Measuring Topic Quality in Latent Dirichlet Allocation

Sergey Nikolenko Sergei Koltsov Olessia Koltsova

Steklov Institute of Mathematics at St. Petersburg

Laboratory for Internet Studies,
National Research University Higher School of Economics, St. Petersburg

Philosophy, Mathematics, Linguistics: Aspects of Interaction 2014
April 25, 2014

Outline

Probabilistic modeling

- Our work lies in the field of probabilistic modeling and Bayesian inference.
- Probabilistic modeling: given a dataset and some probabilistic assumptions, learn model parameters (and do some other exciting stuff).
- Bayes theorem:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}.$$

- General problems in machine learning / probabilistic modeling:
  - find $p(\theta|D) \propto p(\theta)p(D|\theta)$;
  - maximize it w.r.t. θ (maximal a posteriori hypothesis);
  - find predictive distribution $p(x \mid D) = \int p(x \mid \theta)p(\theta \mid D)\mathrm{d}\theta$.

Probabilistic modeling

- Two main kinds of machine learning problems:
  - *supervised*: we have "correct answers" in the dataset and want to extrapolate them (regression, classification);
  - *unsupervised*: we just have the data and want to find some structure in there (example: clustering).
- Natural language processing models with an eye to topical content:
  - usually the text is treated as a bag of words;
  - usually there is no semantics, words are treated as tokens;
  - the emphasis is on statistical properties of how words cooccur in documents;
  - sample supervised problem: text categorization (e.g., naive Bayes);
  - still, there are some very impressive results.

