Rethinking Probabilistic Topic Modeling from the Point of View of Classical Non-Bayesian Regularization

Konstantin Vorontsov
prof., head of Machine Learning & Semantic Analysis Laboratory, IAI MSU

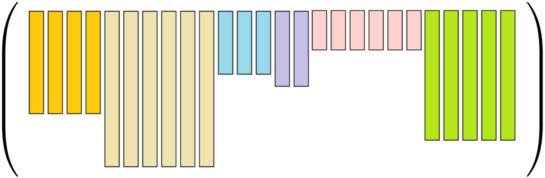
Data Analysis, Optimization and their Applications
MIPT • January 30, 2023

Contents

Maximization of a function with unit simplices constraints

Let $\Omega = (\omega_j)_{j \in J}$ be a set of non-negative normalized vectors ω_j having dimensions $|I_j|$ respectively, $\omega_j = (\omega_{ij})_{i \in I_j}$:

$\Omega =$



Problem: maximize the function $f(\Omega)$ on unit simplices:

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad j \in J; \\ \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J. \end{cases}$$

Necessary extremum conditions and the simple-iteration method

Define normalization operator: $p_i = \underset{i \in I}{\mathrm{norm}}(x_i) = \dfrac{\max(x_i, 0)}{\sum_k \max(x_k, 0)}$

Lemma. Let $f(\Omega)$ be continuously differentiable function on Ω. If ω_j is the local maximum of $f(\Omega)$ and $\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} > 0$ for some i, then ω_j satisfies the system of equations

$$\omega_{ij} = \underset{i \in I_j}{\mathrm{norm}}\left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right).$$

- For numerical solution, the simple-iteration method can be used
- Vectors $\omega_j = 0$ must be discarded as degenerate solutions
- Iterations are similar to gradient maximization of $f(\Omega)$:

$$\omega_{ij} := \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}},$$

differing in "norm" projection and absence of η parameter

Proof of the Lemma on Maximization on unit simplices

Problem: $f(\Omega) \to \max\limits_{\Omega}; \quad \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J.$

The Lagrangian of the optimization problem:

$$\mathscr{L}(\Omega; \mu, \lambda) = f(\Omega) + \sum_{j \in J} \lambda_j \Bigl(\sum_{i \in I_j} \omega_{ij} - 1 \Bigr) - \sum_{j \in J} \sum_{i \in I_j} \mu_{ij} \omega_{ij}.$$

The Karush–Kuhn–Tucker conditions for the vector ω_j:

$$\frac{\partial f(\Omega)}{\partial \omega_{ij}} = \lambda_j - \mu_{ij}, \quad \mu_{ij}\omega_{ij} = 0, \quad \mu_{ij} \geqslant 0.$$

Multiply both sides of the equation by ω_{ij}:

$$A_{ij} \equiv \omega_{ij} \frac{\partial f(\Omega)}{\partial \omega_{ij}} = \omega_{ij} \lambda_j.$$

By the condition of the Lemma, $\exists i$: $A_{ij} > 0$. Then $\lambda_j > 0$.

If $\frac{\partial f(\Omega)}{\partial \omega_{ij}} < 0$ for some i, then $\mu_{ij} > 0 \;\Rightarrow\; \omega_{ij} = 0$.

Thus, $\omega_{ij}\lambda_j = (A_{ij})_+; \; \lambda_j = \sum\limits_i (A_{ij})_+ \;\Rightarrow\; \omega_{ij} = \operatorname*{norm}\limits_i (A_{ij}).$ ∎

Theorem on the simple-iteration method convergence

$$\omega_{ij}^{t+1} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij}^{t} \frac{\partial f(\Omega^t)}{\partial \omega_{ij}^{t}} \right)$$

Theorem. Let $f(\Omega)$ be a continuously differentiable upper bounded function, and all Ω^t satisfy the following conditions starting from some iteration t^0:

- $\forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t = 0 \to \omega_{ij}^{t+1} = 0$ (keeping zeros)
- $\exists \varepsilon > 0 \;\; \forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t \notin (0, \varepsilon)$ (separation from zero)
- $\exists \delta > 0 \;\; \forall j \in J \;\; \exists i \in I_j \;\; \omega_{ij}^t \frac{\partial f(\Omega^t)}{\partial \omega_{ij}} \geqslant \delta$ (nondegeneracy)

Then $f(\Omega^{t+1}) > f(\Omega^t)$ and $\left|\omega_{ij}^{t+1} - \omega_{ij}^t\right| \to 0$ under $t \to \infty$.

Irkhin I. A., Vorontsov K. V. Convergence of the algorithm of additive regularization of topic models // Trudy Instituta Matematiki i Mekhaniki UrO RAN, 2020

Probabilistic Topic Modeling (PTM): the problem setting

Given:

- W, a finite set (vocabulary) of terms (words, tokens)
- D, a finite set (collection) of documents
- n_{dw} = how many times term w appears in document d

Find: Probabilistic Topic Model (PTM)

$$p(w|d) = \sum_{t\in T} p(w\,|\,\cancel{d},t)\,p(t|d) = \sum_{t\in T} \varphi_{wt}\theta_{td}$$

where $\varphi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$ are model parameters

Log-likelihood maximization:

$$L(\Phi,\Theta) = \ln \prod_{d,w} p(w|d)^{n_{dw}} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

with constraints $\varphi_{wt} \geqslant 0$, $\sum_w \varphi_{wt} = 1$, $\theta_{td} \geqslant 0$, $\sum_t \theta_{td} = 1$

Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.

Some interpretations of the PTM problem setting

1. **Soft bi-clustering** by topical clusters $t \in T$
of both documents: $p(t|d)$, and terms: $p(t|w) = p(w|t)\frac{p(t)}{p(w)}$

2. **Vector representations** (topical embeddings) which are probabilistic, interpretable, sparse: $p(t|d)$, $p(t|w)$, $p(t|d,w)$, etc.

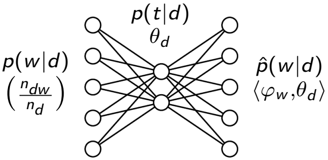
3. **Matrix factorization** $\left(\frac{n_{dw}}{n_d}\right) \approx \Phi\Theta$ which is low-rank, non-negative (stochastic), approximate

4. **Auto-encoder** of documents $p(w|d)$ into embeddings $p(t|d)$:

encoder $f_\Phi \colon \frac{n_{dw}}{n_d} \to \theta_d$
decoder $g_\Phi \colon \theta_d \to \Phi\theta_d$

the reconstruction problem:

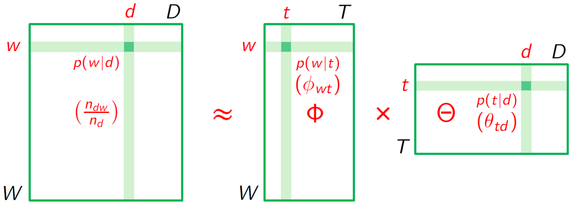
$$\sum_d \mathrm{KL}\left(\frac{n_{dw}}{n_d} \,\Big\|\, \langle \varphi_w, \theta_d \rangle\right) \to \min_{\Phi,\Theta}$$



5. **Probabilistic language model** $p(w|d)$

The ill-posed problem of matrix factorization

The problem of nonnegative (stochastic) matrix factorization:



If (Φ, Θ) is a solution, then (Φ', Θ') is also the solution:

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta)$, where $\operatorname{rank} S = |T|$
- $\mathscr{L}(\Phi', \Theta') = \mathscr{L}(\Phi, \Theta)$ for other linearly independent solutions
- $\mathscr{L}(\Phi', \Theta') \geqslant \mathscr{L}(\Phi, \Theta) - \varepsilon$ for approximate solutions

Adding *regularizing criteria* should constrict the set of solutions.

