Pestova

HSE Saint-Petersburg

08.11.2022

Topic modeling: motivation

Suppose you're given a massive corpora and asked to carry out the following tasks:

- Organize the documents into thematic categories
- Describe the evolution of those categories over time
- Enable a domain expert to analyze and understand the content
- Find relationships between the categories
- Understand how authorship influences the content

Topic Modeling: A method of (usually unsupervised) discovery of latent or hidden structure in a corpus

- Applied primarily to text corpora, but techniques are more general
- Provides a modeling toolbox
- Has prompted the exploration of a variety of new inference methods to accommodate large-scale datasets

From words to topics: Multinomial distribution over words

Original Text

Карл у Клары украл кораллы, Клара у Карла украла кларнет.

Словарь (распределение)

Карл	у	Клара	украсть	коралл	кларнет
Карл	у	Клара	украсть		
2	2	2	2	1	1
0.2	0.2	0.2	0.2	0.1	0.1

Corpus as a mixture of topics (distributions)

Topics — events

Topic							всего
	Карл	у	Клара	украсть	коралл	кларнет	
Карл	0,1	0,1	0,1	0,1	0,1	0	**0,5**
Клара	0,1	0,1	0,1	0,1	0	0,1	**0,5**

Topics — common and differences

Тема							всего
	Карл	у	Клара	украсть	коралл	кларнет	
Общее	0,2	0,2	0,2	0,2	0	0	**0,8**
Разное	0	0	0	0	0,1	0,1	**0,2**

Probabilistic topic models

We will look at

- pLSA (Probabilistic Latent semantic analysis)
- LDA (Latent Dirichlet Analysis)

Some abbreviations

- D - set of the documents d(collection of docs, the corpus)
- W - set of all words w(vocabulary)
- T - set of topics t(some latent variable, we do not know)

Topic:

- latent variable (describes words distribution in a corpus)
- represents multinomial distribution over words

we want to find probabilites: $p(w|t)$ and $p(t|d)$

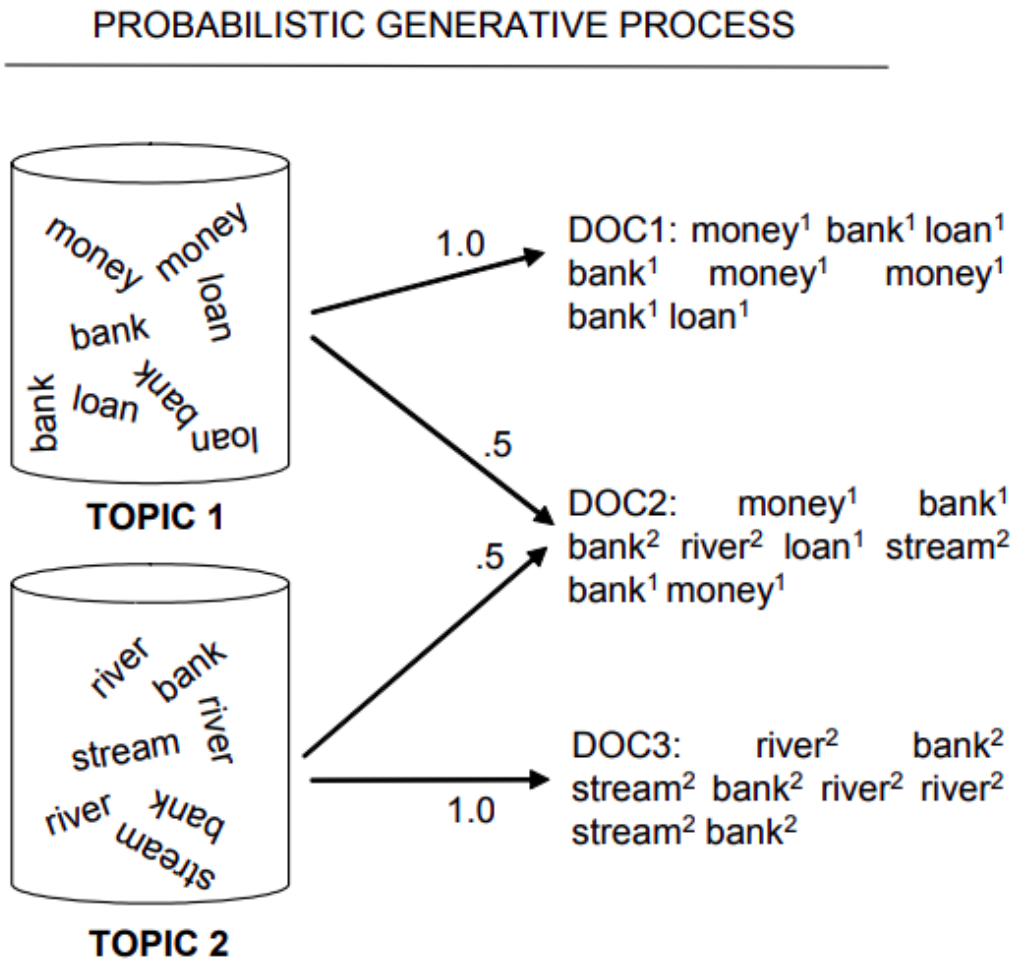
- toy example of topic family as a distribution over words (p(w|topic=family)): [p(dad)=0.3, p(mom)=0.35, p(son)=0.1, p(daughter)=0.1, p(home)=0.25, p(all other words)=0]
- example of a document as a distribution over topics (p(t|doc1)): [p(family topic)=0.5, p(work topic)=0.3, p(hobby topic)=0.2, p(all other topics)=0]

