Fast and Modular Regularized Topic Modeling

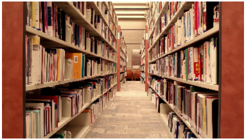
Denis Kochedykov (J.P.Morgan, USA–China)
Murat Apishev (Moscow State University, Russia)
Lev Golitsyn (Integrated Systems, Russia)
Konstantin Vorontsov (MIPT, Russia)

The 21st Conference of Open Innovations Association FRUCT
Seminar on Intelligence, Social Media and Web
Helsinki, Finland • 6–10 November 2017

Contents

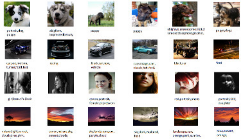
Topic modeling applications

exploratory search in digital libraries

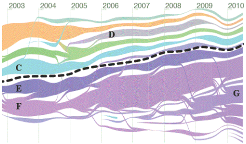


personalized search in social media

multimodal search for texts and images

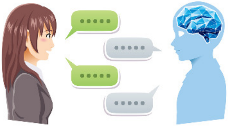


topic detection and tracking in news flows



navigation in big text collections

dialog manager in chatbot intelligence



Example. Multilingual topic model of Wikipedia

216 175 of Russian–English parallel not-aligned 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.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Example. Multilingual topic model of Wikipedia

216 175 of Russian–English parallel not-aligned 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.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

What is a “topic” in a text collection

Intuitively,

- *Topic* is a specific terminology of a particular domain area
- *Topic* is a set of terms that often co-occur in documents

More formally,

- *topic* is a probability distribution over terms (words, tokens): $p(w|t)$ is the frequency of term w in topic t
- *document profile* is a probability distribution over *topics*: $p(t|d)$ is the frequency of topic t in document d

When writing term w in document d author thought of topic t.

Topic model uncovers the set T of latent topics in a text collection.

Problem setup

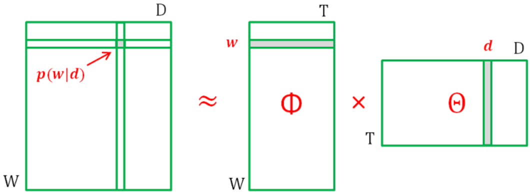
Given: a set of terms W, a set of documents D,
n_{dw} = how many times term w appears in document d

Find: parameters $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$ of the topic model

$$p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td} = \sum_{t \in T} p(w|t)p(t|d).$$

subject to $\phi_{wt} \geqslant 0$, $\sum_w \phi_{wt} = 1$, $\theta_{td} \geqslant 0$, $\sum_t \theta_{td} = 1$.

This is a problem of *nonnegative matrix factorization*:



Well-posed and ill-posed problems in the sense of Hadamard (1923)

The problem is *well-posed* if

- a solution exists,
- the solution is unique,
- the solution is stable w.r.t. initial conditions.

Jacques Hadamard
(1865–1963)

Matrix factorization is an *ill-posed* inverse problem.
If (Φ, Θ) is a solution, then (Φ', Θ') is also the solution:

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

Additional *regularizing criteria* should narrow the set of solutions.

ARTM — Additive Regularization of Topic Model

Maximum log-likelihood with additive combination of regularizers:

$$\sum_{d,w} n_{dw} \ln \sum_{t} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}, \qquad R(\Phi,\Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta)$$

where τ_i are regularization coefficients.

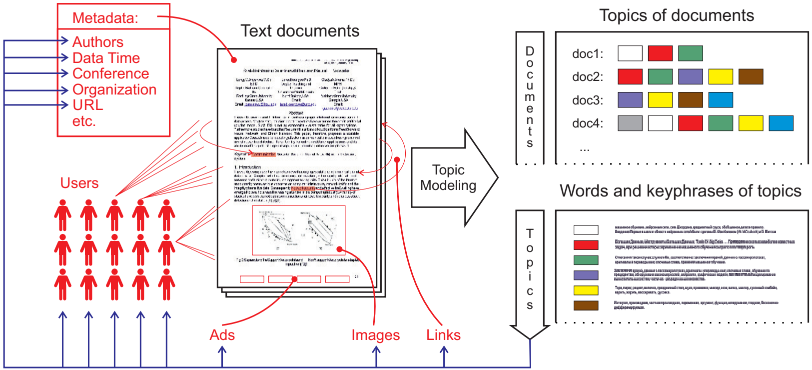
EM-algorithm is a simple iteration method for solving the system

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

where $\operatorname*{norm}_{t\in T} x_t = \frac{\max\{x_t,0\}}{\sum_{s\in T}\max\{x_s,0\}}$ is vector normalization.

Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topic distributions of terms $p(w|t)$ and other modalities: $p(\text{author}|t)$, $p(\text{time}|t)$, $p(\text{category}|t)$, $p(\text{tag}|t)$, $p(\text{link}|t)$, $p(\text{object-on-image}|t)$, $p(\text{user}|t)$, etc.



Multimodal extension of ARTM

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

Maximum multimodal log-likelihood with regularization:

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

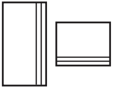
EM-algorithm is a simple iteration method for the system

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

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

Regularizers for the interpretability of topics

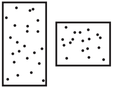
background



Smoothing background topics $B \subset T$:

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

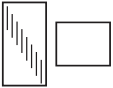
sparse



Sparsing subject domain topics $S = T \setminus B$:

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

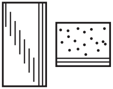
decorrelated



Making topics as different as possible:

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

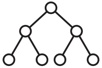
interpretable



Making topics more interpretable
by combining the above regularizers

Many Bayesian PTMs can be reinterpreted as regularizers in ARTM

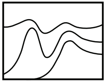
hierarchy



Hierarchical links between topics t and subtopics s:

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

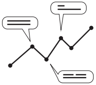
temporal



Topics dynamics over the modality of time intervals i:

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

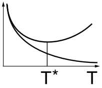
regression



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

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

n of topics

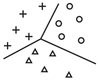


Sparsing $p(t)$ for topic selection:

$$R(\Theta) = -\tau \sum_{t \in T} \frac{1}{|T|} \ln p(t), \quad p(t) = \sum_{d} p(d) \theta_{td}.$$

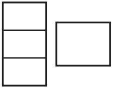
Special cases of the multimodal topic modeling

supervised



The modalities of classes or categories
for text classification and categorization.

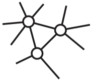
multilanguage



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

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

graph



The modality of graph vertices v with doc sets D_v:

$$R(\Phi) = -\frac{\tau}{2} \sum_{(u,v)\in E} S_{uv} \sum_{t\in T} n_t^2 \left(\frac{\phi_{vt}}{|D_v|} - \frac{\phi_{ut}}{|D_u|}\right)^2.$$

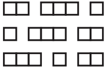
geospatial

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

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

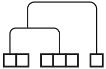
Beyond the “bag-of-words” restrictive hypothesis

n-gram



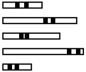
The modalities of n-grams, collocations, named entities

syntax



The modality of n-grams after SyntaxNet preprocessing

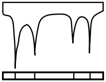
coherence



Modeling co-occurrence data n_{uv} for biterms (u, v):

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

segmentation



E-step regularization affecting $p(t|d, w)$ distributions for segmentation and sentence topic models

BigARTM: open source for fast modular topic modeling

BigARTM features:

- Parallel + online + multimodal + regularized Topic Modeling
- Out-of-core one-pass processing of large text collections
- Built-in library of regularizers and quality measures

BigARTM community:

- 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

Why BigARTM simplifies topic modeling for applications

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

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

Bayesian modeling requires maths and coding at each stage.

ARTM introduces the modular LEGO-style technology, packing each our requirement into a ready-to-use unified building block.

Benchmarking BigARTM vs. Gensim and Vowpal Wabbit

- 3.7M articles from Wikipedia, 100K unique words

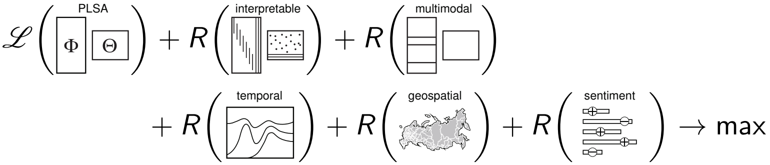
		$T = 50$		$T = 200$	
	procs	time, m	perplexity	time, m	perplexity
BigARTM	1	42	5117	83	3347
BigARTM async	1	25	5131	53	3362
VowpalWabbit	1	50	5413	154	3960
Gensim	1	142	4945	637	3241
BigARTM	4	12	5216	26	3520
BigARTM async	4	7	5353	16	3634
Gensim	4	88	5311	315	3583
BigARTM	8	8	5648	15	3929
BigARTM async	8	5	6220	10	4309
Gensim	8	88	6344	288	4263

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

Mining ethnical discourse in social media

Goal: find topics about inter-ethnic relations using 300 ethnonyms as seed words or modality

The bag-of-regularizers:



Result: the number of ethnically relevant topics augmented from 45 for baseline model (LDA) to 83 for ARTM.