ARTM — Additive Regularization for Topic Modeling

Maximize log-likelihood with regularization criteria $R_i(\Phi, \Theta)$:

$$\sum_{d,w} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \qquad R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta)$$

EM-algorithm is a simple-iteration method for the system of equations with auxiliary variables $p_{tdw} = p(t|d,w)$:

$$\begin{array}{rl} \text{E-step:} & \begin{cases} p_{tdw} = \operatorname*{norm}_{t\in T}\big(\varphi_{wt}\theta_{td}\big) \\ \varphi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum_{d\in D} n_{dw} p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\Big) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{cases} \\ \text{M-step:} & \end{array}$$

K.Vorontsov. Additive regularization for topic models of text collections. 2014.

Proof (by Lemma on Maximization on unit simplices)

Apply the Lemma to the regularized log-likelihood:

$$f(\Phi,\Theta) = \sum_{d,w} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

$$\varphi_{wt} = \operatorname*{norm}_{w\in W}\left(\varphi_{wt}\frac{\partial f}{\partial \varphi_{wt}}\right) = \operatorname*{norm}_{w\in W}\left(\varphi_{wt}\sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right) =$$

$$= \operatorname*{norm}_{w\in W}\left(\sum_{d\in D} n_{dw} p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right);$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\left(\theta_{td}\frac{\partial f}{\partial \theta_{td}}\right) = \operatorname*{norm}_{t\in T}\left(\theta_{td}\sum_{w\in W} n_{dw}\frac{\varphi_{wt}}{p(w|d)} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) =$$

$$= \operatorname*{norm}_{t\in T}\left(\sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right).$$

■

Two most cited topic models are special cases of ARTM

PLSA, Probabilistic Latent Semantic Analysis [Hofmann, 1999]:

$$R(\Phi, \Theta) = 0.$$

M-step gives frequency estimates of conditional probabilities:

$$\varphi_{wt} = \operatorname*{norm}_{w}\bigl(n_{wt}\bigr), \qquad \theta_{td} = \operatorname*{norm}_{t}\bigl(n_{td}\bigr).$$

LDA, Latent Dirichlet Allocation [Blei, Ng, Jordan, 2001]:

$$R(\Phi, \Theta) = \sum_{t,w} \beta_w \ln \varphi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}.$$

M-step gives shifted frequency estimates, $\beta_w > -1$, $\alpha_t > -1$:

$$\varphi_{wt} = \operatorname*{norm}_{w}\bigl(n_{wt} + \beta_w\bigr), \qquad \theta_{td} = \operatorname*{norm}_{t}\bigl(n_{td} + \alpha_t\bigr).$$

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet allocation. NIPS 2001.

Bayesian vs classical (non-Bayesian) regularization

Bayesian inference of posterior distribution $p(\Omega|X)$ being usually cumbersome and approximate is used only for Ω point estimate:

$$\text{Posterior}(\Omega|X,\gamma) \;\propto\; p(X|\Omega)\,\text{Prior}(\Omega|\gamma)$$

$$\Omega := \arg\max_{\Omega} \text{Posterior}(\Omega|X,\gamma)$$

Maximum a posteriori estimation (MAP) gives a point estimate Ω directly without posterior inference:

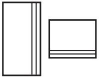
$$\Omega := \arg\max_{\Omega}\bigl(\ln p(X|\Omega) + \ln \text{Prior}(\Omega|\gamma)\bigr)$$

Multicriteria additive regularization (ARTM) generalizes MAP to non-probabilistic regularizers as well as the weighted sum of regularizers, without violating the convergence properties:

$$\Omega := \arg\max_{\Omega}\Bigl(\ln p(X|\Omega) + \sum_{i=1} \tau_i R_i(\Omega)\Bigr)$$

Regularizers for the interpretability of topics

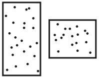
background



LDA: Smoothing background topics $B \subset T$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_{w} \beta_w \ln \varphi_{wt} + \alpha_0 \sum_{d} \sum_{t \in B} \alpha_t \ln \theta_{td}$$

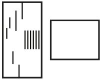
sparse



"Anti-LDA": Sparsing subject domain topics $S = T \backslash B$:

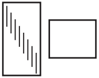
$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w} \beta_w \ln \varphi_{wt} - \alpha_0 \sum_{d} \sum_{t \in S} \alpha_t \ln \theta_{td}$$

seed words



Smoothing relevant topics with seed words vocabulary or query documents

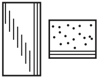
decorrelated



Making topics as different as possible:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_{w} \varphi_{wt} \varphi_{ws}$$

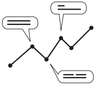
interpretable



Making topics more interpretable by combining regularizers: Decorrelation + Smoothing + Sparsing

Many Bayesian PTMs can be restated as ARTM regularizers

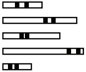
regression



Linear predictive model $\hat{y}_d = \langle v, \theta_d \rangle$ for documents:

$$R(\Theta, v) = -\tau \sum_{d \in D} \Bigl(y_d - \sum_{t \in T} v_t \theta_{td} \Bigr)^2$$

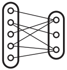
biterm



Using word co-occurrence data n_{uv}:

$$R(\Phi) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} n_t \varphi_{ut} \varphi_{vt}$$

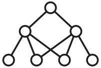
relational



Using document links or citations data n_{dc}:

$$R(\Theta) = \tau \sum_{d,c \in D} n_{dc} \sum_{t \in T} \theta_{td} \theta_{tc}$$

hierarchy

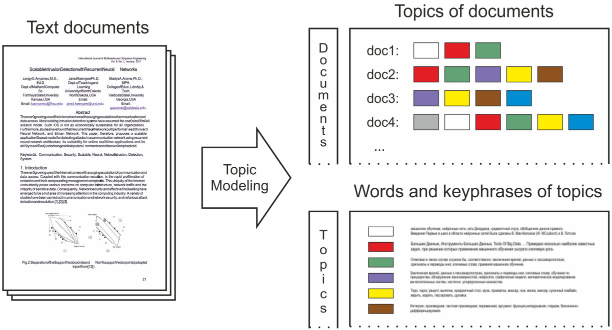


Hierarchical links between topics t and subtopics s:

$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \varphi_{ws} \psi_{st}$$

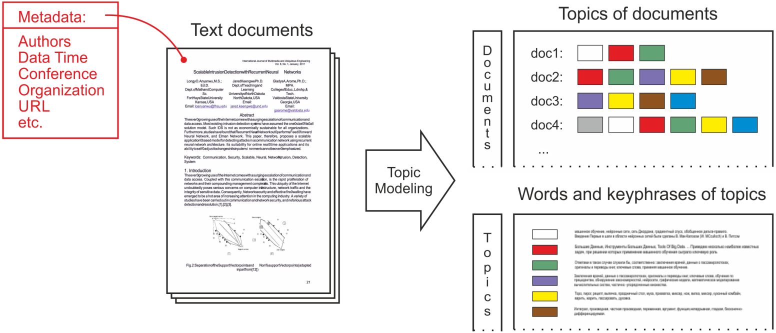
Multimodal Probabilistic Topic Modeling

Topic may generate terms of multiple *modalities*: $p(\text{word}|t)$, $p(\text{n-gram}|t)$, $p(\text{entity}|t)$, $p(\text{tag}|t)$,



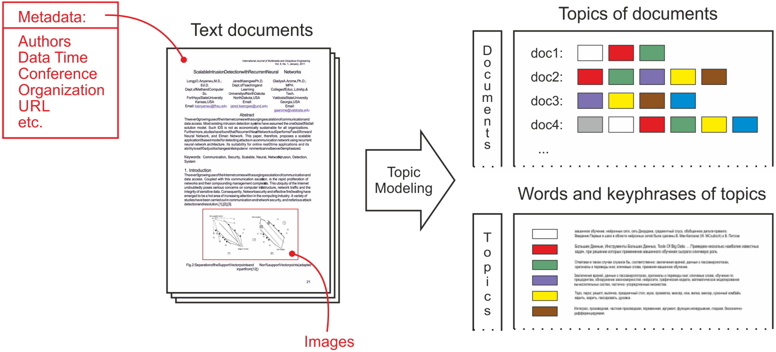
Multimodal Probabilistic Topic Modeling

Topic may generate terms of multiple *modalities*: $p(\text{word}|t)$, $p(\text{n-gram}|t)$, $p(\text{entity}|t)$, $p(\text{tag}|t)$, $p(\text{author}|t)$, $p(\text{time}|t)$, $p(\text{source}|t)$,