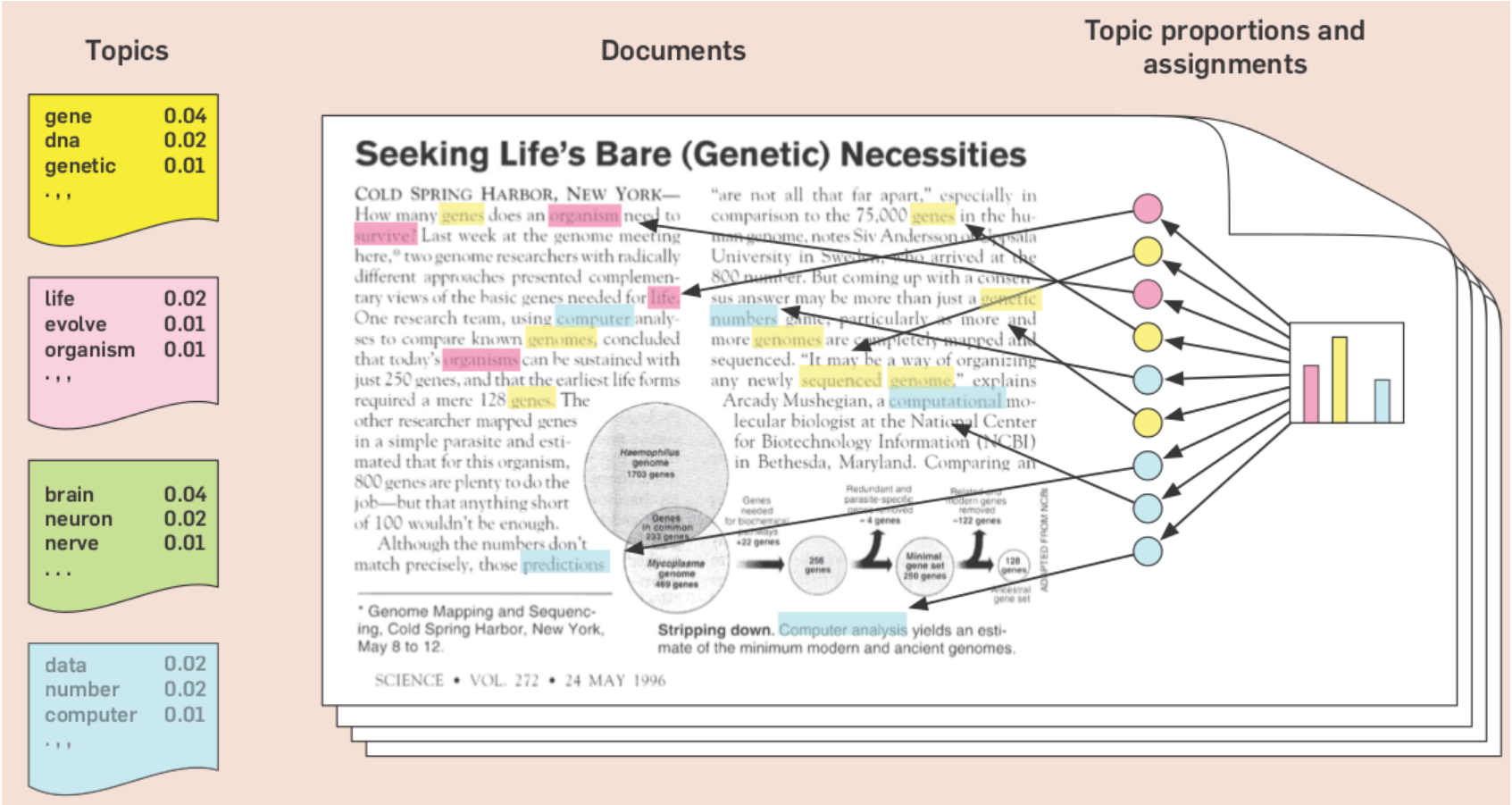
Topic modeling

- Suppose that you want to study a large text corpus.
- You want to identify specific topics that are discussed in this dataset and then either study the topics that are interesting for you or just look at their general distribution, do topical information retrieval etc.
- However, you do not know the topics in advance.
- Thus, you need to somehow extract what topics are discussed and find which topics are relevant for a specific document, in a completely unsupervised way because you do not know anything except the text corpus itself.
- This is precisely the problem that *topic modeling* solves.

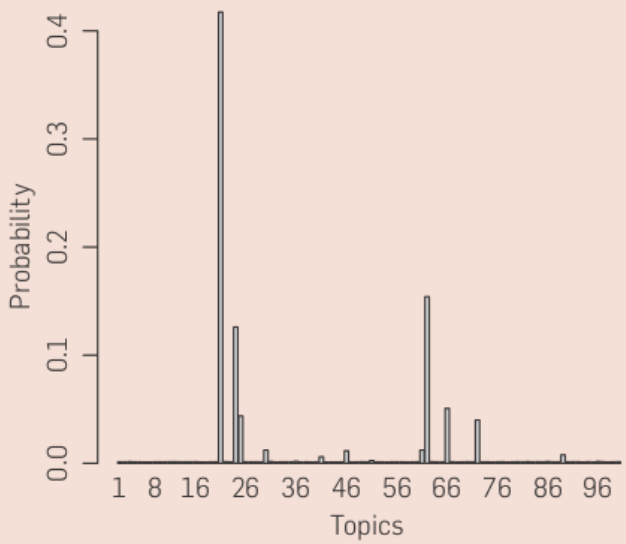
LDA

- Latent Dirichlet Allocation, LDA: the modern model of choice for topic modeling.
- In naive approaches to text categorization, one document belongs to one topic (category).
- In LDA, we (quite reasonably) assume that a document contains several topics:
  - a topic is a (multinomial) distribution on words (in the bag-of-words model);
  - a document is a (multinomial) distribution on topics.

Pictures from [Blei, 2012]