Generative process

if we know distributions $p(w|t)$ and $p(t|d)$, then we can generate a document:

- sample a topic for a document from $p(t|d)$
- sample words from $p(w|t)$

The generative process

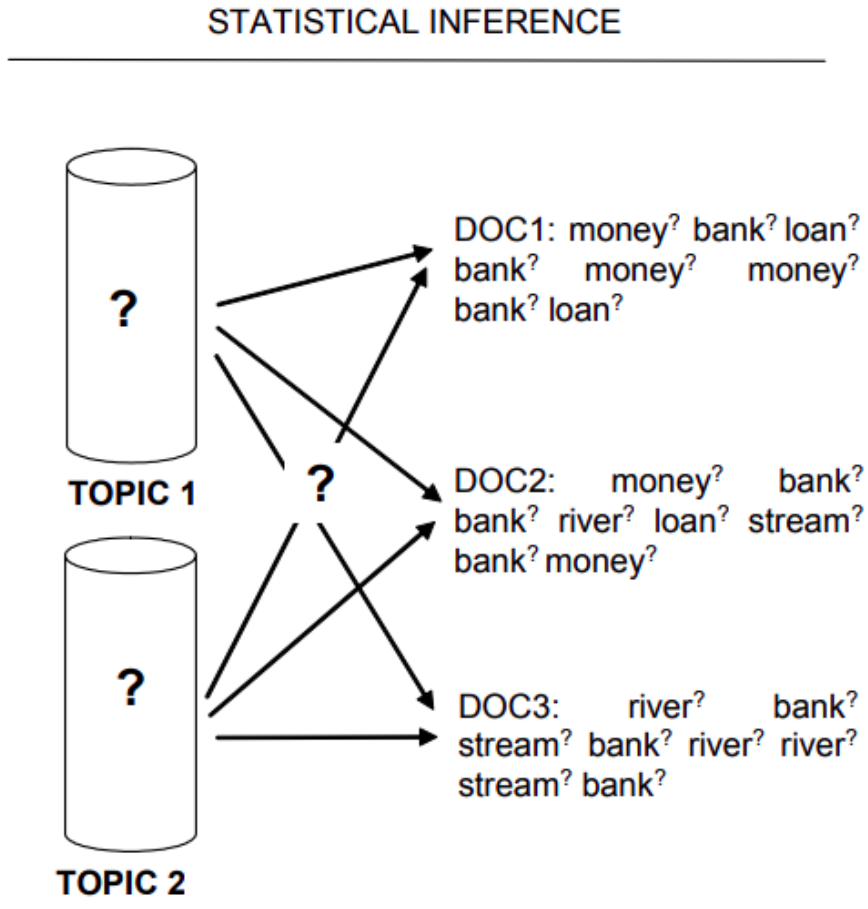


Steyvers, M. & Griffiths, T. (2006). Probabilistic topic models. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum

But we do not know $p(w|t)$ and $p(t|d)$, and we can only see W and D in our data (we have some corpus that consists of some number of documents).
And our task is to find this distributions $p(w|t)$ and $p(t|d)$.
If we find them, then

- $p(w|t)$ will help us to understand the topics, their interpretation
- $p(t|d)$ will help us to understand the documents - what topics are presented in some doc, in what proportions.

The inference problem



Steyvers, M. & Griffiths, T. (2006). Probabilistic topic models. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning.* Laurence Erlbaum

Basic probabilistic topic models

Basic probabilistic topic models are based in the following assumptions:

- the order of documents in the collection is not important
- the order of words in the document is not important, document is — «bag of words»;
- words found in most documents are not important for defining topics, they are usually excluded from the dictionary and called stop words;
- a word in different forms is one and the same word;

Basic probabilistic topic models

Basic probabilistic topic models are based in the following assumptions:

- a collection of documents can be viewed as a simple selection of document-word pairs (d, w), $d \in D$, $w \in W_d$.
- each topic $t \in T$ is described by an unknown distribution $p(w|t)$ on the set of words $w \in W$;
- each document $d \in D$ is described by an unknown distribution $p(t|d)$ on the set of topics $t \in T$;
- conditional independence hypothesis: $p(w|t, d) = p(w|t)$.
- To build a topic model means to find the matrices $\Phi = ||p(w|t)||$ and $\Theta = ||p(t|d)||$ from collection D.

Probabilistic Latent semantic analysis (pLSA)

Hoffman 1999
Topic:

- latent variable (describes words distribution in a corpus)
- represents multinomial distribution over words

Generative model pLSA

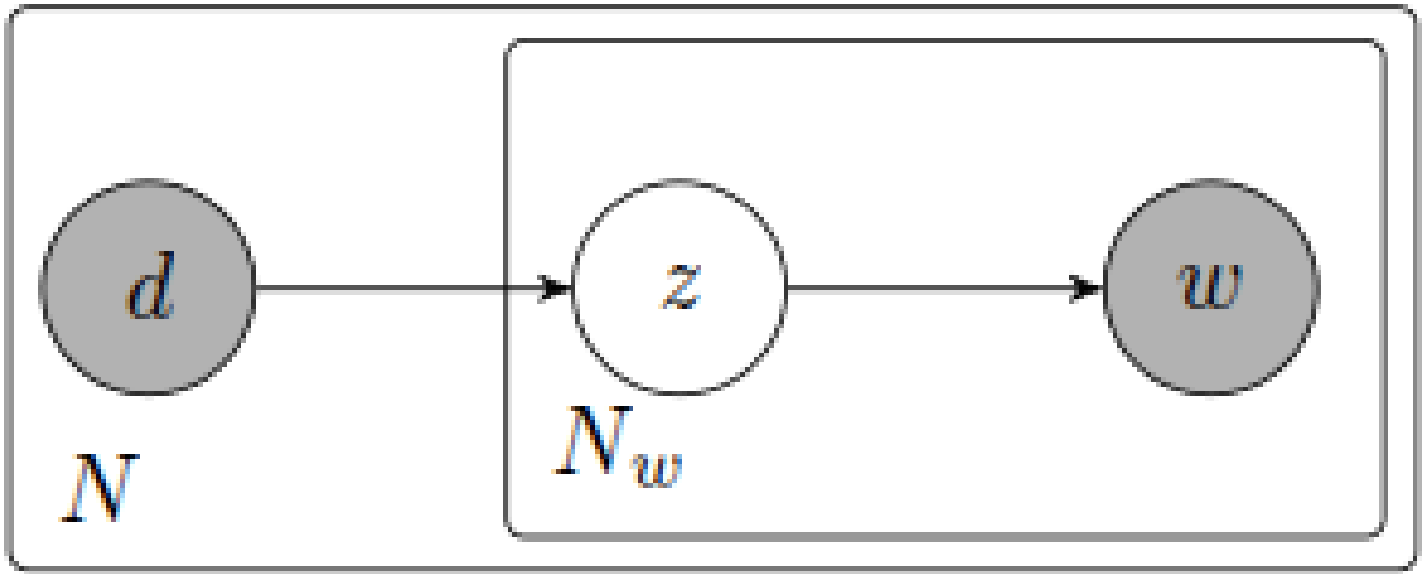
For each word in each document:

1. Select <randomly> topic z based on the probability distribution of the topics in the collection.
2. Select <randomly> a word based on distribution probabilities of words in topic z.

Properties:

- one word may belong to several topics (we have distribution $p(w|t)$ $t \in T$)
- there can be several topics in one document;
- the collection has one common topic probability distribution. (!!!)

Graphical model pLSA



pLSA: how it is trained

we observe $p(w|d)$ (as we see documents that consists of words) the model can be trained to maximize the likelihood (maximum likelihood estimation) of:

$$p(w|d) = \Sigma_{t \in T} p(w|d,t)p(t|d) = \Sigma_{t \in T} p(w|t)p(t|d)$$

(our probabilistic model)

$$p(w|d,t) = p(w|t)$$

(conditional independence assumption)

pLSA: disadvantages

- Slow convergency on the big collections of documents
- The PLSA algorithm is characterized by overfitting, as well as non-uniqueness and instability of solutions.
- The algorithm does not highlight non-topic words. In real text, there are terms that do not explicitly refer to any of the topics. Accounting for such terms is possible with the help of robust thematic models, in which noise and background components are added.
- It cannot assign probabilities to new documents.

Latent Dirichlet Allocation (LDA)

Assumptions:

- There are K topics in the collection.
- Each document is represented as a mixture of topics.
- «Topic» — multinomial distribution over words. Each word in the vocabulary has some weight (probability) in each topic.)

Properties:

- one word may belong to several topics (we have distribution $p(w|t)$ $t \in T$)
- there can be several topics in one document;
- Each document has its own topic distribution (!!!), in one document only some of the topics are represented

How LDA differ from pLSA in simple words

- pLSA - one topic distribution for the whole collection, LDA - each document has its own topic distribution
- other differences more related to training (regularization, optimization method (Gibbs sampling), etc.)