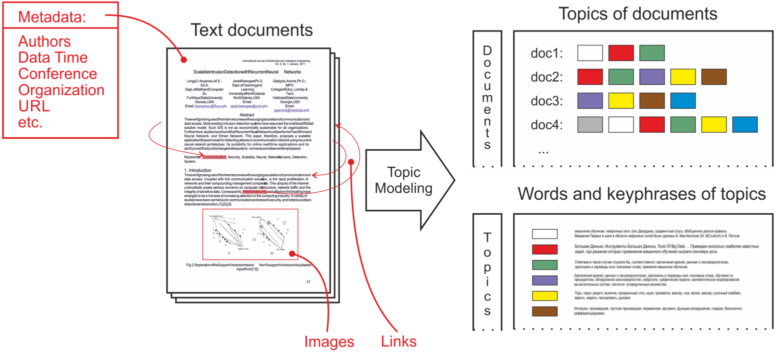
Multimodal Probabilistic Topic Modeling

Topic may generate terms of multiple *modalities*: $p(\text{word}|t)$, $p(\text{n-gram}|t)$, $p(\text{entity}|t)$, $p(\text{tag}|t)$, $p(\text{author}|t)$, $p(\text{time}|t)$, $p(\text{source}|t)$, $p(\text{object}|t)$,



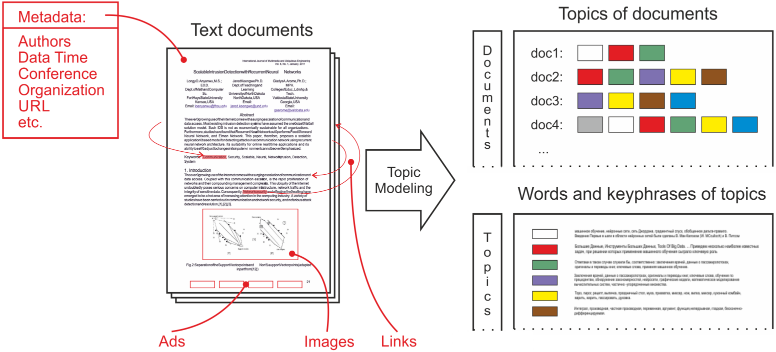
Multimodal Probabilistic Topic Modeling

Topic may generate terms of multiple *modalities*: $p(\text{word}|t)$, $p(\text{n-gram}|t)$, $p(\text{entity}|t)$, $p(\text{tag}|t)$, $p(\text{author}|t)$, $p(\text{time}|t)$, $p(\text{source}|t)$, $p(\text{object}|t)$, $p(\text{link}|t)$,



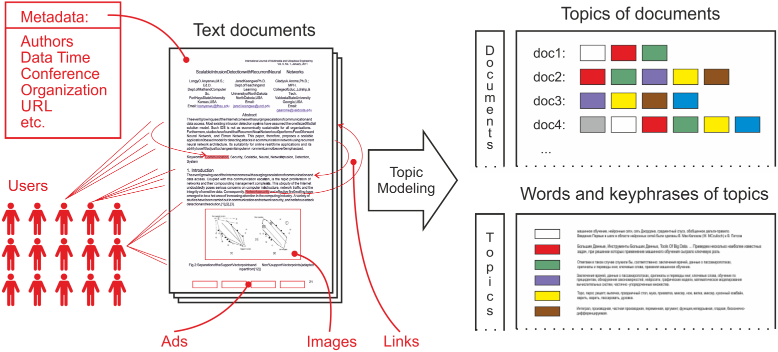
Multimodal Probabilistic Topic Modeling

Topic may generate terms of multiple *modalities*: $p(\text{word}|t)$, $p(\text{n-gram}|t)$, $p(\text{entity}|t)$, $p(\text{tag}|t)$, $p(\text{author}|t)$, $p(\text{time}|t)$, $p(\text{source}|t)$, $p(\text{object}|t)$, $p(\text{link}|t)$, $p(\text{banner}|t)$,



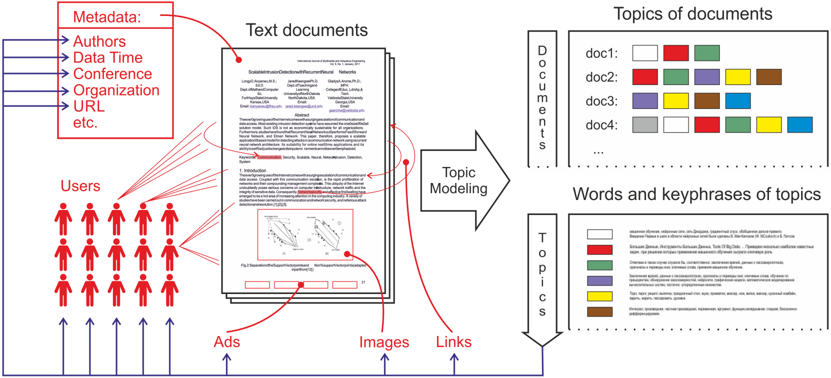
Multimodal Probabilistic Topic Modeling

Topic may generate terms of multiple *modalities*: $p(\text{word}|t)$, $p(\text{n-gram}|t)$, $p(\text{entity}|t)$, $p(\text{tag}|t)$, $p(\text{author}|t)$, $p(\text{time}|t)$, $p(\text{source}|t)$, $p(\text{object}|t)$, $p(\text{link}|t)$, $p(\text{banner}|t)$, $p(\text{user}|t)$



Multimodal Probabilistic Topic Modeling

Topic may generate terms of multiple *modalities*: $p(\text{word}|t)$, $p(\text{n-gram}|t)$, $p(\text{entity}|t)$, $p(\text{tag}|t)$, $p(\text{author}|t)$, $p(\text{time}|t)$, $p(\text{source}|t)$, $p(\text{object}|t)$, $p(\text{link}|t)$, $p(\text{banner}|t)$, $p(\text{user}|t)$



Multimodal extension of ARTM

W^m is a vocabulary of *terms* of m-th *modality*, $m \in M$.

Maximize the sum of modality log-likelihoods with regularization:

$$\sum_{m\in M} \lambda_m \sum_{d\in D} \sum_{w\in W^m} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi, \Theta) \;\to\; \max_{\Phi,\Theta}$$

EM-algorithm is a simple-iteration method for the system

$$\begin{cases} \text{E-step:} & p_{tdw} = \underset{t\in T}{\operatorname{norm}}\bigl(\varphi_{wt}\theta_{td}\bigr) \\ \text{M-step:} & \varphi_{wt} = \underset{w\in W^m}{\operatorname{norm}}\Bigl(\sum_{d\in D} \lambda_{m(w)} n_{dw} p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\Bigr) \\ & \theta_{td} = \underset{t\in T}{\operatorname{norm}}\Bigl(\sum_{w\in d} \lambda_{m(w)} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Bigr) \end{cases}$$

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova, A.Ianina. Non-Bayesian additive regularization for multimodal topic modeling of large collections. 2015.

Example. Multilingual topic model of Wikipedia

Dataset: 216 175 pairs of parallel Russian–English articles.
Top 10 words and their probabilities $p(w|t)$ in %:

topic #68				topic #79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Assessors evaluated 396 topics from 400 as paired and interpretable

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova. BigARTM: open source library for regularized multimodal topic modeling of large collections. 2015.