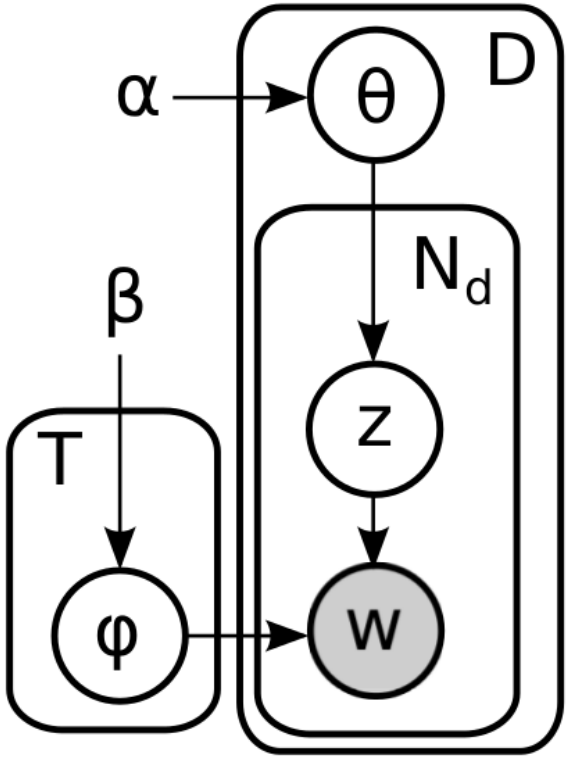
Pictures from [Blei, 2012]



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

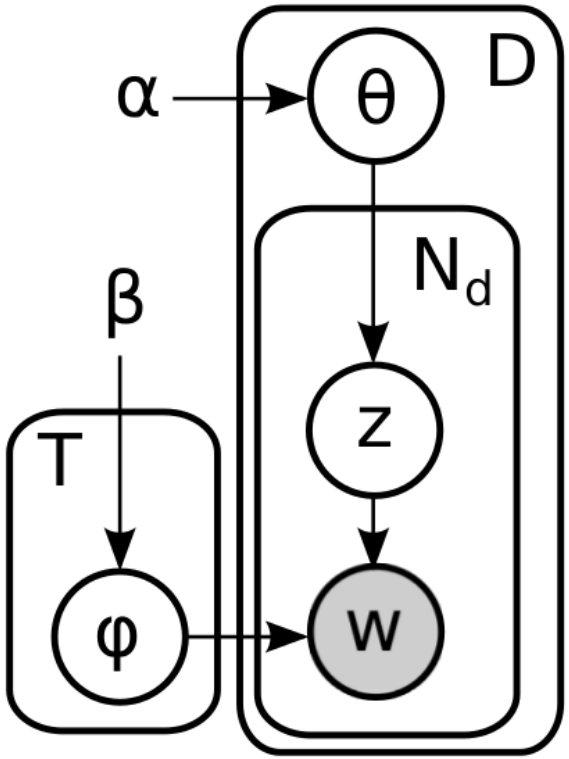
LDA

- LDA is a hierarchical probabilistic model:
  - on the first level, a mixture of topics φ with weights z;
  - on the second level, a multinomial variable θ whose realization z shows the distribution of topics in a document.
- It's called Dirichlet allocation because we assign Dirichlet priors α and β to model parameters θ and φ (conjugate priors to multinomial distributions).



LDA

- Generative model for the LDA:
  - choose document size $N \sim p(N \mid \xi)$;
  - choose distribution of topics $\theta \sim \mathrm{Dir}(\alpha)$;
  - for each of N words w_n:
    - choose topic for this word $z_n \sim \mathrm{Mult}(\theta)$;
    - choose word $w_n \sim p(w_n \mid \varphi_{z_n})$ by the corresponding multinomial distribution.



- So the underlying joint distribution of the model is

$$p(\theta, \varphi, \boldsymbol{z}, \boldsymbol{w}, N \mid \alpha, \beta) = p(N \mid \xi) p(\theta \mid \alpha) p(\varphi \mid \beta) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid \varphi, z_n).$$

LDA: inference

- The inference problem: given $\{\boldsymbol{w}\}_{\boldsymbol{w}\in D}$, find

$$p(\theta, \varphi \mid \boldsymbol{w}, \alpha, \beta) \propto \int p(\boldsymbol{w} \mid \theta, \varphi, \boldsymbol{z}, \alpha, \beta) p(\theta, \varphi, \boldsymbol{z} \mid \alpha, \beta) \mathrm{d}\boldsymbol{z}.$$

- There are two major approaches to inference in complex probabilistic models like LDA:
  - *variational approximations* simplify the graph by approximating the underlying distribution with a simpler one, but with new parameters that are subject to optimization;
  - *Gibbs sampling* approaches the underlying distribution by sampling a subset of variables conditional on fixed values of all other variables.