Generative process of LDA

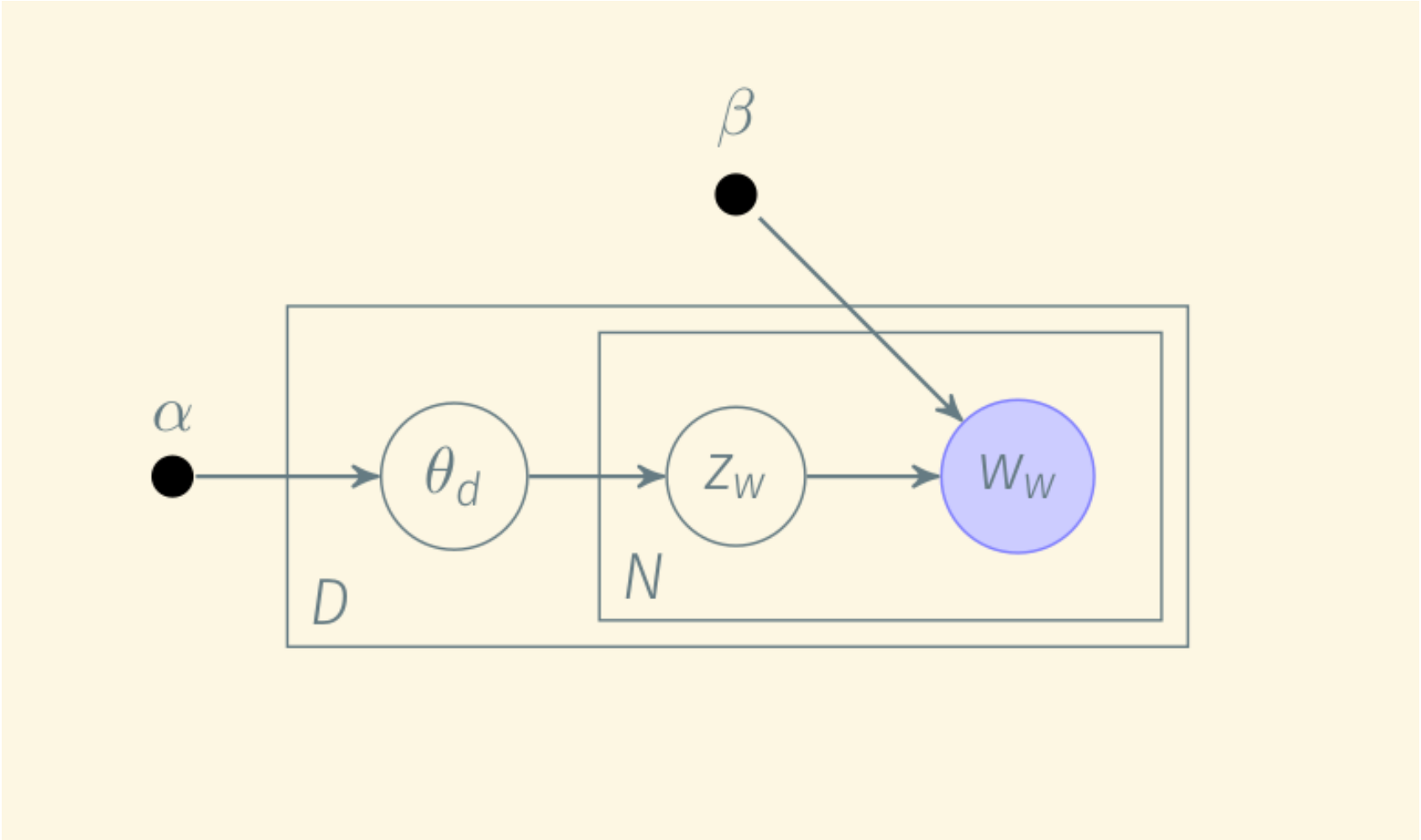
For each **topic** $1...k$:

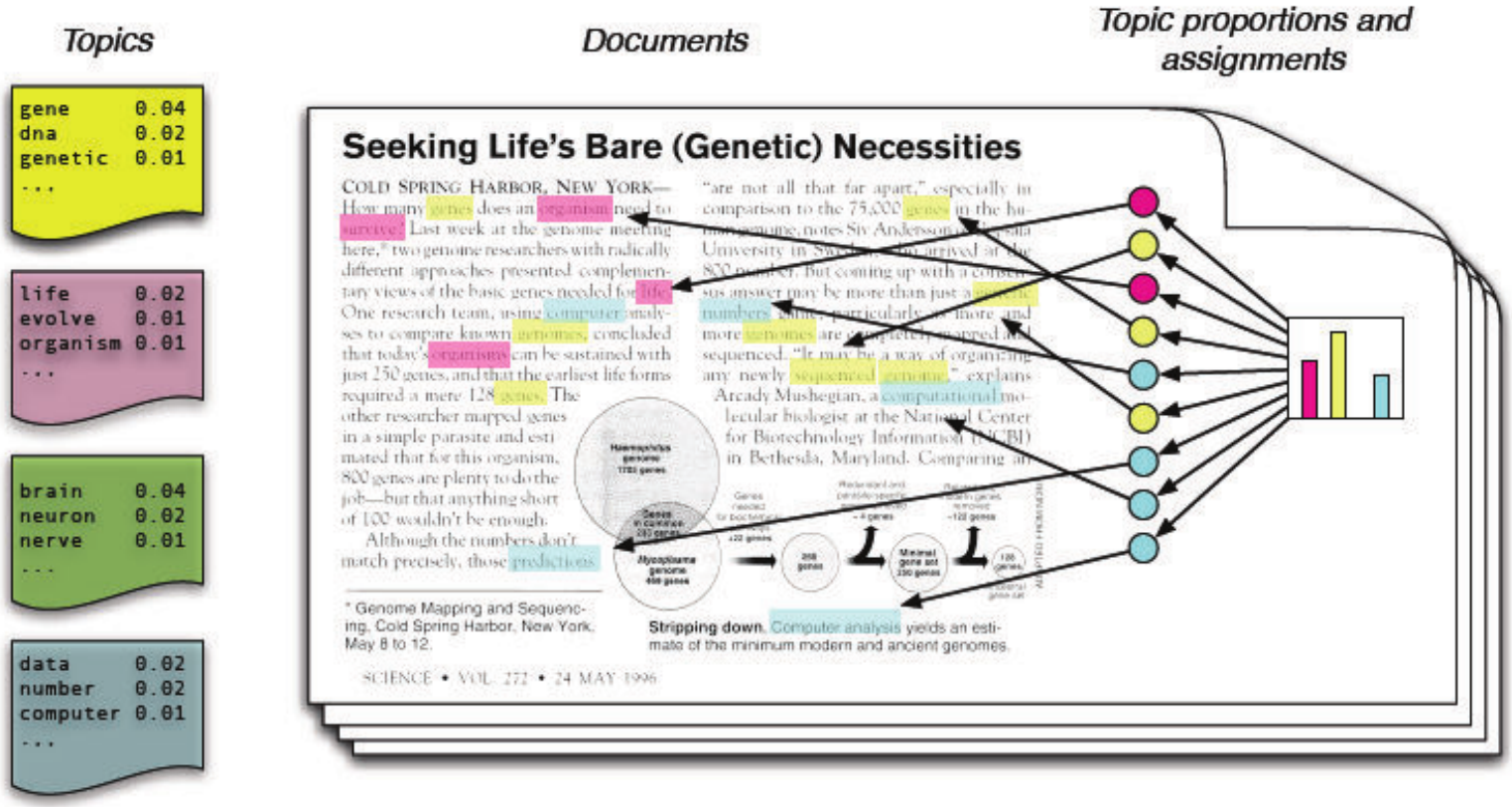
- Draw a **multinomial over words** $\beta_k \sim Dir(\eta)$

For each **document** $1...d$:

- Draw a **multinomial over topics** $\theta_d \sim Dir(\boldsymbol{\alpha})$
- For each word $w_{d,n}$:
    - Draw a topic $Z_{d,n} \sim Mult(\theta_d)$ with $Z_{d,n} \in [1..K]$
    - Draw a word $w_{d,n} \sim Mult(\beta_{Z_{d,n}})$

Graphical model of LDA





Slide by David Blei

Pros and cons of LDA

Pros:

- It provides better performances than LSA and pLSA.
- Unlike pLSA, LDA can assign a probability to a new document thanks to the document-topic Dirichlet distribution.
- As a probabilistic module, LDA can be embedded in more complex models or extended. There are extensions that address limitations from previous works.

Cons:

- The number of topics must be known beforehand.
- The bag-of-words approach disregards the semantic representation of words in a corpus, similarly to LSA and pLSA.
- It requires an extensive pre-processing phase to obtain a significant representation from the textual input data.
- Studies report LDA may yield too general or irrelevant topics.
- Results may also be inconsistent across different executions.

LDA: hyperparameters

- K - the number of topics
- *alpha* and *beta* - parameters of Dirichlet distribution

LDA: hyperparameters

- https://datascienceplus.com/topic-modeling-and-latent-dirichlet-allocation-lda/

Topic modeling: texts preparation

1. Preprocessing
2. Segmentation

Preprocessing: stop-words

- Static list:
- Dynamic list:
  - Too frequent words (N most frequent)
  - Too rare words (threshold: appear less than in N docs)
  - Too short (less then N letters)

More preprocessing

In TM, it can be a good idea to delete even more words. For example:

- Not nouns (ot not nouns and adjectives)
- Proper nouns (names)

Segmentation

Size of the document is important.