Example. Multilingual topic model of Wikipedia

Dataset: 216 175 pairs of parallel Russian–English articles.
Top 10 words and their probabilities $p(w|t)$ in %:

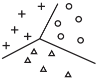
topic #88				topic #251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Assessors evaluated 396 topics from 400 as paired and interpretable

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova. BigARTM: open source library for regularized multimodal topic modeling of large collections. 2015.

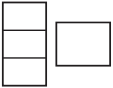
Special cases of the multimodal topic modeling

supervised



The modalities of classes or categories for text classification or categorization

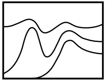
multilanguage



The modalities of languages with translation dictionary $\pi_{uwt} = p(u|w,t)$ for the $k \to \ell$ language pair:

$$R(\Phi, \Pi) = \tau \sum_{u \in W^k} \sum_{t \in T} n_{ut} \ln \sum_{w \in W^\ell} \pi_{uwt} \varphi_{wt}$$

temporal



Topics dynamics over the modality of time intervals i:

$$R(\Phi) = -\tau \sum_{i \in I} \sum_{t \in T} \left| \varphi_{it} - \varphi_{i-1,t} \right|$$

geospatial

The modality of geolocations g with proximity $S_{gg'}$:

$$R(\Phi) = -\frac{\tau}{2} \sum_{g,g' \in G} S_{gg'} \sum_{t \in T} n_t^2 \left(\frac{\varphi_{gt}}{n_g} - \frac{\varphi_{g't}}{n_{g'}} \right)^2$$

Transaction data

Data may contain not only pairs (d, w) but also *transactions* represented by triples, . . . , n-tuples of terms of multiple modalities

- **Social network data:**
  (d, u, w) — the user u wrote the word w in the blog d
- **Advertising network data:**
  (u, d, b) — the user u clicked on the banner b on the page d
- **Recommender system data:**
  (u, m, s) — the user u rated the movie m in the situation s
- **Banking and retail data:**
  (b, s, g) — the buyer u bought the goods g from the seller s
- **Passenger flight data:**
  (u, a, b, c) — customer u flies from a to b by airline c

The problem is *giving* an observable set of transactions
find the latent distribution $p(t|v)$ of topics t for each term v

Hypergraph ARTM of transaction data: problem statement

V^m is the term vocabulary of modality $m \in M$
$V = V^1 \sqcup \dots \sqcup V^M$ is joint vocabulary of all modalities
Hypergraph $\Gamma = \langle V, E \rangle$ is a set E of subsets of V
$(d, x) \in E$ is an edge with $x \subset V$ and *container* vertex $d \in V$

Given:

E_k, an observable set of edges (transactions) of the type k;
n_{kdx}, the number of transactions (d, x) of the type k

Find: a generative topic model of edges of all types:

$$p(x|d) = \sum_{t \in T} \underbrace{p(t|d)}_{\theta_{td}} \prod_{v \in x} \underbrace{p(v|t)}_{\varphi_{vt}}$$

Log-likelihood maximization:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} n_{kdx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vt} + R(\Phi, \Theta) \to \max_{\Phi,\Theta}$$

Hypergraph ARTM of transaction data: EM-algorithm

Log-likelihood maximization:

$$\sum_{k\in K} \tau_k \sum_{(d,x)\in E_k} n_{kdx} \ln \sum_{t\in T} \theta_{td} \prod_{v\in x} \varphi_{vt} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

EM-algorithm is a simple-iteration method for the system of equations with auxiliary variables $p_{tdx} = p(t|d,x)$:

$$\begin{cases} \text{E-step:} & p_{tdx} = \underset{t\in T}{\operatorname{norm}}\Bigl(\theta_{td} \prod_{v\in x} \varphi_{vt}\Bigr) \\ \text{M-step:} & \varphi_{vt} = \underset{v\in V^m}{\operatorname{norm}}\Bigl(\sum_{k\in K} \tau_k \sum_{(d,x)\in E_k} [v\in x]\, n_{kdx} p_{tdx} + \varphi_{vt} \frac{\partial R}{\partial \varphi_{vt}}\Bigr) \\ & \theta_{td} = \underset{t\in T}{\operatorname{norm}}\Bigl(\sum_{k\in K} \tau_k \sum_{(d,x)\in E_k} n_{kdx} p_{tdx} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\Bigr) \end{cases}$$

Proof (by Lemma on Maximization on unit simplices)

Let's apply the Lemma to log-likelihood with a regularizer R:

$$\varphi_{vt} = \operatorname*{norm}_{v\in V_m}\left(\varphi_{vt}\sum_{k\in K}\tau_k\sum_{dx\in E_k} n_{kdx}\frac{\theta_{td}}{p(x|d)}\frac{\partial}{\partial\varphi_{vt}}\prod_{u\in x}\varphi_{ut} + \varphi_{vt}\frac{\partial R}{\partial\varphi_{vt}}\right) =$$

$$= \operatorname*{norm}_{v\in V_m}\left(\sum_{k\in K}\sum_{dx\in E_k}\tau_k n_{kdx}[v\in x]p_{tdx} + \varphi_{vt}\frac{\partial R}{\partial\varphi_{vt}}\right)$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\left(\theta_{td}\sum_{k\in K}\tau_k\sum_{x\in d} n_{kdx}\frac{1}{p(x|d)}\prod_{v\in x}\varphi_{vt} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\right) =$$

$$= \operatorname*{norm}_{t\in T}\left(\sum_{k\in K}\sum_{x\in d}\tau_k n_{kdx}p_{tdx} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\right)$$

■

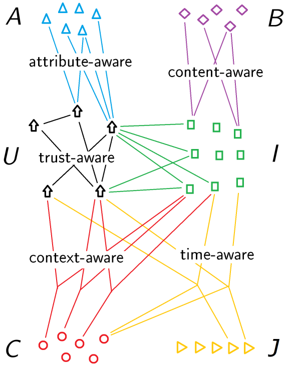
K. Vorontsov. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization. Optimization and Its Applications. 2023 (to appear)

Transaction data in Recommender Systems

U is a finite set (vocabulary) of users
I is a finite set (vocabulary) of items
A is a finite set of user attributes (social, region, tags, etc)
B is a finite set of item properties or content elements
C is a finite set of situation context
J is a finite set of time intervals

Transaction types in RecSys:

n_{ui} — user u chose item i
n_{ua} — user u has attribute a
n_{ib} — item i has property b
n_{uv} — user u trusts the user v
n_{uib} — user u tagged item i by b
n_{uic} — user u chose i in context c
n_{uicj} — u chose i в c at time j



Hypergraph topic models of natural language

The edge of a hypergraph can be a subset of terms that are semantically related and generated by a common topic:

- sentence / syntax tree branch / noun phrase / syntagma
- fact as a Subject-Predicate-Object (SPO) triple
- pair of synonyms, hyponym–hypernym, meronym–holonym
- lexical chain
- comment text and its author

The model gives interpretable topical embeddings:

- $p(t|d)$ for a document of container d
- $p(t|w) = \varphi_{wt}\frac{p(t)}{p(w)}$ for a term w
- $p(t|d, x)$ for a transaction, a phrase, a fact, etc.

Topic models of sentences and short texts: TwitterLDA, senLDA

S_d is a set of sentences in the document d
n_{sw} = how many times term w appears in the sentence s

Topic model of a sentence s:

$$p(s|d) = \sum_{t\in T} p(t|d) \prod_{w\in s} p(w|t)^{n_{sw}} = \sum_{t\in T} \theta_{td} \prod_{w\in s} \varphi_{wt}^{n_{sw}}$$

Maximization of the regularized log-likelihood

$$\sum_{d\in D} \sum_{s\in S_d} \ln \sum_{t\in T} \theta_{td} \prod_{w\in s} \varphi_{wt}^{n_{sw}} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

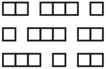
is a special case of hypergraph topic model with sentences considered as edges (transactions).

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee Peng Lim et al. Comparing Twitter and traditional media using topic models. ECIR 2011.

G.Balikas, M.-R.Amini, M.Clausel. On a topic model for sentences. SIGIR 2016.