LDA: inference

- Both variational approximations and Gibbs sampling are known for the LDA; we will need the collapsed Gibbs sampling:

$$p(z_w = t \mid \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) \propto q(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) =$$
$$= \frac{n^{(d)}_{-w,t} + \alpha}{\sum_{t' \in T} \left(n^{(d)}_{-w,t'} + \alpha \right)} \frac{n^{(w)}_{-w,t} + \beta}{\sum_{w' \in W} \left(n^{(w')}_{-w,t} + \beta \right)},$$

  where $n^{(d)}_{-w,t}$ is the number of times topic t occurs in document d and $n^{(w)}_{-w,t}$ is the number of times word w is generated by topic t, not counting the current value z_w.
- Gibbs sampling is usually easier to extend to new modifications, and this is what we will be doing.

LDA extensions

- Numerous extensions for the LDA model have been introduced:
  - *correlated topic models* (CTM): topics are codependent;
  - *Markov topic models*: MRFs model interactions between topics in different parts of the dataset (multiple corpora);
  - *relational topic models*: a hierarchical model of a document network structure as a graph;
  - *Topics over Time*, *dynamic topic models*: documents have timestamps (news, blog posts), and we model how topics develop in time (e.g., by evolving hyperparameters α and β);
  - *DiscLDA*: each document has a categorical label, and we utilize LDA to mine topic classes related to the classification problem;
  - *Author-Topic model*: information about the author; texts from the same author will share common words;
  - a lot of work on *nonparametric* LDA variants based on Dirichlet processes (no predefined number of topics).

Outline

Quality of the topic model

- We want to know how well we did in this modeling.
- Problem: there is no ground truth, the model runs unsupervised, so no cross-validation.
- Solution: hold out a subset of documents, then check their likelihood in the resulting model.
- Alternative: in the test subset, hold out half the words and try to predict them given the other half.

Quality of the topic model

- Formally speaking, for a set of held-out documents D_{test}, compute the likelihood

$$p(\mathbf{w} \mid D) = \int p(\mathbf{w} \mid \Phi, \alpha\mathbf{m}) p(\Phi, \alpha\mathbf{m} \mid D) \mathrm{d}\alpha \mathrm{d}\Phi$$

  for each held-out document $\mathbf{w}$ and then maximize the normalized result

$$\text{perplexity}(D_{\text{test}}) = \exp\left(-\frac{\sum_{\mathbf{w} \in D_{\text{test}}} \log p(\mathbf{w})}{\sum_{\mathbf{w} \in D_{\text{test}}} N_d}\right).$$

- It is a nontrivial problem computationally, but efficient algorithms have already been devised.

Quality of individual topics

- However, this is only a general quality measure for the entire model.
- Another important problem: quality of individual topics.
- Qualitative studies: is a topic interesting? We want to help researchers (social studies, media studies) identify "good" topics suitable for human interpretation.

Quality of individual topics

- Recent studies agree that topic *coherence* is a good candidate.
- For a topic t characterized by its set of top words W_t, coherence is defined as

$$c(t, W_t) = \sum_{w_1, w_2 \in W_t} \log \frac{d(w_1, w_2) + \epsilon}{d(w_1)},$$

  where $d(w_i)$ is the number of documents that contain w_i, $d(w_i, w_j)$ is the number of documents where w_i and w_j cooccur, and ϵ is a smoothing count usually set to either 1 or 0.01.
- I.e., a topic is good if its words cooccur together often.

Quality of individual topics

- However, in our studies we have found that coherence does not work so well.
- We see two reasons:
  1. many topics that have good coherence are composed of common words that do not represent any topic of discourse per se; these common words do indeed cooccur often but coherence does not distinguish between high frequency words and informative words;
  2. in modern user-generated datasets (e.g., in the blogosphere), many topics stem from copies, reposts, and discussions of a single text that either directly copy or extensively cite the original text; thus, words that appear in this text have very good cooccurrence statistics even if they are rather meaningless words.