- Divide long texts (for ex, novels)
- Merge short texts (for ex, text messages)
- Optimal(?) text length — 100—1000 words (from abstract to article)

The main principle is the unity of context.

Output of LDA

You will see some lists like this:
(most probable words for each topic) So, you will probably try to interpret all this topics, drop the noisy ones and then try to do smth else.
You also can obtain: the probability of each topic in the document.
So, you can:

- calculate the topics proportions in the documents/in some groups/in different time periods
- look at the co-occurence of the topics in the documents
- etc.

Interpretation

eye face lip hand glance eyebrow hair voice smile nose forehead
взглядывать щека подымать темный плечо строгий широкий
рот повертываться черный ухо палец открытый словно
выражение высокий бледный густой весить прямой подбородок
звать угол чувствовать круглый вспыхивать похожий
покраснеть сводить слегка несколько спокойно дело eyelash
левый живой поглядеть успевать

Interpretation

Interpretation:

eye face lip hand glance eyebrow hair voice smile nose forehead взглядывать щека подымать темный плечо строгий широкий рот повертываться черный ухо палец открытый словно выражение высокий бледный густой весить прямой подбородок звать угол чувствовать круглый вспыхивать похожий покраснеть сводить слегка несколько спокойно дело eyelash левый живой поглядеть успевать

Interpretation

Interpretation: portrait descriptions

eye face lip hand glance eyebrow hair voice smile nose forehead
взглядывать щека подымать темный плечо строгий широкий
рот повертываться черный ухо палец открытый словно
выражение высокий бледный густой весить прямой подбородок
звать угол чувствовать круглый вспыхивать похожий
покраснеть сводить слегка несколько спокойно дело eyelash
левый живой поглядеть успевать

Model evaluation

- Human judgment of
  - Topics interpretability
  - sufficient generality / specificity of topics (based on the task)
- Measures of topics coherence (perplexity/coherence)
- * Assessment using metrics of the final task, if such a task exists

Model evaluation: Varieties of bad topics

Chained every word is connected to every other word through some pairwise word chain, but not all word pairs make sense. fatty ← acids → nucleic

Intruded either two or more unrelated sets of related words, joined arbitrarily, or an otherwise good topic with a few "intruder" words.

Random no clear, sensical connections between more than a few pairs of words

Unbalanced the top words are all logically connected to each other, but the topic combines very general and specific terms

[0] mimno2011optimizing

Model evaluation: Topic coherence

Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference.

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(w_m^{(t)}, w_l^{(t)}) + 1}{D(w_l^{(t)})}$$

, где

- $D(w)$ — document frequency of the word w
- $D(w, w')$ — joint document frequency of the words w и w'
- $V^{(t)} = (v_1^{(t)}, \ldots, v_M^{(t)})$ — list of top-M most probable words in the topic t

What is K (the number of topics)?

David Mimno
@dmimno

"Topic" models are just machines for finding groups of words that occur together. Themes are one of many ways to produce those groups, but they are not defined by them. To say that there is one "optimal" "topic" model is insulting to the complexity of human communication.

3:59 AM · Oct 28, 2021 · Twitter for iPhone

LDA: summary

- Distributional Hypothesis
- Corpus
- Defining the documents
- Selection and normalization of words
- training LDA model, tuning hyperparameters)
  - Document-term matrix
  - Topic as a multinomial distribution over words
  - Distribution of topics in the documents with Dirichlet
- Interpretation

Some links with simple (or not) explanations:

- about lda and other algorithms: https://towardsdatascience.com/topic-modeling-with-lsa-plsa-lda-nmf-bertopic-top2vec-a
- nice explanation of α and β parameters in LDA model in the beginning of the algorithm: https://datascienceplus.com/topic-modeling-and-latent-dirichlet-allocation-lda/
- topic coherence with examples of tuning hyperparameters (though in Python) : https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocat but there are explanations of topic coherence. BUT such tuning hyperparameters with topic coherence measures may be not as good idea - it does not guarantee that interpretability of topics will increase.
- about TM in simple words in russian: https://sysblok.ru/knowhow/kak-ponjat-o-chem-tekst-ne-chitaja-ego/