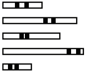
Beyond the “bag-of-words” restrictive assumption

n-gram



Modalities of *n*-grams, named entities, collocations extracted by external text preprocessors

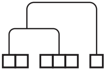
biterm



Modeling co-occurrence data n_{uv} of word pairs (u, v):

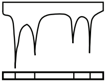
$$R(\Phi) = \tau \sum_{u,v} n_{uv} \ln \sum_t n_t \varphi_{ut} \varphi_{vt}$$

syntax



Phrases extracted by a syntax parser for hypergraph TM
Modalities of part of speech, part of a sentence

segmentation



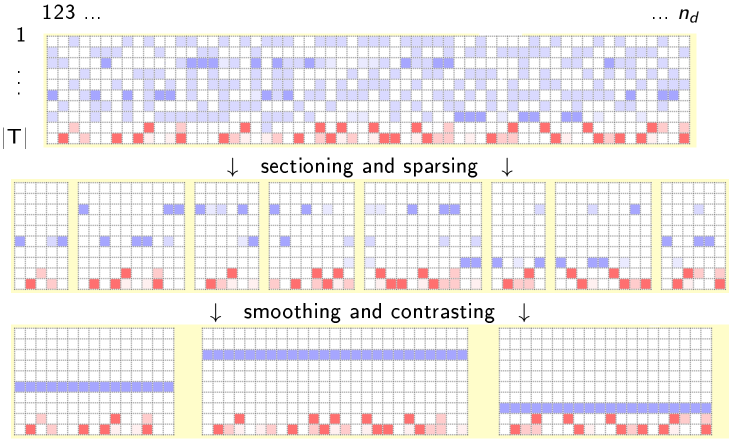
Detecting thematically homogeneous segments in sequential text

D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov. Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.

The segment topical structure of text and intratext processing

Consider a document $d = \{w_1, \dots, w_{n_d}\}$ of a size n_d

Matrix of *word-in-document topics* $p(t|d, w_i)$ of a size $T \times n_d$:



E-step regularization as $p(t|d,w)$ post-processing

Log-likelihood maximization with $\Pi = \big(p_{tdw} = p(t|d,w)\big)_{T\times D\times W}$:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Pi(\Phi,\Theta),\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}.$$

EM-algorithm is a simple-iteration method for the system

$$\begin{array}{ll}
\text{E-step:} & \left\{\begin{array}{l}
p_{tdw} = \operatorname*{norm}_{t\in T}\big(\varphi_{wt}\theta_{td}\big) \\
\tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R}{\partial p_{tdw}} - \sum\limits_{z\in T} p_{zdw}\frac{\partial R}{\partial p_{zdw}}\right)\right) \\
\varphi_{wt} = \operatorname*{norm}_{w\in W}\left(\sum\limits_{d\in D} n_{dw}\tilde{p}_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right) \\
\theta_{td} = \operatorname*{norm}_{t\in T}\left(\sum\limits_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)
\end{array}\right. \\
\text{M-step:} &
\end{array}$$

Proof sketch, in three steps

1. For the function $p_{tdw}(\Phi,\Theta) = \frac{\varphi_{wt}\theta_{td}}{\sum_z \varphi_{wz}\theta_{zd}}$ and any $z \in T$

$$\varphi_{wt}\frac{\partial p_{zdw}}{\partial \varphi_{wt}} = \theta_{td}\frac{\partial p_{zdw}}{\partial \theta_{td}} = p_{tdw}\bigl([z=t] - p_{zdw}\bigr).$$

2. Introduce an auxiliary function of variables Π, Φ, Θ:

$$Q_{tdw}(\Pi,\Phi,\Theta) = \frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{zdw}}.$$

$\tilde{R}(\Phi,\Theta) = R(\Pi(\Phi,\Theta),\Phi,\Theta)$ does not depend on p_{tdw} for $w \notin d$, thus

$$\varphi_{wt}\frac{\partial \tilde{R}}{\partial \varphi_{wt}} = \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} + \sum_{d\in D} p_{tdw} Q_{tdw}; \quad \theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}} = \theta_{td}\frac{\partial R}{\partial \theta_{td}} + \sum_{w\in d} p_{tdw} Q_{tdw}.$$

3. Substitute this equations into the M-step formulas:

$$\varphi_{wt} = \operatorname*{norm}_{w\in W}\Bigl(\sum_{d\in D} n_{dw} p_{tdw} + \sum_{d\in D} Q_{tdw} p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\Bigr);$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(\sum_{w\in d} n_{dw} p_{tdw} + \sum_{w\in d} Q_{tdw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Bigr).$$

■

Any E-step post-processing is equivalent to a regularizer $R(\Pi)$

So, any differentiable regularizer $R(\Pi, \Phi, \Theta)$ induces a unique transformation $p_{tdw} \to \tilde{p}_{tdw}$ that is performed before the M-step.

The converse is also true:

Theorem. Let the vector $(\tilde{p}^k_{tdw})_{t \in T}$ satisfying the normalization condition $\sum_t \tilde{p}^k_{tdw} = 1$ be substituted in the M-step formulas instead of the vector $(p^k_{tdw})_{t \in T}$ for each (d, w): $n_{dw} > 0$ at the k-th iteration of the EM-algorithm.
That is equivalent to adding a smoothing–sparsing regularizer:

$$R(\Pi) = \sum_{d \in D} \sum_{w \in d} n_{dw} \sum_{t \in T} (\tilde{p}^k_{tdw} - p^k_{tdw}) \ln p_{tdw}.$$

Conclusion: E-step post-processing takes into account the order of terms in the document bypassing the “bag of words” hypothesis.

Fast EM-algorithm with single-pass through the document

Log-likelihood maximization under $\Theta = \Theta(\Phi)$ constraint:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td}(\Phi) + R\big(\Phi, \Theta(\Phi)\big) \to \max_{\Phi}$$

EM-algorithm is a simple-iteration method for the system

$$p_{tdw} = \operatorname*{norm}_{t \in T} \big(\varphi_{wt} \theta_{td}(\Phi)\big); \qquad n_{td} = \sum_{w \in d} n_{dw} p_{tdw} + \theta_{td}(\Phi) \frac{\partial R}{\partial \theta_{td}}$$

$$\tilde{p}_{tdw} = p_{tdw} + \frac{\varphi_{wt}}{n_{dw}} \sum_{z \in T} \frac{n_{zd}}{\theta_{zd}(\Phi)} \frac{\partial \theta_{zd}}{\partial \varphi_{wt}}$$

$$\varphi_{wt} = \operatorname*{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} \tilde{p}_{tdw} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right)$$

I.Irkhin, V.Bulatov, K.Vorontsov. Additive regularization of topic models with fast text vectorization. Computer Research and Modeling, 2020.

Proof (by Lemma on Maximization on unit simplices)

The M-step maximization problem:

$$Q(\Phi) = \sum_{d\in D} \sum_{u\in W} \sum_{z\in T} n_{du} p_{zdu} \big(\ln \varphi_{uz} + \ln \theta_{zd}(\Phi)\big) + R\big(\Phi, \Theta(\Phi)\big) \to \max_{\Phi}$$

Apply the Lemma for the regularized log-likelihood Q:

$$\varphi_{wt} \frac{\partial Q}{\partial \varphi_{wt}} = \sum_{d\in D} n_{dw} p_{tdw} + \sum_{d,z,u} n_{du} p_{zdu} \frac{\varphi_{wt}}{\theta_{zd}} \frac{\partial \theta_{zd}}{\partial \varphi_{wt}} + \varphi_{wt} \sum_{d,z} \frac{\partial R}{\partial \theta_{zd}} \frac{\partial \theta_{zd}}{\partial \varphi_{wt}} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} =$$

$$= \sum_{d\in D} n_{dw} \left(p_{tdw} + \frac{\varphi_{wt}}{n_{dw}} \sum_{z\in T} \frac{1}{\theta_{zd}} \underbrace{\left(\sum_{u\in d} n_{du} p_{zdu} + \theta_{zd} \frac{\partial R}{\partial \theta_{zd}} \right)}_{n_{zd}} \frac{\partial \theta_{zd}}{\partial \varphi_{wt}} \right) + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} =$$

$$= \sum_{d\in D} n_{dw} \underbrace{\left(p_{tdw} + \frac{\varphi_{wt}}{n_{dw}} \sum_{z\in T} \frac{n_{zd}}{\theta_{zd}} \frac{\partial \theta_{zd}}{\partial \varphi_{wt}} \right)}_{\tilde{p}_{tdw}} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}$$