Quality of individual topics

- To alleviate these drawbacks, we propose a modification of the basic coherence metric that takes into account informative content by substituting tf-idf scores instead of the number of [co]occurrences.
- Namely, we define *tf-idf coherence* as

$$c_{\text{tf-idf}}(t, W_t) = \sum_{w_1, w_2 \in W_t} \log \frac{\sum_{d: w_1, w_2 \in d} \text{tf-idf}(w_1, d)\text{tf-idf}(w_2, d) + \epsilon}{\sum_{d: w_1 \in d} \text{tf-idf}(w_1, d)},$$

where tf-idf is computed with augmented frequency,

$$\text{tf-idf}(w, d) = \text{tf}(w, d) \times \text{idf}(w) = \left(\frac{1}{2} + \frac{f(w, d)}{\max_{w' \in d} f(w', d)}\right) \log \frac{|D|}{|\{d \in D : w \in d\}|},$$

and $f(w, d)$ is how many times term w occurs in document d.

- Intuitively, we skew the metric towards topics with high tf-idf scores in top words.

Topics with top coherence scores

- Topics with common words have excellent coherence:

(1)	(2)	(3)	(4)	(5)
say	just	issue	just	author
tell	stay	solution	instance	fact
know	solve	problem	example	article
just	problem	situation	often	say
need	moment	side	say	issue
nothing	know	relation	have	write

(6)	(7)	(8)	(9)	(10)
life	have	century	just	right
know	instance	appear	know	law
just	image	beginning	say	Russian
live	example	history	nothing	state
see	system	well-known	general	citizen
say	follow	end	see	federation

Topics with top tf-idf coherence scores

- W.r.t. tf-idf coherence we see much more interesting topics:

(1)	(2)	(3)	(4)	(5)
(64)	(58)	(42)	(63)	(75)
terr. attack	pope	church	butter	Korea
explosion	Vatican	orthodox	sugar	North
Boston	Roman	temple	add	DPRK
terrorist	church	cleric	flour	South
brother	cardinal	faith	dough	Kim
police	catolic	church^{adj}	recipe	Korean
Boston^{adj}	Benedict	patriarch	egg	nuclear

(6)	(7)	(8)	(9)	(10)
(48)	(34)	(61)	(28)	(25)
add	Cyprus	Syria	military	war
butter	bank	Syrian	army	German^{adj}
onion	Russian	Al	service	Germany
meat	Cypriot	country	general	German
pepper	Euro	Muslim	officer	Hitler
minute	financial	fighter	mil. force	Soviet
dish	money	Arab	defense	world

Experimental evaluation

- We also conducted an experimental evaluation of topic quality based on lists of top words.
- For each topic, we asked the subjects (among them media studies experts) two binary questions:
  (1) Do you understand why the words in this topic have been united together, do you see obvious semantic criteria that unite the words in this topic?
  (2) If you have answered "yes" to the first question: can you identify specific issues/events that documents in this topic might address?
- (plus an open question: please sum up a topic in a few words)

Experimental evaluation

- We compared quality metrics with the *area-under-curve* (AUC) measure: the share of pairs consisting of a positive and a negative example that the classifier ranks correctly (the positive one is higher).
- Tf-idf coherence is significantly better.

Dataset	# of topics	Question 1			Question 2		
		AUC		Ham.	AUC		Ham.
		coh.	tf-idf		coh.	tf-idf	
March 2012	100	0.66	**0.74**	0.15	0.59	**0.65**	0.24
March 2012	200	0.72	**0.76**	0.19	0.67	**0.73**	0.24
April 2012	100	0.66	**0.74**	0.10	0.59	**0.65**	0.22
September 2012	200	0.67	**0.73**	0.14	0.65	**0.70**	0.25

Summary

- Latent Dirichlet allocation is a probabilistic model that extracts topics from a corpus of documents.
- By training the model, we decompose the word-document matrix into word-topic and topic-document matrices.
- There are many topics, and it is desirable to distinguish interesting ones.
- For this purpose, we propose a new metric (tf-idf coherence) and show that it is better.

Thank you!

Thank you for your attention!