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

Exploratory search in tech news

Goal: exploratory search by long text queries in digital libraries and tech news.

The bag-of-regularizers:

$$\mathscr{L}\Big(\overset{\text{PLSA}}{\boxed{\Phi}\ \boxed{\Theta}}\Big) + R\Big(\overset{\text{interpretable}}{\ }\Big) + R\Big(\overset{\text{multimodal}}{\ }\Big) + R\Big(\overset{\text{n-gram}}{\ }\Big) \to \max$$

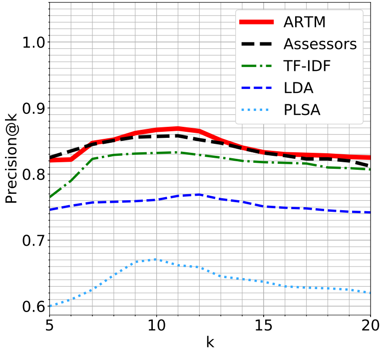
Results:

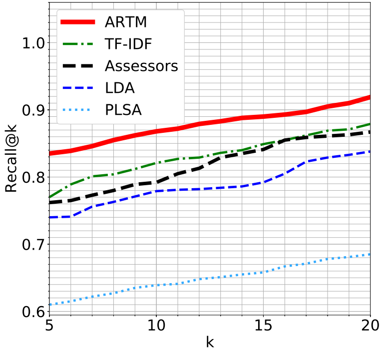
- Precision and Recall augmented +8% on Habrahabr.ru and TechCrunch.com tech news collections.
- Precision and Recall are comparable with assessors' quality.
- The topic-based search engine instantly performs the work that people typically complete in about 30 minutes.

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

Precision and Recall: comparison against baselines

TechCrunch.com text collection, 760K documents
Precision and Recall at top *k* search result positions

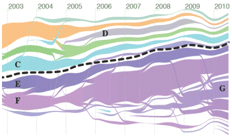




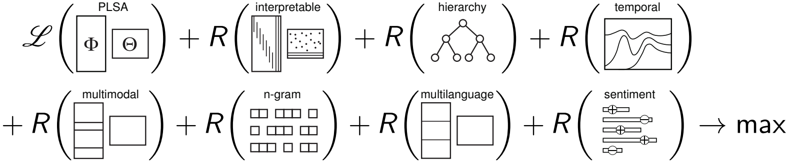
A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Topic detection and tracking in news for media planning

Goal: the development of an interpretable hierarchical temporal dynamic topic model of the news flow.



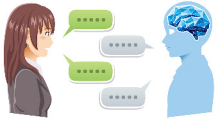
The bag-of-regularizers:



Results: ... (ongoing project)

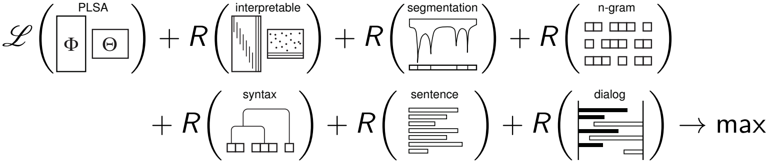
Scenario analysis of call center records

Goals:



- determine typical scenarios of dialogues between operators and customers
- elaborate the quantitative measure of how well operator works
- provide online tips for help operator handle customer's objections

The bag-of-regularizers:



Result: the quality of segmentation augmented from 40% for baselines to 75% for ARTM

Brief summary

- ARTM is a non-Bayesian regularization framework for PTM
- ARTM gives the easy way to formalize and combine PTMs
- ARTM makes it easier to understand and explain PTMs
- ARTM originates the modular "LEGO-style" PTM technology
- BigARTM: open source implementation of ARTM
- Ongoing projects: news, call-center dialogs, bank transactions.

http://bigartm.org

Welcome to use and make contributions!

References

[1] *K.Vorontsov.* Additive regularization for topic models of text collections. Doklady Mathematics, 2014.

[2] *K.Vorontsov, A.Potapenko.* Additive regularization of topic models. Machine Learning, 2015.

[3] *K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova, A.Ianina.* Non-bayesian additive regularization for multimodal topic modeling of large collections. CIKM, 2015.

[4] *K.Vorontsov, A.Potapenko, A.Plavin.* Additive regularization of topic models for topic selection and sparse factorization. SLDS, 2015.

[5] *K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova.* BigARTM: Open source library for regularized multimodal topic modeling of large collections. AIST, 2015.

[6] *O.Frei, M.Apishev.* Parallel non-blocking deterministic algorithm for online topic modeling. AIST, 2016.

[7] *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.

[8] *N.Chirkova, K.Vorontsov.* Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.

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

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

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