■

Averaging word embeddings: $\theta_{td}(\Phi) = \sum_w p_{wd}\, \mathrm{norm}_t(\varphi_{wt} p_t)$

Partial derivatives: $\varphi_{wt} \frac{\partial \theta_{zd}}{\partial \varphi_{wt}} = p_{wd} \tilde{\varphi}_{tw} \big(\delta_{zt} - \tilde{\varphi}_{zw}\big)$

EM-algorithm is a simple-iteration method for the system

$$\tilde{\varphi}_{tw} = \mathop{\mathrm{norm}}_{t\in T}\big(\varphi_{wt} p_t\big); \qquad \theta_{td} = \sum_{w\in d} p_{wd} \tilde{\varphi}_{tw}$$

$$p_{tdw} = \mathop{\mathrm{norm}}_{t\in T}\big(\varphi_{wt} \theta_{td}\big); \qquad n_{td} = \sum_{w\in d} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}$$

$$\tilde{p}_{tdw} = p_{tdw} + \frac{\tilde{\varphi}_{tw}}{n_d} \left(\frac{n_{td}}{\theta_{td}} - \sum_{z\in T} \tilde{\varphi}_{zw} \frac{n_{zd}}{\theta_{zd}} \right)$$

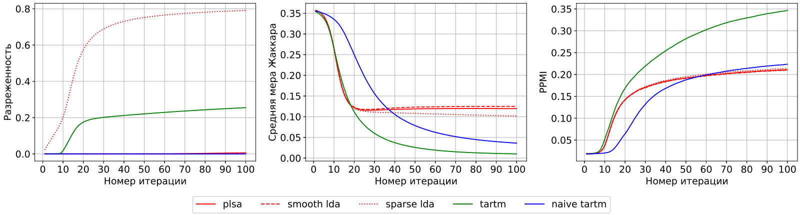
$$\varphi_{wt} = \mathop{\mathrm{norm}}_{w\in W} \left(\sum_{d\in D} n_{dw} \tilde{p}_{tdw} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right)$$

N.B. E-step still takes $O(n_d |T|)$ operations for each d.

Experiment. Verification of the modified EM algorithm

NIPS collection, $|T| = 50$. Models to compare:

- Baselines: PLSA, smooth LDA, sparse LDA
- TARTM (Θless ARTM) is our modified EM-algorithm
- naive TARTM is EM-algorithm with single document iteration

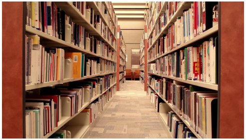


- TARTM clears topics from common words,
- improves sparsity, diversity and coherence of topics

I.Irkhin, V.Bulatov, K.Vorontsov. Additive regularization of topic models with fast text vectorization. Computer Research and Modeling, 2020.

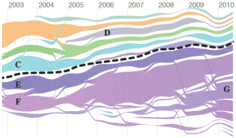
Some of the Topic Modeling applications

exploratory search
in digital libraries

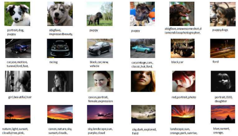


search and recommendation
in topical communities

topic detection and
tracking in news flows

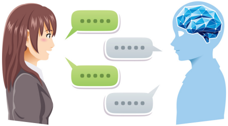


multimodal search
for texts and images



mining the banking
customer behavior

dialog management in
chatbot intelligence

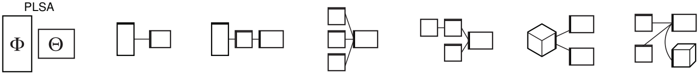


Topic Model for applications in Digital Humanities must be...

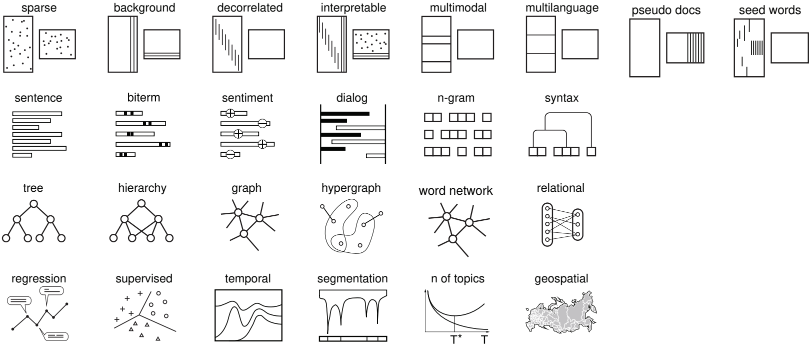
1. **Interpretable** so that each topic could tell about itself
2. **Hierarchical** to subdivide topics into subtopics recursively
3. **Temporal** for topic detection and tracking
4. **Multimodal** with authors, categories, tags, links, users, etc.
5. **Multigram** with n-grams being domain concepts
6. **Multilingual** for cross-lingual information retrieval
7. **Segmented** for thematically heterogeneous documents
8. **Supervised** for processing expert markups and user logs
9. **Determining number of topics** automatically
10. **Creating and labeling topics** automatically
11. **Online** for fast one-pass data processing
12. **Parallel, distributed** for big data processing

Palette of regularizers in ARTM (the list to be continued)

Matrix factorization structures:



Regularizers to constrain the model or use additional data:



ARTM unifies and simplifies topic modeling for applications

<table>
<tr><th>Stages</th><th>Bayesian Inference for PTMs</th><th colspan="2">ARTM</th></tr>
<tr><td>Requirements analysis:</td><td>Requirements analysis</td><td colspan="2">Requirements analysis</td></tr>
<tr><td>Model formalization:</td><td>Generative model design</td><td>predefined criteria</td><td>user-defined criteria</td></tr>
<tr><td>Model inference:</td><td>Bayesian inference for the generative model (VI, GS, EP)</td><td colspan="2">One regularized EM-algorithm for any combination of criteria</td></tr>
<tr><td>Model implementation:</td><td>Researchers coding (Matlab, Python, R)</td><td colspan="2">Production code (C++)</td></tr>
<tr><td>Model evaluation:</td><td>Researchers coding (Matlab, Python, R)</td><td>predefined measures</td><td>user-defined measures</td></tr>
<tr><td>Deployment:</td><td>Deployment</td><td colspan="2">Deployment</td></tr>
</table>

conventions: ::: not unified stages ::: ::: unified stages :::

Bayesian modeling forces new calculus and coding for each model

ARTM introduces the modular "LEGO-style" modeling technology, packing each requirement into a *regularization plug-in*

BigARTM: open source for fast and modular topic modeling

BigARTM features:

- Parallelism + modalities + regularizers + hypergraph
- Out-of-core one-pass processing of large text collections
- Built-in library of regularizers and quality measures

BigARTM community since 2014:

- Open-source https://github.com/bigartm (discussion group, issue tracker, pull requests)
- Documentation http://bigartm.org

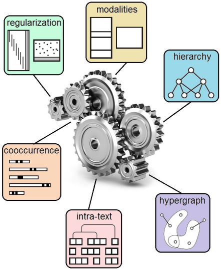
BigARTM license and programming environment:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

The cornerstone features of the BigARTM and TopicNet libraries

BigARTM:

- additive regularization
- multimodal data
- topical hierarchy
- intratext regularization
- hypergraph data



TopicNet:

- choosing regularization strategies for model selection
- automatic logging of all experiments
- collecting a “topic bank” from miltiple models
- visualization of topic modeling results

V. Bulatov, E. Egorov, E. Veselova, D. Polyudova, V. Alekseev, A. Goncharov, K. Vorontsov. TopicNet: making additive regularisation for topic modelling accessible. LREC-2020

Benchmarking BigARTM vs. Gensim and Vowpal Wabbit

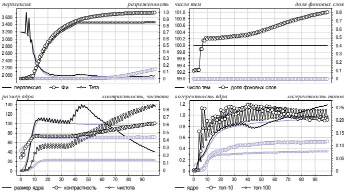
3.7M wiki articles, 100K unique words, time (perplexity)

proc.	$\|T\|$	Gensim	Vowpal Wabbit	BigARTM	BigARTM async
1	50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
1	100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
1	200	637m (3241)	154m (3960)	83m (3347)	53m (3362)
2	50	89m (5056)		22m (5092)	13m (5160)
2	100	143m (4012)		29m (4107)	19m (4144)
2	200	325m (3297)		47m (3347)	28m (3380)
4	50	88m (5311)		12m (5216)	7m (5353)
4	100	104m (4338)		16m (4233)	10m (4357)
4	200	315m (3583)		26m (3520)	16m (3634)
8	50	88m (6344)		8m (5648)	5m (6220)
8	100	107m (5380)		10m (4660)	6m (5119)
8	200	288m (4263)		15m (3929)	10m (4309)

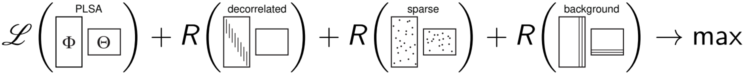
D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov.
Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.

Decorrelation, sparsing and smoothing of topics

Goal: to find a combination of regularizers that improves the interpretability of topics by a set of criteria.



The bag-of-regularizers:



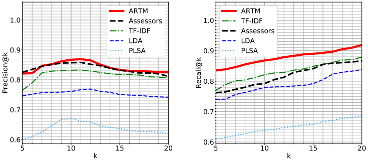
Results:

- sparsity 0 → 95%, coherence 0.25 → 0.96, purity 0.14 → 0.89, contrast 0.43 → 0.52,
- without noticeable damage to perplexity: 1920 → 2020
- successive regularization strategies have been developed

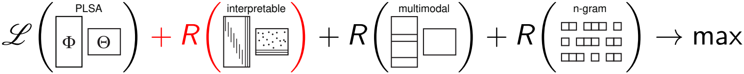
K.Vorontsov, A.Potapenko. Additive regularization of topic models. Mach. Learn., 2015.

Exploratory search in tech news #1

Goal: doc-by-doc exploratory search
— Habr.ru (175K docs)
— TechCrunch.com (760K docs)



The bag-of-regularizers:



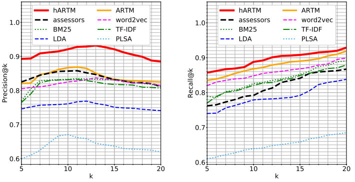
Results:

- Precision and Recall 88% bypass both assessors and baselines (tf-idf, word2vec, PLSA, LDA).
- The topic-based search engine instantly performs the work that people typically complete in about 5–65 minutes.

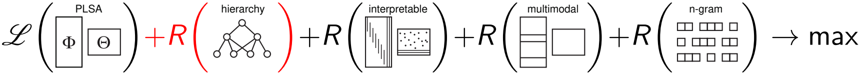
A.Ianina, L.Golitsyn, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Exploratory search in tech news #2

Goal: improving precision and recall of doc-by-doc exploratory search using hierarchical ARTM and cutting off irrelevant topics.



The bag-of-regularizers:



Results:

- Precision and Recall 93% bypass both assessors and baselines (tf-idf, BM25, word2vec, PLSA, LDA, ARTM).
- The optimal dimension of vectors has increased: 200 → 1400 (Habr.ru), 475 → 2800 (TechCrunch.com).

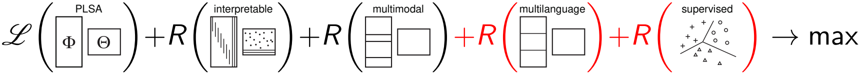
A.Ianina, K.Vorontsov. Regularized multimodal hierarchical topic model for document-by-document exploratory search. FRUCT–ISMW, 2019.

Multilingual search and categorization of scientific papers

Goal: multilingual ARTM for 100 languages using multiple library classification systems UDC (УДК), ГРНТИ, ОЭСР, ВАК

модель	ср.ч. УДК	ср.% УДК	ср.ч. ГРНТИ	ср.% ГРНТИ
Базовая ТМ	0.558	0.165	0.536	0.220
XLM-RoBERTa	0.835	0.179	0.832	0.288
ARTM	0.995	0.225	0.852	0.366

The bag-of-regularizers:



Results:

- the accuracy of multilingual search is 94%
- vocabulary reduction to 11K tokens per language (using BPE) results in the model reduction 128 GB → 4.8 GB.

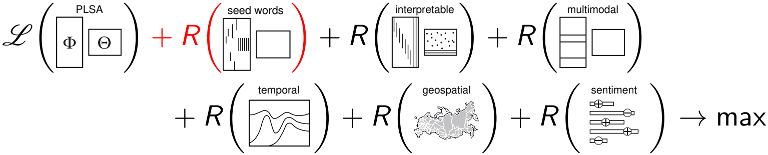
П.Потапова, А.Грабовой, О.Бахтеев, Е.Егоров, Ю.Чехович, К.Воронцов и др. Мультиязыковая автоматическая рубрикация научных документов. 2023 (to appear)

Mining ethnical discourse in social media

Goal: detecting as many topics as possible about nationalities and inter-ethnic relations (using 300 ethnonyms as seed words).

(японцы): японский, япония, корея, китайский, жилища, авария, фукусиму, цунами, сообщать, океан, станция, хатико, район, правительство, атомный,
(норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, воспитание, норвежский, родительский, родить, мальчик, взрослый, опека, сын,
(венесуэльцы): куба, кастро, венесуэла, чавес, президент, уго, мадуро, боливия, фидель, глава, латинский, венесуэльский, лидер, боливарианской, президентский, альенде, гевару,
(китайцы): китайский, россия, производство, китай, продукция, страна, предприятие, компания, технология, военный, регион, производить, производственный, промышленность, российский, экономический, кнр,
(азербайджанцы): русский, азербайджан, азербайджанец, россия, азербайджанский, таксист, диаспора, анапа, народ, москва, страна, армянин, слово, рынок,
(грузины): грузинский, спецназ, военный, август, баташева, российский, спецназовец, миротворец, операция, румын, бригада, миротворческий, абхазия, группа, войска, русский, цхинвале,
(осетины): конституция, осетия, аминат, русский, осетинский, южный, северный, россия, война, республика, вопрос, алахай, российский, население, конфликт,
(цыгане): наркотик, цыган, цыганка, хороший, место, страна, деньга, время, работать, жизнь, жить, рука, дом, цыганский, наркоманка,

The bag-of-regularizers:



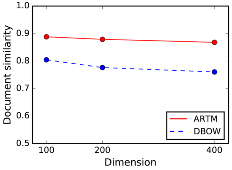
Results: the number of relevant topics 45 (LDA) → 83 (ARTM).

M.Apishev, S.Koltcov, O.Koltsova, S.Nikolenko, K.Vorontsov. Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI, 2016.

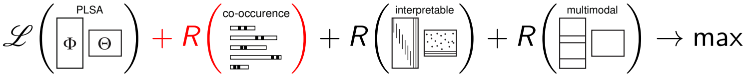
Mining ethnic content online with additively regularized topic models. 2016.

Topic modeling of short texts and probabilistic word embeddings

Goal: sparse interpretable embeddings $p(t|w)$ based on distributional semantics similar to word2vec and WNTM.



The bag-of-regularizers:



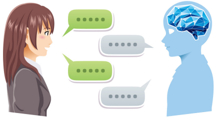
Results:

- Accuracy on document similarity tasks: 0.8 → 0.9
- Performance on word similarity tasks: 0.53→0.58, 0.38→0.61
- Coherence of topics: 0.08 → 0.33
- Modalities improve performance on word similarity tasks

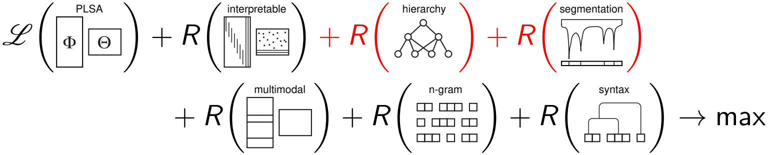
A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.

Intent detection and scenario analysis of call center records

Goal: determine typical topics and scenarios of dialogues between operators and customers;
then build the topical hierarchy of customer intents for further dialogs markup.



The bag-of-regularizers:

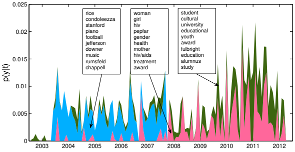


Results: the intent classification accuracy 60% → 66%.

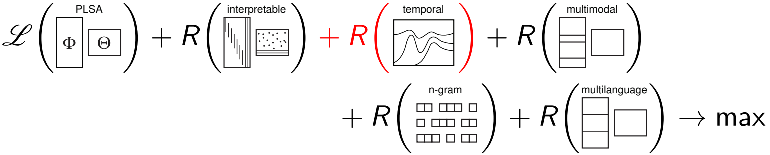
A.Popov, V.Bulatov, D.Polyudova, E.Veselova. Unsupervised dialogue intent detection via hierarchical topic model. RANLP, 2019.

Topic detection and tracking (TD&T) in news flows

Goal: TD&T in the collection of press releases of the Ministries of Foreign Affairs of 4 countries.



The bag-of-regularizers:



Results:

- classification of topics into permanent and events
- coherence of topics: $5.5 \to 6.5$

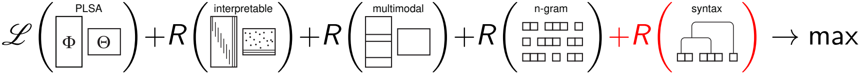
Н.Дойков. Адаптивная регуляризация вероятностных тематических моделей. ВКР бакалавра, ВМК МГУ, 2015.

Unsupervised detection of polarized opinions in political news

Goal: find linguistic-based cues for clustering event topics into polarized opinions

Modalities	*Pr*	*Rec*	*F1*
TF-IDF	0.51	0.95	0.67
SPO	0.59	0.7	0.64
FR	0.86	0.49	0.65
Sent	0.69	0.57	0.66
SPO+FR	0.86	0.68	0.76
SPO+Sent	0.83	0.78	0.81
FR+Sent	0.9	0.52	0.67
All	**0.77**	**0.97**	**0.86**

The bag-of-regularizers:



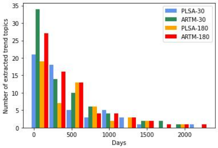
Results:

- detection of opinions within topics: F1-measure = 0.86%
- as a result of the joint use of three modalities: facts as SPO triplets (subject–predicate–object), semantic roles of words from Fillmore's theory, named entity sentiments.

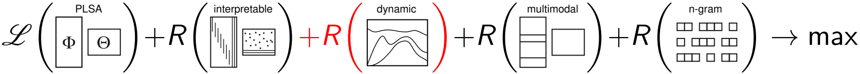
D.Feldman, T.Sadekova, K.Vorontsov. Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining. Dialogue 2020.

Scientific trend detection in big collection of scientific papers

Goal: early detection of trending topics with initial exponential growth in AI/ML research area, 2009–2021.



The bag-of-regularizers:



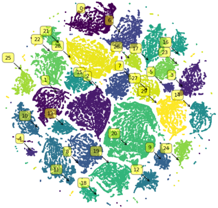
Results:

- automatic detection of 90 from 91 trends in AI/ML area
- 63% of topics are detected in a year, 79% in two years

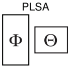
N.Gerasimenko, A.Chernyavskiy, M.Nikiforova, M.Nikitin, K.Vorontsov. Incremental topic modeling for scientific trend detection Doklady RAS, 2022.

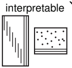
Topic modeling of bank transaction data

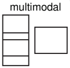
Goal: reveal patterns of consumer behavior
from purchase transaction data;
document = consumer,
word = MCC (Merchant Category Codes).

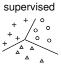


The bag-of-regularizers:

$\mathscr{L}\Big($ 

$\Big) + R\Big($ 

$\Big) + R\Big($ 

$\Big) + R\Big($ 

$\Big) \to \max$

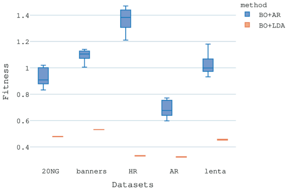
Results:

- topics are interpretable patterns of consumer behavior
- consumer topical behavior profile $p(t|d)$ can be used for predicting gender, age, wealth, interests, etc

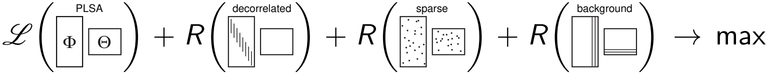
E.Egorov, F.Nikitin, A.Goncharov, V.Alekseev, K.Vorontsov. Topic modelling for extracting behavioral patterns from transactions data. 2019.

Automatic learning of regularization coefficients

Goal: AutoARTM is automatic optimization of hyperparameters such as regularization coefficients, number of iterations, number of topics according to the topic coherence criterion.



The bag-of-regularizers:



Results:

- Significant improvement in topic coherence across 5 datasets
- Genetic algorithm showed the best results

M.Khodorchenko, S.Teryoshkin, T.Sokhin, N.Butakov. Optimization of learning strategies for ARTM-based topic models. LNCS, 2020.

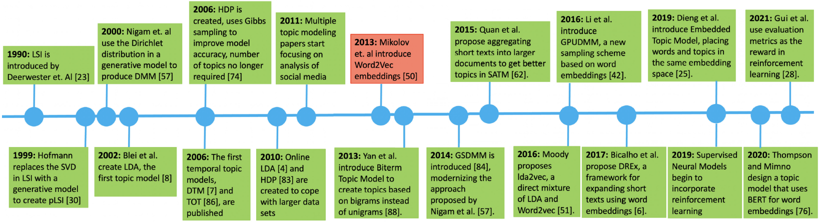
Conclusions and discussion

- 100s of models over 20 years of advances in PTM have been elaborated within overcomplicated Bayesian framework.
- All the while, a high potential of the classical non-Bayesian regularization went almost untested and unnoticed.
- ARTM is a somewhat belated attempt to fill this gap.
- ARTM transforms PTM into «a theory of single Lemma».
- If the community knew about the Lemma, the development of PTM would hardly have followed the Bayesian way... Is not it?
- Neural Topic Models (NTM) is now the main trend in TM.
- The Lemma is applicable for learning neural networks with non-negative normalized vectors as parameters.
- Could non-negativity and normalization constraints be a right direction towards interpretable neural networks?

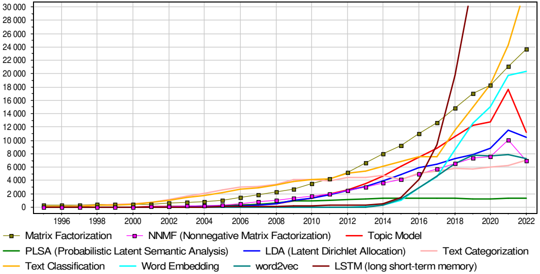
К.Воронцов. Вероятностное тематическое моделирование: теория ARTM и проект BigARTM. 2022.
http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022.

Appendix. The Evolution of Topic Modeling



Citation dynamics:
Topic Modeling and related research areas (from Google Scholar)



Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022.